Medical Image Analysis 79 (2022) 102456

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Reasoning discriminative dictionary-embedded network for fully automatic vertebrae tumor diagnosis

Shen Zhao[a], Bin Chen[b,*], Heyou Chang[c], Bo Chen[d], Shuo Li[d,*]

[a] Department of Artificial Intelligence, Sun Yat-sen University, Guangzhou 510006, China
[b] Orthopedics Department, The First Affiliated Hospital of Zhejiang University, Hangzhou, Zhejiang, China
[c] Nanjing XiaoZhuang University, Nanjing, Jiangsu, China
[d] Western University, London, Canada

ARTICLE INFO

Article history:
Received 20 July 2021
Revised 1 April 2022
Accepted 8 April 2022
Available online 12 April 2022

Keywords:
Vertebrae tumor diagnosis system
Vertebrae recognition
Dictionary embedded deep learning
Graph reasoning

ABSTRACT

Fully automatic vertebrae tumor diagnosis (FAVTD) means using an end-to-end network to directly perform vertebrae recognition and tumor diagnosis from MRI images. FAVTD is clinically crucial for tumor screening and treatment, which helps prevent further metastasis and save the patients' lives. However, FAVTD has not yet been fully attempted due to the challenges raised by tumor appearance variability as well as MRI image field of view (FOV) and/or characteristics diversity. We propose a **RE**asoning **D**iscriminativ**E** di**C**t**I**onary-embe**D**ded n**E**twork (RE-DECIDE) to tackle the challenges in FAVTD. RE-DECIDE contains an elaborated enhanced-supervision recognition network (ERN) and a self-adaptive reasoning diagnosis network (SRDN). ERN is implemented in a feed-forward dictionary learning manner, which encodes each vertebra by the sparse codes and uses the sparse projections of the vertebrae coordinates onto multiple observation axes for supervision. ERN thus provides multiple sparse encodings of all vertebrae (and their ground truths) to enhance supervision, which reinforces the discrimination of different vertebrae and thus improves recognition performance. SRDN first highlights the most informative feature in the recognized vertebrae based on an attention mechanism. It then performs feature interaction, i.e., exchanges features of different vertebrae based on the graph reasoning mechanism. A reasoning controlling strategy is designed to prompt feature interaction in vertebrae with the same diagnosis labels and meanwhile reduces that in vertebrae with different labels, which avoids over-smoothing and improves diagnosis performance. RE-DECIDE is trained and evaluated using a challenging dataset consisting of 600 MRI images; the evaluation results show that RE-DECIDE achieves high performance in both recognition (accuracy: 0.940) and diagnosis (AUC: 0.947) tasks.

1. Introduction

Fully automatic vertebrae tumor diagnosis (FAVTD) means recognizing each vertebra by classifying its label and regressing its bounding box, and meanwhile diagnosing whether it is invaded by tumors in an end-to-end network. The diagnosing procedure is in essence a machine learning classification task. FAVTD is clinically significant because: (1) Early diagnosis and treatment of spinal tumors, the most fatal processes in spine (Weilbaecher et al., 2011), is crucial to prevent further metastasis and save the patients' lives (Mundy, 2002). (2) FAVTD enables direct diagnosis of vertebrae tumors without manual processes such as vertebrae extraction. FAVTD not only eliminates the time-consuming and labor-intensive work but also provides diagnosis performance that is independent on the experience of the clinicians. Thus, FAVTD may clinically assist radiologists as an automated processor for locating lesions, planning treatments, and preventing further metastasis (Soffer et al., 2019). We mainly consider FAVTD in magnetic resonance imaging (MRI) images because of its sensitivity to soft tissues such as spinal tumors (Shah and Salzman, 2011; Wang et al., 2017b; Chmelik et al., 2018).

However, both recognition and diagnosis tasks in FAVTD are challenging in MRI images. (1) For the recognition task, MRI image characteristics (resolution, scales, and image intensity distribution) vary widely due to the different usages of imaging protocols and endorectal coils. For example, the three images in Fig. 1(a1) have different FOV's (S∼T10, S∼T12, and L4∼T10 respectively). The scales, resolutions, and intensities of these three images are also different (e.g., the vertebrae are of different sizes; those in the second image have low resolutions; those in the first image have

* Corresponding authors.
E-mail addresses: ttbin@zju.edu.cn (B. Chen), sli187@uwo.ca (S. Li).

https://doi.org/10.1016/j.media.2022.102456
1361-8415/

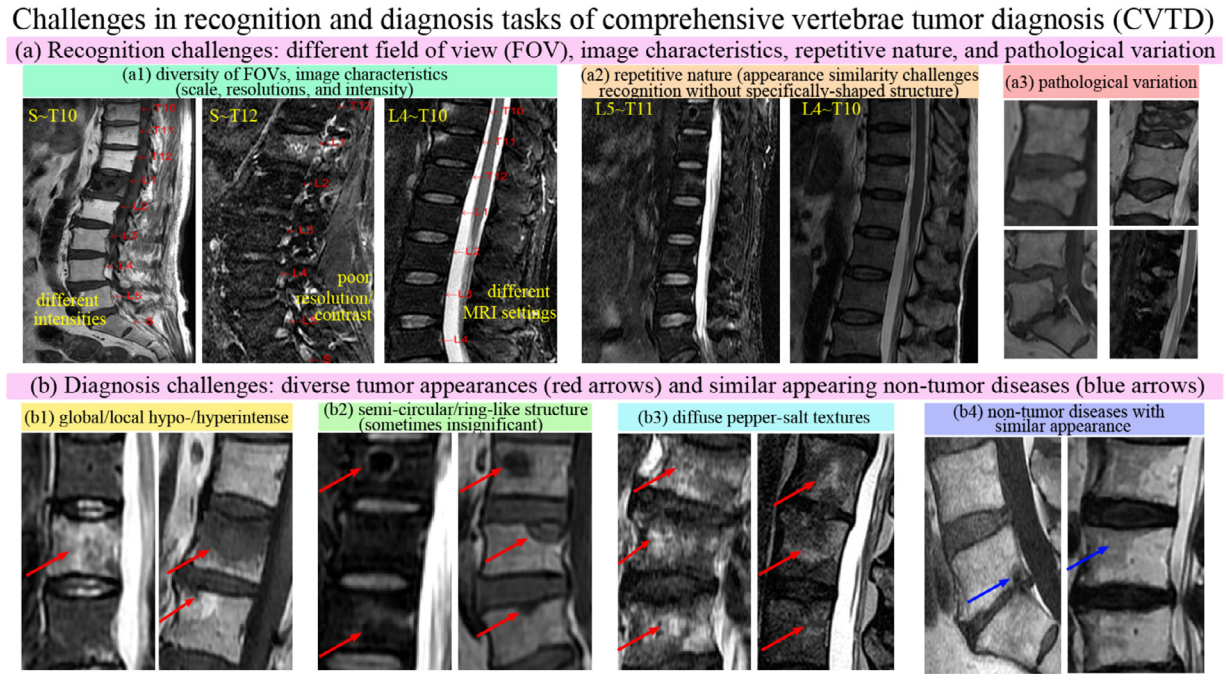

Fig. 1. Both recognition and diagnosis tasks in FAVTD are challenging. (1) Fig. 1(a1∼a3) show the challenges of vertebrae recognition. Fig. 1(a1) shows input images with different FOVs and image characteristics (scales, resolutions, and intensities). For example, the FOVs can be different, i.e., the three images' FOV are respectively S∼T10, S∼T12, and L4∼T10. The scales, resolutions, and intensities can also be different, e.g., the vertebrae in the three images are of different size; the second image has very low resolution; and the vertebral bodies in the first image has higher intensity. Fig. 1(a2) shows vertebrae recognition is challenging because of the vertebrae's repetitive nature, i.e., different vertebrae have similar appearances. This makes it difficult to distinguish an image containing L5∼T11 vertebrae and one containing L4∼T10 (the two images in Fig. 1(a2)). Moreover, in clinical practice, patients often have different body parts examined, which means that there may be no specifically-shaped structure (such as the sacrum). Thus, it is impossible to rely on these structures to assist in recognizing vertebrae with similar appearances, i.e., this adds to the challenge of repetitive nature for distinguishing different vertebrae. Fig. 1(a3) shows that pathological variations can change the appearance of the vertebrae in an unpredictable manner, thus even the same vertebrae can show different appearances. This adds to the challenge of vertebrae recognition. (2) Fig. 1(b1∼b4) shows the challenges of tumor diagnosis. Fig. 1(b1∼b3) shows that tumors have different appearances such as global or local hypo-/hyper intense, semi-circular/ring-like structure, and diffuse pepper-sale textures. Fig. 1(b4) shows that non-tumor diseases such as end-plate osteochondritis may show similar appearances. These issues make tumor diagnosis challenging.

higher intensities). Furthermore, as shown in Fig. 1(a2), the appearances of different vertebrae are similar due to their repetitive nature. This makes it difficult for even experienced physicians to distinguish an image containing different vertebrae (e.g., in the two images in Fig. 1(a2), it is difficult to tell an image containing L5∼T11 vertebrae from one containing L4∼T10). Moreover, in clinical practice, patients often have different body parts examined (i.e., FOV difference), which means that it is not guaranteed that some specifically-shaped structures (such as the sacrum) exist in the input image. Thus, it is impossible to rely on these structures to assist in recognizing vertebrae with similar appearance, i.e., this adds to the challenge of repetitive nature for distinguishing different vertebrae. Besides, as shown in Fig. 1(a3), pathological variations can change the appearance of the vertebrae in an unpredictable manner, thus even the same vertebrae can show different appearances. This makes it more challenging to distinguish the same vertebrae with various appearances from different vertebrae with similar appearances. Wrong recognitions may further result in wrong-site surgery (Zhao et al., 2019b), severe medical malpractice in clinical practice (Makary et al., 2006). (2) For the diagnosis task, the tumor appearance variability raises challenges for distinguishing spinal tumors from other diseases. For example, as shown in Fig. 1(b1)∼(b3), tumors may present local intensity changes in approximately circular or ring-like areas, global hypo-intense, or hyper-intense, and diffuse pepper-salt like textures in images of different MRI modalities and tumor pathology. However, as a contrast, other spinal diseases, such as end-plate osteochondritis (Fig. 1(b4)), may appear similar to tumors (e.g., the left figure of Fig. 1(b4) and the right figure of Fig. 1(b2)) that are even difficult for experienced clinicians to distinguish. Furthermore, the diagnosis task may suffer from massive irrelevant interference information (which may show tumor-like patterns) in the non-vertebrae parts in the input MRI image if the recognition work is not well performed. This shows the necessity of simultaneously performing the recognition and diagnosis tasks in an end-to-end system.

FAVTD in MRI images has not yet been attempted in the existing literature. For closely relative works, some researchers automatically detect metastases from CT images. These methods mainly first extract the spinal region using classical image processing or machine learning methods and then perform tumor diagnosis using the features in the spinal region. For example, (Burns et al., 2013) first identifies the spine canal using region growing, and then detect lesions using the watershed algorithm; (Wiese et al., 2011) uses thresholding and region growing to segment the spine, then uses the watershed algorithm for lesion candidate detection, and lastly uses support vector machines to diagnose metastasis; (Chmelik et al., 2018) uses intensity projections and adaptive filters to locate the spine, then uses 3D CNN's to perform tumor diagnosis. These methods show accurate and promising results in CT images, however, they may not be robust enough to properly handle MRI images.

Several works also perform single diagnosis or vertebrae detection/recognition tasks, which have the potential to be extended into FAVTD. (1) For tumor diagnosis, (Wang et al., 2017b) detects tumors from manually extracted MRI patches using parallel convolutional neural networks (CNN) to deal with different input resolutions. (2) Much more work has been attempted for vertebrae detection/recognition. Here, we define "capturing the vertebrae centroid points" as "detection", whereas "joint vertebrae classification and bounding box regression" as "recognition" to avoid con-

fusion. For vertebrae detection, (Glocker et al., 2013) uses random forests and probabilistic graphical models to regress vertebrae centroid points; (Chen et al., 2015) uses CNN's jointly trained with a shape regression model to extract more robust features for vertebrae detection; (Yang et al., 2017) uses deeply supervised CNN enhanced by message passing to accurately predict pixel-wise probability maps of each vertebrae centroid. These works can precisely capture each vertebrae centroid, however, simultaneously classifying their labels and regressing their bounding boxes may be more clinically meaningful for the succeeding diagnosis procedure (Zhao et al., 2019b; 2021). For vertebrae recognition, (Lootus et al., 2014) presents an accurate method using the deformable part model detector and dynamic programming; (Windsor et al., 2020) proposes a two-stage detecting-labeling CNN for accurate vertebrae recognition in whole MRI scan. However, (Lootus et al., 2014) needs the sacrum to be present, whereas the training set used in (Windsor et al., 2020) contains only MRI images containing the sacrum. Other object detection, recognition, or segmentation methods, such as the active contour methods (Zhao et al., 2017), Faster RCNN (Ren et al., 2015), YOLO detector (Yang and Deng, 2020), and SSD detector (Liu et al., 2016), have also been used for finding human tissues or lesions from medical images (Guo et al., 2021; Gao et al., 2019). In all, the above works have provided a reliable detection and diagnosis (machine learning classification) algorithm, which lays a solid foundation for FAVTD.

Dictionary learning (sparse coding) has the potential to benefit FAVTD because it can obtain discriminative features. For example, (Sun et al., 2019) designs a supervised dictionary learning network for enhanced image classification performance. (Jiang et al., 2013) proposes a label consistency strategy to prompt samples with the same class labels have similar sparse codes. (Zhang et al., 2020) develops an enhanced dictionary learning method for object detection in 3D ultrasound patches. However, two difficulties hinder its application in FAVTD: (1) The dictionary and the sparse codes are generally trained in an alternate manner (Aharon et al., 2006), which is difficult to be integrated into the end-to-end training of CNN's. (2) The ground truth sparse codes are typically difficult to obtain. Traditional dictionary learning uses unsupervised reconstruction to obtain sparse codes, which may not be optimal for the main recognition (Zhao et al., 2019a) and diagnosis tasks (Coates and Ng, 2011). (Liu et al., 2018) tries to combine dictionary learning with CNN's for scene recognition. We (Zhao et al., 2021) also conduct preliminary research on combining dictionary learning with CNN's in an image detection framework. However, these methods only calculate the sparse codes and use them for classification. Nevertheless, the potential of sparse codes to encode the sparsely distributed vertebrae can be further exploited to enhance recognition performance.

Graph reasoning may help to capture relation-aware information to improve diagnosis (classification) accuracy in FAVTD, however, it has not been widely exploited due to over-smoothing issues. For example, (Wang et al., 2018) purposes a non-local network for leveraging spatial and sequential relationships for video classification; (Chen et al., 2019b; Liang et al., 2019) uses graph reasoning between label features for multi-label image recognition or object detection tasks; (Chen et al., 2019a) uses graph reasoning for capturing global relations between distant regions for image classification and semantic segmentation. For tumor diagnosis, when it is difficult to distinguish spinal tumors from similar non-tumor diseases, graph reasoning allows diagnostic features of other vertebrae to help diagnose (Jiang et al., 2020) based on feature similarity. However, since graph reasoning is in essence weighted average between different nodes (i.e., recognized vertebrae in our work), it may cause over-smoothing (i.e., the features of nodes with different labels may be "mixed" together, which may cause features of nodes with different labels to become similar) (Chen et al., 2019b) and on the contrary harm feature discrimination. Thus, it would be interesting to explore a method to leverage graph reasoning for capturing relation dependencies while avoiding over-smoothing.

We propose a novel reasoning discriminative dictionary-embedded network (RE-DECIDE) for FAVTD, which overcomes the MRI FOV and image characteristics variety, vertebrae repetitive nature, and tumor appearance variability in the recognition and diagnosis tasks. Our RE-DECIDE performs two tasks in an end-to-end framework: firstly, vertebrae labels (e.g., T11, T12, L1, L2, L3, etc.) and bounding boxes are predicted by the recognition network, i.e., the recognition task; then, vertebrae diagnostic labels (e.g., tumor/non-tumor) are predicted by the diagnosis network, i.e., the diagnosis task. As shown in Fig. 2:

- To overcome the FOV and image characteristics challenges in the recognition task, an enhanced-supervision recognition network (ERN) is designed to use projection-based sparse codes to encode vertebrae and prompt discrimination of different vertebrae. ERN encodes each vertebra by predicting L sparse codes via feed-forward dictionary learning. The sparse codes are trained to approach the projections of ground truth angular points onto L observation axes (OA). Since the projections of different vertebrae on different OAs exhibit adequate discrepancy, the trained sparse codes have better distinguishability of different vertebrae. Under the projections' guidance, the ensemble of predicted sparse codes helps to distinguish different vertebrae (Xie et al., 2017; Quan et al., 2016).
- To overcome the tumor appearance variability in the diagnosis task, a self-adaptive reasoning diagnosis network (SRDN) is designed to interact between vertebrae features considering their diagnostic labels. This leverages the relational clues between different recognized vertebrae to contribute to each other's final diagnosis predictions. Furthermore, to prompt the validity of graph reasoning and avoid over-smoothing, SRDN leverages attention mechanism to highlight the most informative features for tumor diagnosis; also, it designs a self-adaptive reasoning controlling strategy to facilitate feature interaction between vertebrae of the same diagnosis labels while reducing feature interaction between those of different diagnosis labels. This alleviates the over-smoothing problem while allowing features from easy-to-diagnosis vertebrae to assist in diagnosing the difficult ones.

In all, after adopting HPN (Zhao et al., 2021) for feature extraction and coarse regional proposal localization, our RE-DECIDE elaborately designs two sub-networks, i.e., ERN and SRDN, to individually perform the recognition and diagnosis tasks in FAVTD; ERN and SRDN share some features and forms an end-to-end network.

Our contributions can be summarized as:

(1) For the first time, an accurate computer aided diagnosis (CAD) tool is designed to perform vertebrae recognition and tumor diagnosis together in a fully automatic end-to-end network based on graph reasoning and dictionary learning mechanisms. This work effectively reduces the burden on clinicians to manually analyze the medical data.

(2) Projection-guided dictionary learning is embedded into a CNN-based recognition framework in a forward propagation manner to encode each vertebra by sparse codes. This strategy leverages the projections of vertebrae angular points on different OAs for enhanced supervision, which prompts the discrimination of vertebrae with repetitive appearances in MRI images of different FOV's.

(3) For the first time, we develop a self-adaptive graph reasoning diagnosis (classification) method that can control the feature interaction weights according to graph node labels. This strategy avoids over-smoothing while keeping the advantages of graph rea-

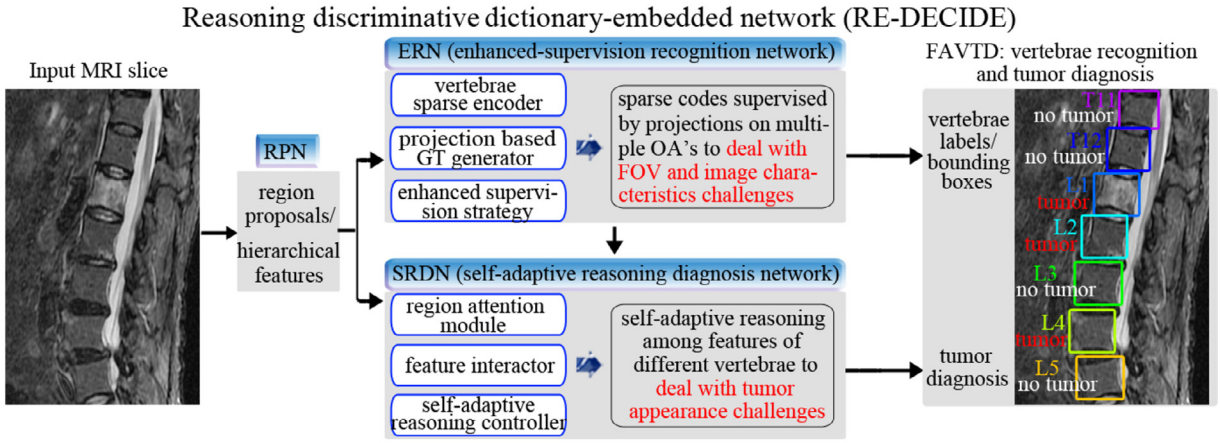

Fig. 2. RE-DECIDE addresses the challenges of comprehensive vertebrae tumor diagnosis by two elaborated models ERN and SRDN. ERN encodes vertebrae by multiple sparse codes, which are supervised by the ground truth projections for enhanced supervision to tackle the recognition challenges. SRDN allows different vertebrae to contribute to each other based on their feature similarity, which assists diagnosis of the "difficult" vertebrae by means of other vertebrae.

soning, which prompts the discrimination between spinal tumors and other similar-appearing diseases.

In this work, we advance our preliminary attempt on vertebrae detection in MICCAI 2020 (Zhao et al., 2020) in the following aspects: (1) We propose a self-adaptive reasoning diagnosis network to improve diagnosis performance. We also conduct elaborate experiments to explore when graph reasoning is beneficial for classification tasks. (2) We propose a region attention module to focus on the most informative diagnostic features. (3) We provide more detailed descriptions and discussions on the dictionary learning module used in (Zhao et al., 2020) for a better demonstration of how the projection-guided enhanced supervision is implemented and how it helps FAVTD. (4) A more comprehensive review of FAVTD (as well as its closely relative work), dictionary learning, and graph reasoning is conducted to provide a panorama of existing work.

2. Methodology

Our reasoning discriminative dictionary-embedded network (RE-DECIDE, Fig. 2) is an end-to-end framework for fully automatic vertebrae tumor diagnosis (FAVTD). RE-DECIDE first adopts a hierarchical proposal network (HPN, Section 2.1) (Zhao et al., 2021) to coarsely locate regions containing vertebrae. Then, two cascading modules are deliberately designed to respectively perform the recognition and diagnosis task: (1) **enhanced-supervision recognition network** (ERN, Section 2.2) takes the coarse regions and the corresponding features as input; it then outputs the recognized vertebrae labels and bounding boxes. ERN contains a **vertebrae sparse encoder** that designs a feed-forward dictionary learning layer to obtain sparse codes encoding each vertebra, and a **projection-based ground truth generator** that leverages the projections of each vertebra on L observation axes for ground truth sparse codes. ERN also develops an **enhanced supervision strategy** for the ensemble of different predictions to improve the generalized vertebrae recognition accuracy and tackle the FOV/characteristics challenges. (2) **self-adaptive reasoning diagnosis network** (SRDN, Section 2.3) takes the recognized vertebrae as input; it finally outputs a diagnostic prediction indicating whether each vertebra is invaded by tumors. SRDN contains a **region attention module** that helps the network to focus on the most informative diagnostic features and a **feature interactor** that allows the diagnostic features of one vertebra to help to diagnose the others in a graph reasoning manner. SRDN also designs a **self-adaptive reasoning strategy** for preventing over-smoothing in the reasoning procedure and alleviating the tumor appearance variability challenges.

2.1. Brief retrospect of hierarchical proposal network (HPN)

2.1.1. Hierarchical proposal network (HPN)

To present a complete and comprehensible workflow of our RE-DECIDE, we firstly briefly retrospect the Hierarchical proposal network (HPN) and show how is it interfaced with the succeeding enhanced-supervision recognition network (ERN) and self-adaptive reasoning diagnosis network (SRDN). HPN takes the original input MRI slice as input. It generates multi-scale anchors at different regular locations, extracts hierarchical image features corresponding to all anchors, and predicts which anchors contain vertebrae as well as the coarse locations of the vertebrae by generating proposals. More detailed knowledge of HPN can be found in our previous work (Zhao et al., 2021). All implementation details of HPN (e.g., the number of layers, blocks, and structure of each block) are the same as in our previous work. The output proposals of HPN include positive and negative proposals. The positive proposals are multi-scale rectangle boxes that coarsely cover class-agnostic vertebrae. The negative proposals are non-vertebral regions in the image (usually, non-maximum suppression is used to select the most difficult negative proposals to accelerate training). Both positive and negative proposals are used to train the succeeding recognition network ERN (Section 2.2), whereas only recognized vertebrae (corresponding to positive proposals) are used to train the diagnosis network SRDN (Section 2.3). Besides the proposals, the hierarchical image features are shared to the succeeding network for recognition and/or diagnosis tasks.

2.1.2. The interface of HPN and ERN

After obtaining the proposals and the hierarchical image features using HPN, ROI aligning (Zhao et al., 2021) is adopted to choose features of the most suitable scale from the hierarchical image features, crop the chosen feature using the proposals, and then resize them to a certain size (7×7 in our work). After ROI aligning, each proposal corresponds to a $7 \times 7 \times 256$ feature map. Then, the feature maps are fed into 2 cascading convolutional layers with "VALID" paddings (the first one has a kernel size of $7 \times 7 \times 256 \times 1024$, and the second $1 \times 1 \times 1024 \times 1024$, stride 1) followed by batch normalization layers and ReLU activation layers; the feature size of the second convolutional layer are thus $1 \times 1 \times 1024$. Lastly, for each proposal, its features are flattened into vectors (denoted as $\mathbf{x}_i \in \mathbb{R}^M$) by squeezing the dimen-

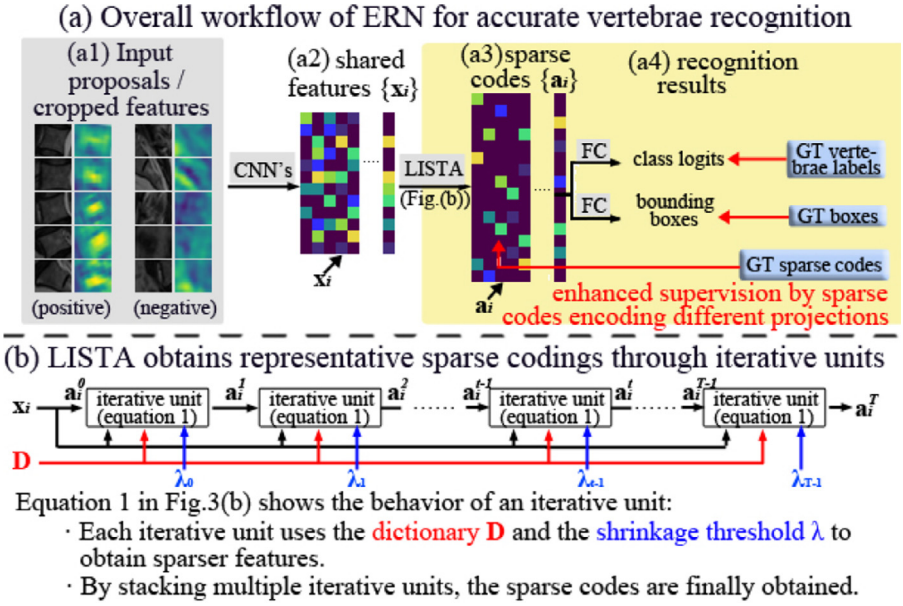

Fig. 3. The enhanced-supervision recognition network (ERN). In ERN, Regional proposals and features are firstly obtained as in (Ren et al., 2015) and (Zhao et al., 2021) (Fig. 3(a1∼a2)). Then, our ERN embeds each vertebrae by L sparse codes using feed-forward dictionary learning methods such as LISTA (Fig. 3(a3)). The sparse codes are then used for predicting the labels and bounding boxes of each vertebra (Fig. 3(a4)). The sparse codes are supervised by the projections of the vertebrae's angular point coordinates to L observation axes (detailed in Fig. 4) for enhanced supervision; and the ensemble of their predictions can improve recognition performance.

sionalities with size 1 and fed into ERN (Section 2.2) for vertebrae recognition (joint vertebrae label classification and bounding box regression).

2.1.3. The interface of HPN, ERN, and SRDN

After recognizing the vertebrae existing in an input MRI slice, a procedure similar to ROI aligning is performed for the diagnosis task. In this procedure, features of the most suitable scale are still chosen from the hierarchical image features, and then the chosen feature is cropped and resized to a certain size (32×32 in our work). The only difference is that the cropping uses the recognized vertebrae bounding boxes (the output of ERN) instead of the proposals. After this procedure, each vertebrae corresponds to a $32 \times 32 \times 256$ diagnostic feature (denoted as $\mathbf{F}_i \in \mathbb{R}^{h \times w \times c}$). Then, $\mathbf{F}_i$s are fed into the region attention module (Section 2.3.2) of SRDN (Section 2.3) to obtain the modulated diagnostic feature $\mathbf{F}'_i$. $\mathbf{F}'_i$s are next fed into a simple network (this network first uses 3 cascading convolutional layers of sizes $3 \times 3 \times 256 \times 512$, $3 \times 3 \times 512 \times 512$, and $3 \times 3 \times 512 \times 1024$, stride 1, each with "SAME" paddings and followed by a batch normalization layer, a ReLU activation layer, and a max-pooling layer; then, it uses 2 convolutional layers with kernel size $4 \times 4 \times 1024 \times 1024$ and $1 \times 1 \times 1024 \times 1024$ with "VALID" paddings followed by batch normalization layers); the output feature size of the network is thus $1 \times 1 \times 1024$. Next, the dimensionalities with size 1 are squeezed to obtain the flattened diagnostic features $\mathbf{y}_i$ for every vertebra. The $\mathbf{y}_i$s are fed to the feature interactor with a self-adaptive reasoning controller (Sections 2.3.3 and 2.3.4) in SRDN to predict whether a vertebra is invaded by tumors.

2.2. Enhanced-supervision recognition network (ERN)

2.2.1. Overall workflow

The inputs of ERN are the regional proposals (including positive and negative proposals) and hierarchical image features provided by the HPN (Fig. 3(a1)). As mentioned in Section 2.1.2, each proposal's features are firstly flattened into vectors (denoted as $\mathbf{x}_i$ for the ith proposal, Fig. 3(a2))) by ROI aligning and some convolutional layers. Then, $\mathbf{x}_i$ are fed into the **vertebrae sparse encoder** (Fig. 3(a3)/(b), Section 2.2.2) to calculated the sparse codes $\mathbf{a}_i$ for each proposal. Meanwhile, the ground truth sparse codes $\mathbf{a}_i^*$ are calculated by the **projection-based ground truth generator** (Fig. 4, Section 2.2.3). The $\mathbf{a}_i$ are finally used to predict the labels and bounding boxes by inverse projection. Losses from label classification, bounding box regression, and sparse code prediction are all used for **enhanced supervision** in network training (Fig. 3(a4), Section 2.2.4).

2.2.2. Vertebrae sparse encoder

The vertebrae sparse encoder is designed to acquire representative sparse codes for recognizing different vertebrae. It designs a feed-forward dictionary learning layer to calculate the sparse code $\mathbf{a}_i$ for each $\mathbf{x}_i$. The subtlety of dictionary learning in recognition tasks is that the objects (vertebrae) are sparsely distributed, i.e., a vertebra has only four angular points, whereas an image has much more pixels. This triggers the thought to use sparse codes $\mathbf{a}_i$ to encode the angular points' positions; furthermore, it also draws forth an interesting idea of enhancing the supervision of sparse codes by predicting multiple $\mathbf{a}_i$'s and supervising them with projections of ground truth angular points onto different OA's (Xue et al., 2019). Meanwhile, it is confirmed in the compressive sensing community that $\mathbf{a}_i$ is able to be obtained by $\mathbf{x}_i$ (the output of CNN's) by minimizing $\frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2^2 + \lambda\|\mathbf{a}_i\|_1$ over $\mathbf{a}_i$. Inspired by the LISTA algorithm (Gregor and LeCun, 2010), we use Eq. (1) (visually demonstrated in Fig. 3(b)) to obtain $\mathbf{a}_i$:

$$\mathbf{a}_i^t = \eta(\mathbf{a}_i^{t-1} + \beta\mathbf{D}^T(\mathbf{x}_i - \mathbf{D}\mathbf{a}_i^{t-1}); \lambda^t),$$
$$\text{where } \eta(\mathbf{r}; \lambda) = \text{sgn}(\mathbf{r}) \max\{|\mathbf{r}| - \lambda, 0\} \quad (1)$$

Eq. (1) demonstrates the principles how the vertebrae sparse encoder iteratively compute the sparse codes $\mathbf{a}_i$ with the vertebrae feature $\mathbf{x}_i$. In Eq. (1), the sparse code of the ith proposal $\mathbf{a}_i^t$ is updated iteratively by the shrinkage function η. η is a thresholding function that processes its input $\mathbf{r}$ element by element: For each element r_j, threshold λ is subtracted from its original absolute value $|r_j|$; if $|r_j| - \lambda < 0$, this element is set to 0 in the next iteration. T iterations (typically $T = 3 \sim 6$) are applied to calculate reliable $\mathbf{a}_i$. The superscript t means the iteration number. In our

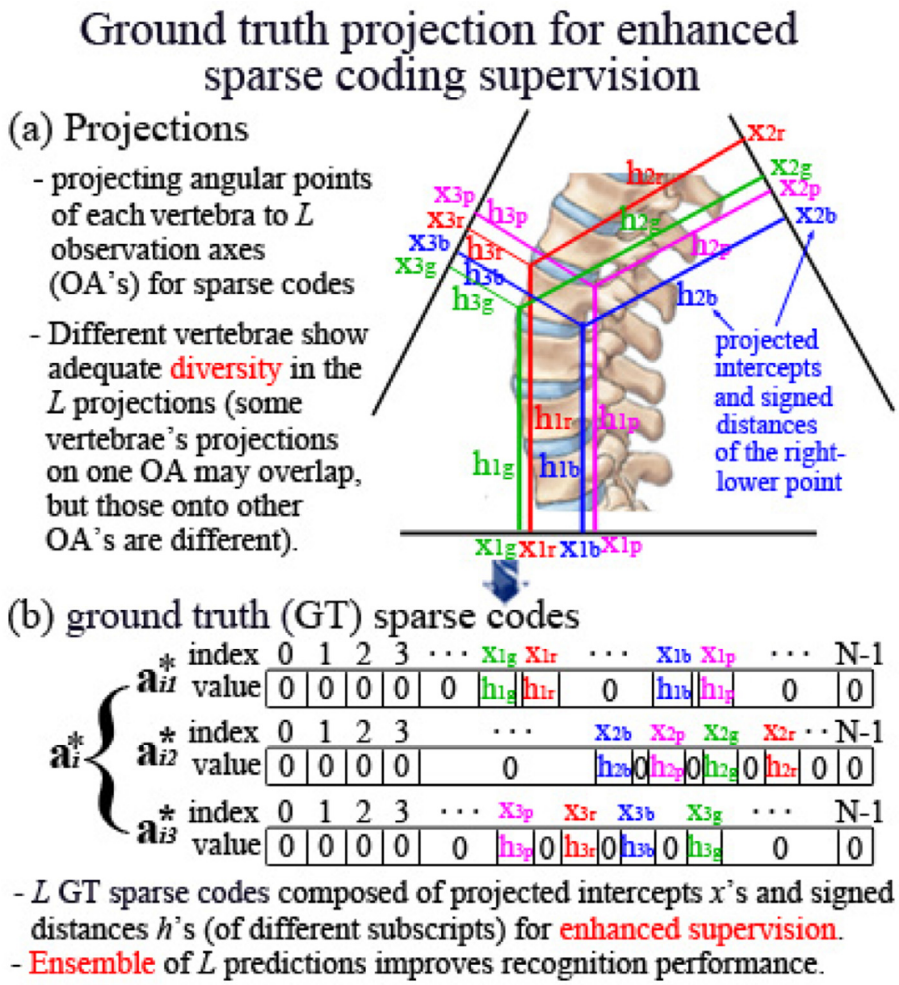

Fig. 4. Detailed illustrations of the projection-based enhanced supervision strategy in ERN. Fig. 4(a) shows how the projection is conducted, and Fig. 4(b) shows how the GT sparse codes are formed. This strategy fully leverages the diversity of projections onto different OAs for improving recognition discrimination.

design, since Eq. (1) is a differentiable architecture, the dictionary $\mathbf{D}$ as well as all λ^t's can be trained together with the preceding CNN's in an end-to-end manner.

2.2.3. Projection-based ground truth generator

The projection-based ground truth (GT) generator is designed for generating ground truths $\mathbf{a}_i^*$ to supervise the sparse codes, which is the key procedure of supervision enhancement. The most straightforward method for generating GT sparse codes is to establish a sparse matrix whose size is the same as the input image. The elements in the matrix corresponding to the vertebrae angular point coordinates are 1 and the others are 0. Then, the sparse matrix can be resized to 1D vectors as GT sparse codes encoding vertebrae positions. However, this strategy may cause the sparse vector to be too large (e.g., its length would reach 262,144 if the input image is 512 × 512). Thus, we consider leveraging the angular point coordinates' projections onto L OAs around the input image. For each vertebra (corresponding to a positive proposal in Fig. 3(a1)), its four angular points would form four intercepts x's and signed distances h's when projected onto one OA; then the x's and h's would form a sparse vector. For example, as shown in Fig. 4(a), when the four angular points are projected to the horizontal OA, four x's and h's are obtained (those with subscript 1r, 1b, 1g, 1p in Fig. 4(a)); they then establish a sparse vector ($\mathbf{a}_{i1}^*$ in Fig. 4(b), where the values at positions $x_{1r}/x_{1b}/x_{1g}/x_{1p}$ of the vector are respectively $h_{1r}/h_{1b}/h_{1g}/h_{1p}$, whereas the other positions are 0). Then, the sparse vectors established by other OA's (e.g., $\mathbf{a}_{i2}^*$ and $\mathbf{a}_{i3}^*$ in Fig. 4(b)) are concatenated together to form the GT sparse code $\mathbf{a}_i^*$. As shown in Fig. 4, the orientations of these axes are uniformly distributed (for clarity, the projections of only one vertebra to $L = 3$ axes are demonstrated). For each non-vertebrae region (corresponding to a negative proposal in Fig. 3(a1)), its ground truth sparse codes are set to zero vectors.

2.2.4. Enhanced supervision strategy

After obtaining the predicted $\mathbf{a}_i$ and the ground truth $\mathbf{a}_i^*$ (for each vertebrae, they are vectors of length "$image\_size \times L$"), we design a loss function for enhanced supervision. Firstly, two sibling fully connected (FC) layers (respectively of sizes "$(image\_size \times L) \times (vertebrae\_class\_numbers \times L)$" and "$(image\_size \times L) \times (vertebrae\_class\_numbers \times L \times 4)$)" are used as inverse projections; they take $\mathbf{a}_i$ as input and separately output L object class probability vectors and $4 \times L$ bounding box coordinates for each class. Then, two other fully connected layers (respectively of sizes "$(vertebrae\_class\_numbers \times L) \times vertebrae\_class\_numbers$" and "$(vertebrae\_class\_numbers \times L \times 4) \times vertebrae\_class\_numbers \times 4$") are used for the ensemble learning for vertebrae classification and bounding box regression. Next, as in our previous work (Zhao et al., 2019b),

message passing method is leveraged for vertebrae class probability calibration. Finally, all these calibrated class probabilities $\mathbf{u}_{i_1,l}$, bounding boxes $\mathbf{v}_{i_2,l}$, and sparse codes $\mathbf{a}_{i_3,l}$ are supervised by the corresponding ground truths $u^*_{i_1}$, $\mathbf{v}^*_{i_2}$, and $\mathbf{a}^*_{i_3,l}$, i.e., the total loss function of the recognition task is:

$$L_r = \frac{\lambda_1}{N_1}\sum_{i_1=1}^{N_1}\sum_{l=1}^{L} L_{ce}(\mathbf{u}_{i_1,l}, u^*_{i_1}) + \frac{\lambda_2}{N_2}\sum_{i_2=1}^{N_2}\sum_{l=1}^{L} L_{sl}(\mathbf{v}_{i_2,l}, \mathbf{v}^*_{i_2}) + \frac{\lambda_3}{N_3}\sum_{i_3=1}^{N_3}\sum_{l=1}^{L} L_{sl}(\mathbf{a}_{i_3,l}, \mathbf{a}^*_{i_3,l}) \quad (2)$$

Eq. (2) means the losses of the recognition task contains three parts, i.e., the vertebrae label classification loss, the vertebrae location regression loss, and the sparse code regression loss. For more details: (1) $L_{ce}(\mathbf{u}_{i_1,l}, u^*_{i_1})$ means the cross entropy loss of the predicted class probabilities $\mathbf{u}_{i_1,l}$ produced by the lth sparse code and the ground truth label $u^*_{i_1}$, N_1 is the total proposal number. (2) $L_{sl}(\mathbf{v}_{i_2,l}, \mathbf{v}^*_{i_2})$ means the average of the smooth L1 loss (Ren et al., 2015) of all elements in vector $\mathbf{v}_{i_2,l} - \mathbf{v}^*_{i_2}$, i.e., the difference between the lth sparse code's prediction of the i_2th vertebra's bounding box coordinates $\mathbf{v}_{i_2,l}$ and the corresponding ground truth $\mathbf{v}^*_{i_2}$, N_2 is the positive proposal number. (3) $L_{sl}(\mathbf{a}_{i_3,l}, \mathbf{a}^*_{i_3,l})$ means the smooth L1 loss of each predicted sparse code and its ground truth, N_3 is the total sparse code number. In this way, our ERN provides L supervision and L predictions (class probabilities, bounding boxes, and sparse codes) for each vertebra for enhanced supervision. The ensemble of the L predictions determine the final recognitions, which improves the recognition discrimination and handles the FOV/image characteristics challenge. (4) The weights $\lambda_1 \sim \lambda_3$ are chosen based on the experience of our previous work (Zhao et al., 2021; 2019b). For λ_1 and λ_2, they are chosen to be 1 as in (Zhao et al., 2021). For λ_3, we set it as 0.05 so that the sparse loss is no larger than half of the main recognition loss (Zhao et al., 2019b), however, we find that the recognition results do not change much when λ_3 varies from 0.05 to 0.5.

2.2.5. Discussions

The reason that ERN helps distinguish different vertebrae is twofold.

loo(1) The projection-based sparse codes make better use of the locational information of different vertebrae to improve discrimination. The projection-based sparse codes help RE-DECIDE to better discriminate different vertebrae than the classical object detectors (such as Faster RCNN, YOLO, and SSD) as well as our previous method that also uses dictionary learning (Zhao et al., 2021). The initial motivation of using projection-based sparse codes is that, sparse coding, generally speaking, has the potential to obtain discriminative features. Thus, our previous work conducts preliminary research on combining dictionary learning with CNNs in an image detection framework. However, it only uses dictionary learning based on an embedded k-sparse autoencoder to prompt the discrimination of the proposal features. Nevertheless, the vertebrae are sparsely distributed in the input image, which triggers the thought to use make better use of the sparse codes to encode the vertebrae and prompt the discrimination of similar-appearing vertebrae of the classical detectors. Thus, our current work, as a contrast, leverages the discrepancy of projections of vertebrae angular points on different OAs to be more aware of the locational information of different vertebrae. As mentioned in (Windsor et al., 2020), the vertebra location can be used to assist in predicting vertebrae labels, while in ERN, the sparse codes formed by projections on different OAs can take fuller advantage of the locational discrepancy of different vertebrae when constructing and supervising the sparse codes. In our ERN, after calculating the sparse codes to encode each vertebra, the sparse codes are supervised using the projection of the ground truth angular point coordinates onto L OAs. This richens the locational discrepancy, for example, the projections of some vertebrae onto one OA overlap, those onto other OAs can still show enough locational discrepancy because the OAs' orientations are diverse (Xue et al., 2019). In this way, the projections onto the L OAs bring more discrepancy for different vertebrae to prompt discrimination, i.e., the recognition performance of distinguishing similar-appearing vertebrae of different labels is thus improved. Using all L GT projections to simultaneously supervise the predicted ones (i.e., enhanced supervision), the final recognitions are determined by the ensemble of the L predictions, which helps lower risks of over-fitting (Quan et al., 2016) compared with the classical detectors such as YOLO, SSD, and Faster RCNN. To summarize, the projection-guided dictionary learning strategy is more beneficial for vertebrae recognition.

(2) Setting the sparse codes of negative proposals to zero vectors helps the recognition network to be more discriminative of positive and negative proposals, i.e., vertebrae and non-vertebrae regions, which is very beneficial in our workflow where message passing is used for correcting the pre-recognition results (detailed in (Zhao et al., 2021)). This strategy helps the network to distinguish vertebrae and non-vertebrae regions because increases the discrepancy of the features of positive and negative regions, i.e., through training, the features of vertebrae regions would gradually approach the ground truth sparse codes obtained by projections; meanwhile those of the non-vertebrae regions would approach 0. This is crucial for the message passing calibration because it can better guarantee a correct neighboring relationship of the pre-recognized vertebrae sequence that is fed into the message passing. In more detail, the validity of the message passing calibration algorithm relies on the neighboring relationship of its inputs (i.e., the pre-recognized vertebrae sequence). The message passing calibration can be summarized as first sorting the pre-recognized vertebrae into a sequence (i.e., the neighboring relationship is determined by this sorting procedure using the relative positions of the pre-recognized boxes), and then performing class probability vector (CPV) calibration on the pre-recognition sequence. If the pre-recognitions correctly distinguish the vertebrae and non-vertebrae regions, each recognized box will correspond to the CPV of an existing vertebra; even if the CPV is wrong, its neighboring relationships are correct, and the neighboring pre-recognitions can use their CPVs to calibrate it. On the other hand, if the pre-recognition mistakes vertebrae with non-vertebrae regions, the sorting procedure may yield wrong neighboring relationships (i.e., inserting a false positive into the sequence or missing a vertebra in the sequence), which will result in the calibration process being invalid. For example, in Fig. 5(a), the existing vertebrae in the image are S1∼T12, but there is a false positive in the pre-recognitions shown by the yellow dashed box, i.e., vertebrae and non-vertebrae regions are mistaken. This results in one more CPV being inserted into the sorted pre-recognition sequence, i.e., a non-vertebrae region will erroneously correspond to one CPV. This ruins the calibration procedure, i.e., the yellow dashed box will be taken as L2, while the labels of the L1 and T12 will be taken as T12 and T11, i.e., they are out by one. Also, in Fig. 5(b), the T12 is missing in the pre-recognitions, i.e., vertebrae and non-vertebrae regions are again mistaken. This results in one CPV being missing in the sorted pre-recognition sequence. The message passing can not calibrate the missing CPV, i.e., the missing vertebrae can not be retrieved. Although in our previous work, a x coordinate threshold is designed to alleviate this problem, the hard threshold may be over-fitted to some datasets.

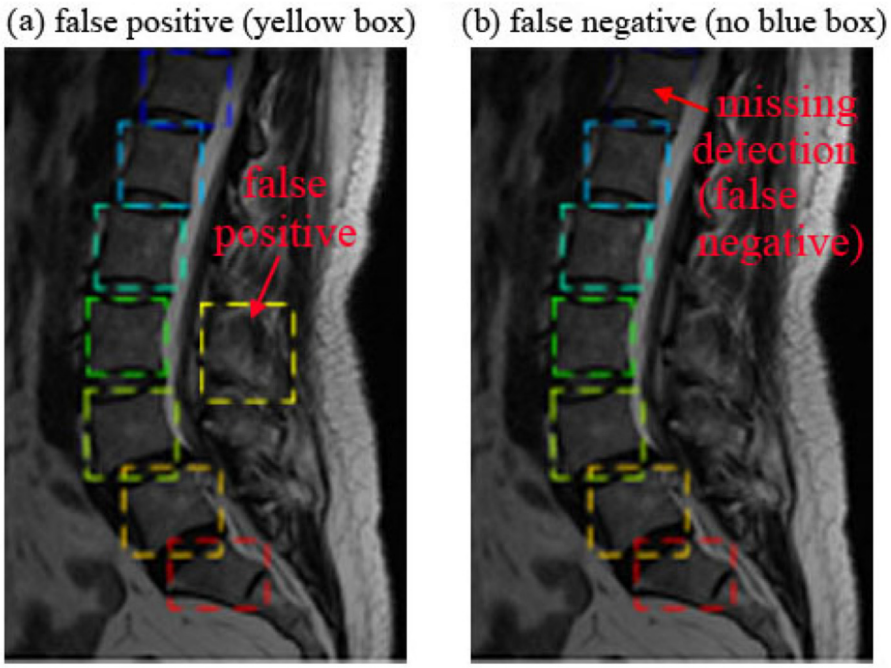

Fig. 5. Discriminating vertebrae with non-vertebrae regions helps message passing calibration. Fig. 5(a) and (b) shows two typical wrong cases where vertebrae with non-vertebrae regions are mistaken, i.e., false positives and false negatives (missing recognitions). Under this circumstance, the message passing is not able to calibrate because its input sequences are wrong. Our ERN prompts discrimination of vertebrae with non-vertebrae regions by setting the sparse codes of negative proposals to zero.

Thus, to summarize, the projection-based sparse codes improve discrimination of positive and negative proposals, which prompts the reliability of message passing and benefits the recognition task.

2.3. Self-adaptive reasoning diagnosis network (SRDN)

2.3.1. Overall workflow

The inputs of SRDN are the recognized vertebrae (i.e., the ERN results, which corresponds only to the positive proposals) and the hierarchical image features (these features are shared between ERN and SRDN). As mentioned in Section 2.1.3, each vertebra's diagnostic features $\mathbf{F}_i \in \mathbb{R}^{h\times w\times c}$ (Fig. 6(a1)) are firstly obtained by a procedure similar to ROI aligning. Then, the region attention module (Fig. 6(a2), Section 2.3.2) is designed to highlight the informative features and obtain $\mathbf{F}_i$' (Fig. 6(a3)). Next, $\mathbf{F}_i$' are flattened into diagnostic features $\mathbf{y}_i \in \mathbb{R}^M$ (Fig. 6(a3)) by the convolutional layers mentioned in Section 2.1.3. Afterwards, $\mathbf{y}_i$ are fed into the **feature interactor** (Fig. 6(a4), Section 2.3.3) to capture relation-aware information between recognized vertebrae via graph reasoning. The interacted features (denoted as $\mathbf{y}'_i$ Fig. 6(a5)) are then used to calculate diagnosis loss. Furthermore, in order to alleviate the over-smoothing problem of graph reasoning, a **self-adaptive reasoning controller** (Fig. 6(a4)/(b), Section 2.3.4) is designed to adjust the reasoning weights according to vertebrae diagnostic labels.

2.3.2. Region attention module

The region attention module (Fig. 6(a2)) is designed to highlight the most informative features for tumor diagnosis. It takes the cropped and resized features $\mathbf{F}_i$ as its input, and outputs the modulated diagnostic feature $\mathbf{F}'_i \in \mathbb{R}^{h\times w\times c}$ by the residual attention network shown in Eq. (3):

$$\mathbf{F}'_i = (1 + f(\mathbf{F}_i)) \otimes \mathbf{F}_i \tag{3}$$

Eq. (3) indicates that the region attention module is essentially a pixel-wise attention mask (weight matrix) for pixels inside the vertebrae bounding boxes, which is implemented in a residual manner. For more detail, in order to calculate the attention mask, we first design a residual unit (the residual unit contains three cascading convolutional layers of sizes $1 \times 1\times 256 \times 64$, $3 \times 3\times 64 \times 64$, and $1 \times 1\times 64 \times 256$, stride 1, each with "SAME" paddings and followed by a batch normalization layer and a ReLU activation layer; shortcut connections (He et al., 2016) is used to add up the input and output of these layers, i.e., the residual connection). Having defined the residual units, the attention masks $f(\mathbf{F}_i)$ is calculated by first feeding the input $\mathbf{F}_i$ into cascading "residual unit - residual unit - max-pooling layers - residual unit - max-pooling layers - residual unit - up-sampling (bilinear interpolation) - residual unit - up-sampling - batch normalization" layers; and then using two cascading convolutional layers (the first is of size $1 \times 1\times 256 \times 256$, stride 1, with "SAME" paddings and followed by a batch normalization layer and a ReLU activation layer, the second is of size $1 \times 1\times 256 \times 1$, stride 1, with "SAME" paddings and followed by a sigmoid activation layer) to convert the channel number to 1. Finally, the attention mask $f(\mathbf{F}_i)$ is element-wise multiplied to $\mathbf{F}_i$ (the symbol $\otimes$ in Eq. (3)) in a residual manner with shortcut connections (i.e., the $1 + f(\mathbf{F}_i)$ term in Eq. (3)). In this way, the region attention module highlights the most informative diagnostic features inside the vertebrae contour while keeping the performance to be no worse than the counterpart without attention.

2.3.3. Feature interactor

The feature interactor (Fig. 6(a4)/(b)) aims at allowing vertebrae with more distinguishing diagnostic features to assist in diagnosing the ones that are "hard" to classify. Our feature interactor constructs an undirected graph using diagnostic features $\mathbf{y}_i$ based on their feature similarity and then performs reasoning among the graph nodes (vertebrae features) using Eq. (4):

$$\mathbf{Y}' = \sigma(\mathbf{E}\psi(\mathbf{Y})\mathbf{W})$$
$$\text{where } \mathbf{E} = \text{norm}\left(\phi(\mathbf{Y})\phi(\mathbf{Y})^T\right) \tag{4}$$

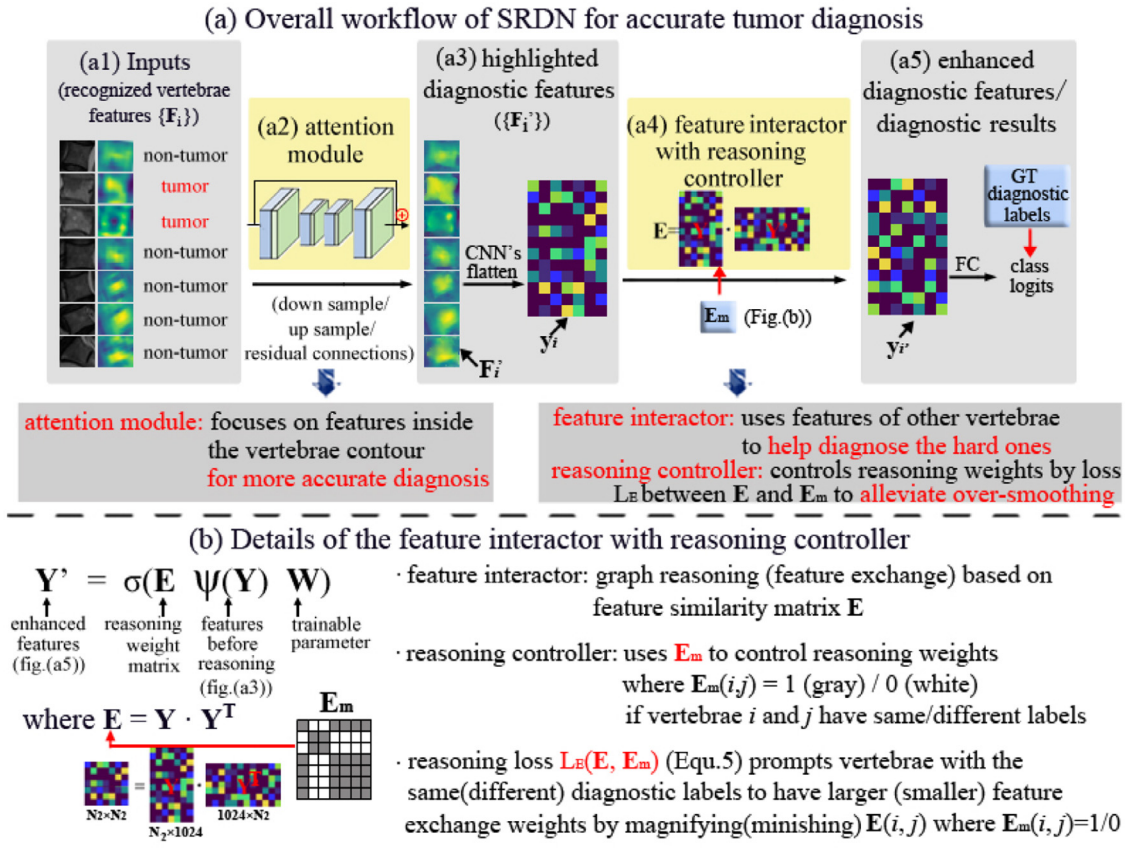

Fig. 6. The self-adaptive reasoning diagnosis network (SRDN). In SRDN, a region attention module is firstly designed to highlight the most informative features inside the recognized vertebrae for a more accurate diagnosis. Then, the vertebrae features are processed by a feature interactor with a self-adaptive reasoning controller. This allows diagnosing the "hard" vertebrae with the help of the features of the relatively "easier" ones while preventing the over-smoothing problem.

Eq. (4) demonstrates the principles of the feature interactor based on graph reasoning. For more details, $\mathbf{Y}$ is the diagnostic feature matrix (the ith column corresponds to $\mathbf{y}_i$), ϕ and ψ are operations adjusting the channel number (e.g., a 1×1 convolutional layer or a fully connected layer; in our work we simply use a 1×1 convolutional layer) with ReLU activation, $\mathbf{W}$ is a fully connected layer to adjust feature dimensions as in (Jiang et al., 2020), and the "norm" operation means column-wise normalization (i.e., $\text{norm}(\mathbf{A}) = \frac{\mathbf{A}}{\sum_k \mathbf{A}(k,:)}$) to force elements in $\mathbf{E}$ to be in range (0, 1]. Eq. (4) shows that feature similarity matrix $\mathbf{E}$ is calculated using pairwise inner product of every vertebrae feature, e.g., its element $E(i, j)$ is calculated by producing ith row of $\phi(\mathbf{Y})$ and the jth column of $\phi(\mathbf{Y}^T)$ element by element and then summing up the productions, which is the inner product of $\mathbf{y}_i$ and $\mathbf{y}_j$. Thus, if the two vertebrae i and j have similar features, then $\mathbf{E}(i, j)$ would be relatively larger. Then, feature interaction is performed by graph reasoning, i.e., calculating $\mathbf{E}\psi(\mathbf{Y})\mathbf{W}$. In this procedure, the resulting vertebrae (nodes) feature would be the weighted sum of its own feature and the other nodes', and the weights (edges) are the elements in a row in $\mathbf{E}$. The more similar the two nodes are, the stronger the corresponding edge (i.e., $\mathbf{E}(i, j)$) is, and the features of these two nodes (vertebrae) would contribute to each other with a higher weight. In this way, for a vertebra that is difficult to classify whether it is invaded by tumors, the other vertebrae's features can be used for assistance (more detailed discussions will be conducted in Section 2.3.5).

2.3.4. Self-adaptive reasoning controller

The self-adaptive reasoning controller is designed for preventing the over-smoothing issue of graph reasoning. After feature interaction by the similarity matrix $\mathbf{E}$, the enhanced diagnostic features $\mathbf{Y}'$ is fused into $\mathbf{Y}$ by lateral concatenations. Then, the fused features $[\mathbf{Y},\ \mathbf{Y}']$ are fed into fully connected layers to calculate the diagnostic logits $\mathbf{c}$, which are supervised by the ground truth diagnostic labels c^* for minimizing the cross-entropy loss $L_c = L_{ce}(\mathbf{c}, c^*)$. However, as analyzed above, we desire vertebrae with similar features to have higher interaction weights, a natural idea to achieve this is to design a strategy to control feature interaction by promoting vertebrae with the same label to have similar features. Thus, we design a self-adaptive reasoning controller to formulate this strategy as a loss function:

$$\begin{aligned} L_E &= exp(-(E_{same} - E_{different})) \\ &= exp[-(\sum_{i,j}(\mathbf{E_m} \otimes \mathbf{E} - (\mathbf{1} - \mathbf{E_m}) \otimes \mathbf{E}))] \end{aligned} \tag{5}$$

Eq. (5) designs a loss term to control the feature similarity matrix $\mathbf{E}$, i.e., it prompts vertebrae with the same (different) diagnostic labels to have larger (smaller) feature exchange weights to control graph reasoning. For more details, $\mathbf{E_m}$ is a 0–1 matrix, $\mathbf{1}$ is an all 1 matrix, both $\mathbf{E_m}$ and $\mathbf{1}$ have the same shape with $\mathbf{E}$. $\otimes$ means element-wise production. The $\sum_{i,j}$ operation in Eq. (5) means summing up all elements in the matrix, thus, the first term in Eq. (5) means summing up the elements in $\mathbf{E}$ where corresponding $\mathbf{E_m}$ equals 1, i.e., E_{same} would be the sum of reasoning weights of vertebrae with the same diagnostic labels. In this way, the loss term $-(E_{same} - E_{different})$ encourages positions in $\mathbf{E}$ corresponding to vertebrae with the same labels to be larger (i.e., vertebrae with the same diagnostic labels to have larger feature exchange weights as shown in Fig. 6(b)). To guarantee the loss is larger than 0, an exp expression is used. Furthermore, since the

network has no *a priori* idea about how many vertebrae are recognized, the self-adaptive reasoning loss is only valid when the training of the recognition network comes to stability (i.e., after ∼20000 steps). Also, if all the vertebrae in an image have the same diagnostic labels (i.e., all of them are invaded by tumors or vise versa), this loss is set to 0 to guarantee the validity of training. The self-adaptive reasoning loss L_E is minimized together with the diagnose loss L_c, i.e., the total loss of the diagnosis task is:

$$L_d = L_c + L_E = \frac{\lambda_4}{N_4}\sum_{i=1}^{N_4} L_{ce}(\mathbf{c}_i, c_i^*) + \lambda_5 exp(-(E_{same} - E_{different})) \quad (6)$$

Eq. (6) means the loss of the diagnosis task contains two parts, i.e., the diagnostic classification loss and the self-adaptive reasoning loss. For more details: (1) L_c is the cross-entropy loss of the predictions $\mathbf{c}_i$ and the diagnostic ground truth c_i^*, N_4 is the number of the recognized vertebrae. (2) The weight λ_5 is set to restrict L_E to be no larger than $0.5L_c$ as in our previous work (Zhao et al., 2019b). All losses are minimized using Momentum Optimizer.

2.3.5. Discussions

The reason that SRDN benefits diagnosis performance is also twofold.

(1) The attentional mechanism highlights the most informative features for the diagnosis task. Based on the recognized bounding boxes, the attentional mechanism further highlights the informative information (e.g., tumor-like patterns inside the vertebrae contour) while suppresses useless features (e.g., intervertebral disc portions inside the recognized boxes, which may be mistaken as tumors by the diagnosis network) by learning an amplification factor $f(\mathbf{F}_i)$ and modifying each element in $\mathbf{F}_i$ in a residual manner. For more details, the attention module in Section 2.3.2 is a "down-sample - up-sample" architecture (corresponding to the max-pooling layers and the bilinear interpolation layers); during the down-sampling steps, the network can increase the receptive field and collect global diagnostic features of the whole vertebrae; after reaching the lowest resolution, the up-sampling steps retrieve the input resolution and combine the global diagnostic features with the original fine-grained diagnostic features by lateral connections (Wang et al., 2017a). The final attention masks $f(\mathbf{F}_i)$ is a matrix whose size is the same with the input features; $f(\mathbf{F}_i)$'s values are in range 0∼1 because sigmoid activation is used as the last layer. Thus, $f(\mathbf{F}_i)$ plays the role of feature selection, i.e., when the value of a pixel in $f(\mathbf{F}_i)$ approaches 1, then the feature of this pixel is highlighted; while when that of $f(\mathbf{F}_i)$ approaches 0, then the feature of this pixel is diminished. The residual connection can avoid feature value degrading in deeper networks. The attention mask $f(\mathbf{F}_i)$ is automatically learned during training; in our work as well as other object recognition tasks, if the object mask is available, it can be used for supervising the attention mask for improved attentional performance. Similar approaches have also been leveraged in (Wang et al., 2017a; Pang et al., 2019) for feature selection in other medical image analysis tasks.

(2) The feature interactor with self-adaptive reasoning control strategy can find meaningful relational clues from different vertebrae for tumor diagnosis, i.e., for a "hard" vertebra that is difficult to diagnose, features of other "easier" vertebrae could be used for assistance. Intuitively, if an oncologist finds it difficult to tell tumors from similar-appearing disease (e.g., end-plate osteochondritis) for the "hard" vertebra, but he finds some other "easier" vertebrae in the input image are obviously invaded by tumors, then he would infer that the current vertebra is very likely to be also invaded by tumors, i.e., the oncologist uses the feature clues of the "easier" vertebrae to diagnose the "hard" one based on some clinical knowledge. This triggers the thought to design our feature interactor (Section 2.3.3) to mimic this process, i.e., if the "hard" vertebra's features are similar to those of some "easier" ones, then the features of the "easier" ones are added to those of the "hard" ones in a graph reasoning manner shown in Eq. (4). In this way, the resulting node (vertebrae) features are the weighted sum of its own feature and the other vertebrae's, i.e., the relational clues contained in the features of the "easier" vertebrae may be used to assist in diagnosing the "hard" ones based on their feature similarity. However, naively performing graph reasoning among vertebrae features can also result in over-smoothing (i.e., diagnostic features of all vertebrae approach their average), which may on the contrary decrease diagnostic discrimination. To deal with this problem, we propose a self-adaptive reasoning controller to prompt/suppress the feature exchange of vertebrae with the same/different diagnostic label, i.e., the exchange of vertebrae features is supervised by the diagnostic labels. In more details, during training, the reasoning controller (Section 2.3.4) achieves this by increasing/decreasing the feature exchange weight among vertebrae with the same/different diagnostic labels, i.e., the $E(i,j)$ will be trained to be larger if the vertebrae i and j have the same diagnostic label and vice versa. Thus, if the diagnostic label of the "hard" vertebra and those of the "easier" vertebrae are the same, the features of the "easy" vertebrae would be added to the "hard" vertebrae with a larger weight, and these vertebrae would be prompted to have similar diagnostic features. On the contrary, graph reasoning is discouraged among vertebrae with different diagnostic labels. This hinders the feature exchange among vertebrae with different diagnostic labels, which alleviates the over-smoothing problem. In this way, the features of the "hard" vertebrae would approach those of the "easy" ones with the same diagnostic labels by controlled graph reasoning, i.e., the diagnostic features of the "easy" vertebrae are used to assist in diagnosing the "hard" ones. Thus, to summarize, the feature interactor with self-adaptive reasoning controller prompts the graph reasoning procedure to be more reasonable and benefits the diagnosis task.

3. Experiments and discussions

Dataset, implementations, ground truth annotations, and evaluation metrics. RE-DECIDE has been intensively evaluated using a dataset containing 600 challenging spinal MRI images of ∼163 patients. The dataset contains arbitrary MRI images of thoracic, lumbar, and sacrum vertebrae of 6 different FOVs. Our data has been approved by the Research Ethics Board of Western University (REBID: 17656E). For each patient, 3∼4 slices where all existing vertebrae are not severely distorted are chosen from 3D scans and resized to 512 × 512. The training/testing dataset preparation is similar to our previous work, i.e., we use the standard five-fold cross-validation for evaluation. For more details: (1) To evaluate the recognition performance of RE-DECIDE, we construct our training/testing datasets using MRI images of 6 different FOV's as in our previous work (Zhao et al., 2020; 2021). The number of images of each FOV is kept approximately the same in the training/testing dataset in each fold to provide sufficient training data for each FOV. For each FOV, the training and testing images of each fold are randomly selected. (2) To evaluate the diagnostic performance of RE-DECIDE, we construct our training/testing datasets using 4600 vertebrae (among which 818 of them are invaded by different types of tumors) to mimic the scenario encountered in clinical practice where tumors show a large appearance variety. The vertebrae invaded/not invaded by tumors are completely randomly split into the training and testing set.

As in our previous work (Zhao et al., 2019b; 2020; 2021), RE-DECIDE is implemented in Python 3.6 on Tensorflow 1.13.0. For the hyper-parameters that exist in the previous work (such as the batch size, the initial learning rate, the learning rate decay fac-

tor, the learning momentum, and the weights λ_1 and λ_2), they are kept the same as our previous work. For the hyper-parameters that are new in the current work (such as the weights λ_3 and λ_5), as mentioned in Sections 2.2.4 and 2.3.4, we follow the strategy in (Zhao et al., 2019b) to force the additional losses, e.g., the self-adaptive reasoning loss, to be no larger than half of the main recognition/diagnostic losses. We do not carry our fine-tuning on the hyper-parameters in our work, however, our RE-DECIDE framework proves to work well with these hyper-parameters.

An experienced oncologist for spinal tumors has carefully labeled all vertebrae invaded by tumors twice with a temporal interval of one month. The second annotation is blinded to his initial annotation, which is used to assess intra-operator variability. We define the vertebrae invaded by tumors as a "positive" diagnostic label; whereas the vertebrae not invaded by tumors as a "negative" diagnostic label. If the two annotations have the same labels, the ground truth label is decided to be the manual label; if the two annotations have different labels (one indicates tumor and the other not), the vertebra is assigned a "positive" label (invaded by tumors) for training. Also, for the vertebrae that the oncologist suspects (even if he is not that sure) to be invaded by tumors, they are given the positive label (i.e., labeled as tumor). This follows the clinical practice that mistaking a tumor as non-tumor is more severe than mistaking a non-tumor as tumor. In our dataset, the oncologist feels uncertain about 15% of the vertebrae, which indicates our dataset is very challenging. All implementation details are the same as our previous work (please refer to Section 3.1 of Zhao et al., 2021).

The evaluation metrics are the same with the conference version, i.e., we use standard five-fold cross-validation on seven frequently used metrics (image recognition accuracy (IRA), identification rate (IDR), and mAP_{75} for recognition task; and ROC curve, area under curve (AUC), accuracy, precision, and recall for diagnosis task). In more details: (1) For recognition, IRA is defined as the percentage of images with all its vertebrae correctly detected, i.e., the ratio correctly recognized images/all images; IDR is defined as the accuracy of the individual vertebra classification, i.e., the percentage of individual vertebrae that have been correctly detected; mAP_{75} is a comprehensive metric that considers the precision and recall of the recognition task (not the same as the precision and recall in the diagnosis task, which will be defined below), as well as the IoU (Intersection-over-union) of the predicted bounding box with the ground truth boxes. More detailed explanations of these three metrics can be seen in Section 3.2.2 of (Zhao et al., 2021). (2) For diagnosis (machine learning classification), accuracy means the ratio of correctly classified vertebrae to the total vertebrae; precision is the ratio of correctly classified "positive" vertebrae (i.e., vertebrae invaded by tumors) to the total classified "positive" vertebrae; recall is the ratio of correctly classified "positive" vertebrae to all "positive" vertebrae; ROC curve is a curve considering the false positive rate and true positive rate (recall); AUC is the area under the ROC curve. These five metrics have been widely used in classification tasks (Pedregosa et al., 2011).

Qualitative and quantitative demonstrations for recognition and diagnosis tasks. Experimental results show RE-DECIDE achieves satisfactory performance for both recognition and diagnosis tasks.

(1) For the recognition task, Fig. 7(a) qualitatively shows that RE-DECIDE can recognize vertebrae of different categories despite the image FOV, characteristics, and vertebrae appearance variety. The recognized vertebrae bounding boxes (dashed) overlap well with the ground truth boxes (solid) of the correct labels (colors). For quantitative evaluation, the black, red, and blue bars in Fig. 7(b) show high IRA (overall: 0.940±0.023, individual FOV's: >0.9), IDR (overall: 0.955±0.016, individual FOV's: >0.9), and mAP_{75} (overall: 0.949±0.021, individual FOV's: >0.927), which means that the recognition work produces very few wrongly classified, missing, or false positive recognitions. Even for the most difficult FOV (L4∼T10, which is prone to be confused with L5∼T11 FOV), the recognition IRA and IDR still reaches 0.9 (although relatively lower than the other FOVs).

(2) For the diagnosis task, Fig. 7(a) also qualitatively demonstrates that RE-DECIDE provides diagnostic predictions that (PRED) is generally the same as the diagnostic ground truth (GT). For quantitative evaluation, Fig. 7(c) shows that our work achieves a satisfactory tumor diagnosis ROC curve with an AUC of 0.947. If we regard vertebrae with predicted tumor probability greater than 0.5 as "positive" (suffering from tumors), we get a diagnostic accuracy of 0.931, precision of 0.837, and recall of 0.760. The accuracy is high, which means that our classifier in general achieves good performance. The recall is relatively lower, which may be due to the data imbalance in the diagnosis task (the "positive" vertebrae are far less than the "negative" vertebrae) of our dataset; also, there are some "hard" vertebrae that even the two annotations of the oncologist are not the same (which may be difficult for the diagnose network to classify and therefore affect the precision and recall). Nevertheless, our RE-DECIDE still achieves better performance than the compared methods (which will be demonstrated below); also, these results are comparable with those of the oncologist (rows 1 and 7 in Table 1).

There are some similarities between the oncologist's diagnosis and the classification procedure of RE-DECIDE. For some "hard" vertebrae such as the L3 of the left bottom figure in Fig. 7(a), the positive diagnostic label (i.e., labeled as invaded by tumors) is determined based on our labeling criteria (i.e., for the vertebrae that the oncologist suspects to be invaded by tumors, they are given the positive label even if he is not that sure). Actually, the oncologist claims he is actually not that sure about whether it is invaded by tumors based on its appearance. However, the diagnoses of other vertebrae can be used for help, e.g., the oncologist prefers the L3 to be invaded because the other vertebrae are invaded. In more detail, there is an insignificant ring-like structure (pointed out by the red arrow in the enlarged figure (Fig. 7(d))); also, the left upper corner of the vertebrae seems to be "eroded" and shows a blurry edge (pointed out by the orange arrow in Fig. 7(d)). These may be the hints (although these hints may not be strong enough for diagnosis) that "L3 is invaded by tumors"; based on these hints as well as the fact that some other vertebrae (e.g., L1, T10) are obviously invaded, the oncologist, at last, prefers L3 to be invaded by tumor. Our SRDN may be able to use the self-adaptive graph reasoning strategy to mimic the oncologists diagnosis procedure. In SRDN, the feature interactor with self-adaptive reasoning controller constructs the feature similarity matrix $\mathbf{E}$ to guide feature exchange between different vertebrae. Each element $E(i, j)$ means the feature exchange weight of vertebrae i and j. Via network training, $E(i, j)$ would be larger if vertebrae i and j have the same diagnostic labels. This strategy prompts the vertebrae with the same diagnostic labels to have more similar features; it also allows these vertebrae to exchange features with relatively large weights. In this way, even if a "hard" vertebra does not have a very significant visual appearance indicating tumors (or non-tumors), its latent features can be guided to approach the vertebrae whose features are significant. For example, for the "hard" L3 vertebra which is difficult to diagnose based on its visual appearance, the features of the other vertebrae that are invaded by tumors would be added to its feature, which causes the probability that the L3 is invaded by tumors to increase. In this way, the proposed network to some extent mimics the oncologist's diagnostic procedure where other vertebrae's features are used to assist diagnosis.

Intra-comparison experiments. We first respectively remove ERN and SRDN for ablation experiments. Then, to further explore

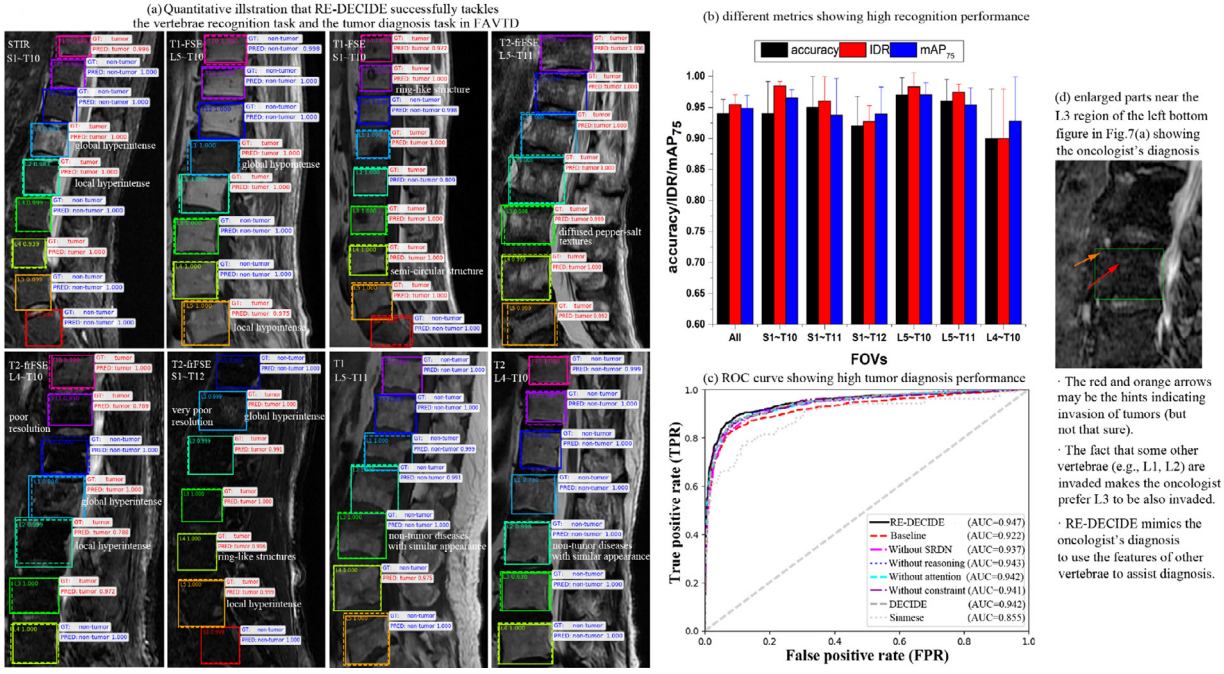

Fig. 7. Different metrics showing the effectiveness of our work. (a) Our network can accurately recognize vertebrae from images of different FOV and characteristics; it can also distinguish tumors from non-tumor diseases of various appearances. (b) Different metrics demonstrating high recognition performance. (c) The ROC curve and AUC value showing high diagnosis performance. (d) Enlarged parts near the L3 region of the left bottom figure in Fig. 7(a) showing the oncologist's diagnosis procedure.

Table 1
Superiority of RE-DECIDE to state-of-the-art methods and the ablation experiments. The first three columns (IRA, IDR, and mAP_{75}) are used to evaluate the recognition performance; meanwhile, the last four columns, (AUC, accuracy, precision, and recall) are used to evaluate the diagnosis performance. The results show our RE-DECIDE achieves high recognition performance with the help of the projection-based sparse codes and the enhanced supervision strategy in ERN; RE-DECIDE also achieves high diagnosis performance with the help of the region attention module and feature interactor with self-adaptive reasoning controller in SRDN.

Method	IRA	IDR	mAP_{75}	AUC	accuracy	precision	recall
RE-DECIDE (Our work)	0.940±0.023	0.955±0.016	0.949±0.021	0.947±0.069	0.931±0.002	0.837±0.042	0.760±0.055
baseline	0.908±0.028	0.930±0.034	0.922±0.031	0.922±0.078	0.921±0.005	0.802±0.049	0.741±0.074
Without SRDN	0.937±0.025	0.954±0.022	0.946±0.032	0.937±0.065	0.922±0.008	0.790±0.044	0.762±0.035
Without reasoning in SRDN	0.939±0.032	0.956±0.025	0.944±0.030	0.943±0.064	0.929±0.009	0.830±0.023	0.756±0.054
Without attention in SRDN	0.932±0.028	0.951±0.026	0.942±0.031	0.942±0.070	0.925±0.011	0.826±0.054	0.734±0.042
Without reasoning constraint in SRDN	0.942±0.032	0.957±0.029	0.945±0.034	0.941±0.073	0.928±0.008	0.828±0.016	0.759±0.057
intra-observer	–	–	–	–	0.964	0.874	0.838
DECIDE (Zhao et al., 2020)	0.936±0.028	0.954±0.019	0.947±0.014	0.942±0.068	0.926±0.082	0.821±0.028	0.743±0.039
Siamese (Wang et al., 2017b)	–	–	–	0.855±0.121	0.884±0.089	0.774±0.197	0.623±0.225
Hi-scene (Zhao et al., 2019b)	0.878±0.048	0.930±0.053	0.923±0.039	–	–	–	–
DI2IN (Yang et al., 2017)	0.803±0.149	0.904±0.115	–	–	–	–	–
Faster-RCNN (Ren et al., 2015)	0.750±0.138	0.869±0.104	0.848±0.146	–	–	–	–
SSD (Liu et al., 2016)	0.725±0.168	0.814±0.187	0.789±0.208	–	–	–	–
YOLO-v3 (Yang and Deng, 2020)	0.933±0.045	0.950±0.021	0.944±0.017	–	–	–	–

the effects of the region attention module, feature interactor, and self-adaptive reasoning controller in SRDN, we respectively remove these components (ERN is enabled in these experiments) and compare the diagnostic performance to answer two interesting questions: (1)*How can the regional attention module help tumor diagnosis?* (2) *Under which circumstance can graph reasoning help the diagnosis task?*

For ablation experiments, as shown in Table 1: (1) If both ERN and SRDN are removed (row 2 in Table 1), the baseline recognition-diagnosis framework achieves an acceptable performance for both recognition (IRA: 0.908±0.028, IDR: 0.930±0.034, mAP_{75}: 0.922±0.031) and diagnosis (AUC: 0.922±0.078, accuracy: 0.921±0.005, precision: 0.802±0.049, recall: 0.741±0.074) tasks. This performance shows that the baseline framework is capable of distinguishing different vertebrae and telling those invaded by tumors from those not despite the MRI image characteristic variety, i.e., it has the potential to be used for FAVTD. Also, the benefit of ERN and SRDN for recognition and diagnosis performances are demonstrated. (2) When ERN is enabled, the recognition performance shows an increase of IRA: ∼1.8%, IDR: ∼0.9%, mAP_{75}: ∼2.4% (columns 1∼3 of rows 1 and 2 in Table 1). This demonstrates the projection-based sparse codes and the enhanced supervision strategy in ERN help tackle the recognition challenges. (3) When SRDN is enabled, the diagnosis performance increases by AUC: ∼2.5%, accuracy: ∼1.0%, precision: ∼3.5%, recall: ∼1.9% (columns 4∼7 of rows 1 and 2 in Table 1). This demonstrates in general SRDN benefits tumor diagnosis by exchanging features between vertebrae; compared with the baseline method, SRDN can on average correctly diagnose ∼9.2 more vertebrae out of the 920 vertebrae in the testing dataset.

To further answer the two above-mentioned questions mentioning the components in SRDN, we carry out four experiments shown in columns 3∼6 in Table 1. In these experiments, we: (1) disable both regional attention module and feature interactor in

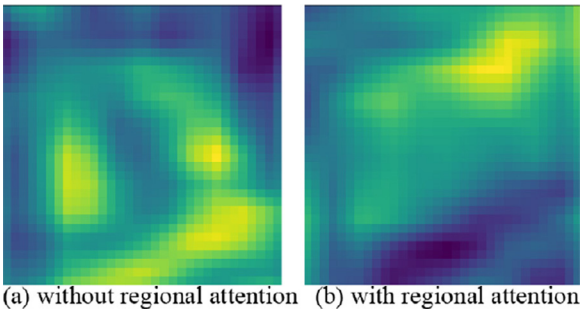

Fig. 8. The regional attention module highlights the most informative parts for diagnosis (Fig. 8(b)). As a contrast, locations outside the vertebrae could be wrongly highlighted without regional attention (Fig. 8(a)).

SRDN (row 3, this case is actually "baseline+ERN'); (2) respectively disable the feature interactor and regional attention module (rows 4 and 5); (3) disable the self-adaptive reasoning controller while enable feature interacting (row 6). The results are analyzed below:

(1) By comparing the diagnostic results in rows 1, 2, 3, and 5, it is shown that the attention mechanism can in general improve diagnosis performance (row 1 VS rows 2/3/5). This first illustrates the effectiveness of the regional attention module, which helps focus on the informative features inside the vertebrae contours. This is also shown in Fig. 8, which shows that without the attention module, the network may wrongly highlight regions outside the vertebrae (Fig. 8(a)) and give a wrong diagnosis. Nevertheless, the attentional mechanism solves this problem (Fig. 8(b)).

(2) By conducting experiments in rows 1, 4 (where the feature interactor based on graph reasoning is disabled), and 6 (where the feature interactor is enabled, but the self-adaptive reasoning controller is disabled), we further explore the effect of graph reasoning on diagnosis performance in condition that the attention module is enabled. To our surprise, the diagnostic results in row 6 are worse than those in row 4. This indicates that, although existing literature (Wang et al., 2017b; Chen et al., 2019b) claim that graph reasoning can prompt image classification performance, we find that using graph reasoning without constraints may also harm classification, which is probably due to the over-smoothing problem (actually, in the experiment in row 6, we found that features of all vertebrae after reasoning are almost the same). However, row 1 gives a better performance than rows 4 and 6, which means that guiding graph reasoning with weight constraints may produce better results. Thus, adding constraints to graph reasoning (which can be implemented as losses in CNN-based architectures) provides a heuristic idea for overcoming the over-fitting problem in image classification besides the re-weighted scheme based on manually selected weights used in (Chen et al., 2019b).

Another interesting finding is that although SRDN generally improves diagnosis performance, row 3 (without SRDN) shows relatively high recall. This may be caused by the data near the decision boundary of the diagnosis task, which accounts for quite a few proportions in our relatively challenging dataset. As the training goes on, the boundary is continually moving. If it moves towards where 'the network tends to classify vertebrae as invaded by tumors', the results may contain more true positives but less true negatives (i.e., higher recall but lower precision) and vice versa. In the case in row 3, the decision boundary moves to such a location that tends to produce positive diagnoses; as a result, the recall is high at the expense of relatively lower accuracy and precision. In contrast, SRDN prompts the boundary to move towards locations where more vertebrae are correctly classified, i.e., SRDN is beneficial to diagnosis and yields higher accuracy, precision, and AUC.

Inter-comparison experiments. We compare our work with state-of-the-art works performing recognition (Zhao et al., 2019b; Yang et al., 2017; Ren et al., 2015) or diagnosis (Wang et al., 2017b) task. As shown in rows 1, 8∼12 in Table 1, RE-DECIDE outperforms all compared methods as well as the preliminary version DECIDE in both recognition (first three columns) and diagnosis (last four columns) tasks. (1) For recognition, IRA, IDR, and mAP_{75} all benefit from the enhanced supervision provided by the embedded dictionary. For all different FOV's (including the most difficult FOV T10∼L4), the recognition performance is higher than 90% (Fig. 3(b)). Furthermore, by comparing with (Yang et al., 2017) that uses dictionary learning as post-processing for landmark refinement, our embedded dictionary is more beneficial because it exploits the projections and introduces the ensemble of multiple predictions to improve vertebrae recognition discrimination. The advantage of ERN is also shown by comparing RE-DECIDE with Hi-scene (Zhao et al., 2019b); ERN provides more discrimination and shows its robustness against the vertebrae appearance changes brought about by the tumors. (2) For diagnosis, the superiority of RE-DECIDE to the ablation experiment (Wang et al., 2017b) shows that on one hand, SRDN improves the diagnostic discrimination; on the other hand, the fully automatic end-to-end recognition-diagnosis framework benefits tumor diagnosis. This framework not only eliminates the tedious manual vertebrae extraction but also shares the rich hierarchical features to the diagnosis network. In this way, the workflow of our RE-DECIDE reinforces mutual benefits between two tasks and improves tumor diagnosis performance.

Limitations. We list the limitations of our work for future research. (1) Our SDRN tries to mimic the oncologists to use other vertebrae's diagnostic features to assist in diagnosing the "hard" ones. However, this assistance is still not exactly the same as the oncologists' diagnosis procedure, e.g., the oncologists decide whether to use other vertebrae's diagnostic features to assist diagnosis based on their rich clinical experience; while SRDN decides this based on the weight matrix $\mathbf{E}$ calculated by the inner production of vertebrae features extracted by the antecedent CNNs. This, in some rare cases, may lead to the propagation of wrong features to other vertebrae via graph reasoning, i.e., if the antecedent CNNs provide correct features, the propagation of these features can enhance performance, however, if wrong features are provided, the features of vertebrae of different diagnostic labels may also be relatively similar, which may also decrease the diagnosis performance due to the feature exchange of their vertebrae. This may happen in the test phase of some splits in our experiments. Other graph reasoning strategies, as well as other specialized methods such as metric learning, could further improve the diagnosis performance. Also, our Re-DECIDE is not intended for tumors farther from the vertebrae bodies. Moreover, our RE-DECIDE concerns giving a binary prediction for whether a vertebra is invaded by tumors; it does not discuss the suspected case where the oncologists prefer not to make a determinate diagnosis. (2) The mutual effects between the recognition task and the diagnosis task may bring difficulties to the interpretability of FAVTD. Although, as mentioned above, the mutual effects are beneficial to both tasks, we observe that changes in SRDN may also affect the recognition and vice versa. For example, the recognition performance in rows 1, 3, 4, 5, 6, and 8 should be approximately the same, however, it still shows a change of 1% in IRA. Similar issues also happen in the diagnosis task. De-coupling the training of the two tasks may better clarify the analysis and enhance interpretability.

4. Conclusion

In this paper, we have designed a reasoning discriminative dictionary-embedded network (RE-DECIDE) as a novel computer aided diagnosis (CAD) tool for fully automatic vertebrae tumor diagnosis (FAVTD) from MRI images. RE-DECIDE contains two novel designs: (1) ERN uses feed-forward dictionary learning and

projection-based enhanced supervision strategy to obtain discriminative representations for vertebrae recognition and tackle the FOV/characteristics challenges; (2) SRDN uses attentional module and self-adaptive graph reasoning strategy for tumor diagnosis and alleviates the tumor appearance variability challenges. The effectiveness of RE-DECIDE, as well as its advantage to the state-of-the-art, have been demonstrated by extensive experiments. Readers are welcome to ask for the codes used in this work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aharon, M., Elad, M., Bruckstein, A., 2006. K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process. 54 (11), 4311–4322.

Burns, J.E., Yao, J., Wiese, T.S., Muñoz, H.E., Jones, E.C., Summers, R.M., 2013. Automated detection of sclerotic metastases in the thoracolumbar spine at CT. Radiology 268 (1), 69–78.

Chen, H., Shen, C., Qin, J., Ni, D., Shi, L., Cheng, J.C.Y., Heng, P.-A., 2015. Automatic localization and identification of vertebrae in spine CT via a joint learning model with deep neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 515–522.

Chen, Y., Rohrbach, M., Yan, Z., Shuicheng, Y., Feng, J., Kalantidis, Y., 2019. Graph-based global reasoning networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 433–442.

Chen, Z.-M., Wei, X.-S., Wang, P., Guo, Y., 2019. Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5177–5186.

Chmelik, J., Jakubicek, R., Walek, P., Jan, J., Ourednicek, P., Lambert, L., Amadori, E., Gavelli, G., 2018. Deep convolutional neural network-based segmentation and classification of difficult to define metastatic spinal lesions in 3D CT data. Med. Image Anal. 49, 76–88.

Coates, A., Ng, A.Y., 2011. The importance of encoding versus training with sparse coding and vector quantization. ICML.

Gao, Z., Chung, J., Abdelrazek, M., Leung, S., Hau, W.K., Xian, Z., Zhang, H., Li, S., 2019. Privileged modality distillation for vessel border detection in intracoronary imaging. IEEE Trans. Med. Imaging 39 (5), 1524–1534.

Glocker, B., Zikic, D., Konukoglu, E., Haynor, D.R., Criminisi, A., 2013. Vertebrae localization in pathological spine CT via dense classification from sparse annotations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 262–270.

Gregor, K., LeCun, Y., 2010. Learning fast approximations of sparse coding. In: Proceedings of the 27th International Conference on International Conference on Machine Learning, pp. 399–406.

Guo, S., Xu, L., Feng, C., Xiong, H., Gao, Z., Zhang, H., 2021. Multi-level semantic adaptation for few-shot segmentation on cardiac image sequences. Med. Image Anal. 73, 102170.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Jiang, C., Wang, S., Liang, X., Xu, H., Xiao, N., 2020. ElixirNet: relation-aware network architecture adaptation for medical lesion detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, pp. 11093–11100.

Jiang, Z., Lin, Z., Davis, L.S., 2013. Label consistent k-SVD: learning a discriminative dictionary for recognition. IEEE Trans. Pattern Anal. Mach. Intell. 35 (11), 2651–2664.

Liang, Z., Yang, M., Deng, L., Wang, C., Wang, B., 2019. Hierarchical depthwise graph convolutional neural network for 3D semantic segmentation of point clouds. In: 2019 International Conference on Robotics and Automation (ICRA). IEEE, pp. 8152–8158.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C., 2016. SSD: single shot multibox detector. In: European conference on computer vision. Springer, pp. 21–37.

Liu, Y., Chen, Q., Chen, W., Wassell, I., 2018. Dictionary learning inspired deep network for scene recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32.

Lootus, M., Kadir, T., Zisserman, A., 2014. Vertebrae detection and labelling in lumbar MR images. In: Computational Methods and Clinical Applications for Spine Imaging. Springer, pp. 219–230.

Makary, M.A., Sexton, J.B., Freischlag, J.A., Millman, E.A., Pryor, D., Holzmueller, C., Pronovost, P.J., 2006. Patient safety in surgery. Ann. Surg. 243 (5), 628.

Mundy, G.R., 2002. Metastasis to bone: causes, consequences and therapeutic opportunities. Nat. Rev. Cancer 2 (8), 584–593.

Pang, S., Su, Z., Leung, S., Nachum, I.B., Chen, B., Feng, Q., Li, S., 2019. Direct automated quantitative measurement of spine by cascade amplifier regression network with manifold regularization. Med. Image Anal. 55, 103–115.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Quan, Y., Xu, Y., Sun, Y., Huang, Y., Ji, H., 2016. Sparse coding for classification via discrimination ensemble. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5839–5847.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28, 91–99.

Shah, L.M., Salzman, K.L., 2011. Imaging of spinal metastatic disease. Int. J. Surg. Oncol. 2011.

Soffer, S., Ben-Cohen, A., Shimon, O., Amitai, M.M., Greenspan, H., Klang, E., 2019. Convolutional neural networks for radiologic images: a radiologists guide. Radiology 290 (3), 590–606.

Sun, X., Nasrabadi, N.M., Tran, T.D., 2019. Supervised deep sparse coding networks for image classification. IEEE Trans. Image Process. 29, 405–418.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X., 2017. Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164.

Wang, J., Fang, Z., Lang, N., Yuan, H., Su, M.-Y., Baldi, P., 2017. A multi-resolution approach for spinal metastasis detection using deep siamese neural networks. Comput. Biol. Med. 84, 137–146.

Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803.

Weilbaecher, K.N., Guise, T.A., McCauley, L.K., 2011. Cancer to bone: a fatal attraction. Nat. Rev. Cancer 11 (6), 411–425.

Wiese, T., Burns, J., Yao, J., Summers, R.M., 2011. Computer-aided detection of sclerotic bone metastases in the spine using watershed algorithm and support vector machines. In: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. IEEE, pp. 152–155.

Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A., 2020. A convolutional approach to vertebrae detection and labelling in whole spine MRI. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 712–722.

Xie, H., Li, J., Xue, H., 2017. A survey of dimensionality reduction techniques based on random projection. arXiv preprint arXiv:1706.04371.

Xue, Y., Bigras, G., Hugh, J., Ray, N., 2019. Training convolutional neural networks and compressed sensing end-to-end for microscopy cell detection. IEEE Trans. Med. Imaging 38 (11), 2632–2641.

Yang, D., Xiong, T., Xu, D., Huang, Q., Liu, D., Zhou, S.K., Xu, Z., Park, J., Chen, M., Tran, T.D., et al., 2017. Automatic vertebra labeling in large-scale 3D CT using deep image-to-image network with message passing and sparsity regularization. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 633–644.

Yang, Y., Deng, H., 2020. GC-YOLOv3: you only look once with global context block. Electronics 9 (8), 1235.

Zhang, Y., He, X., Tian, Z., Jeong, J.J., Lei, Y., Wang, T., Zeng, Q., Jani, A.B., Curran, W.J., Patel, P., et al., 2020. Multi-needle detection in 3D ultrasound images using unsupervised order-graph regularized sparse dictionary learning. IEEE Trans. Med. Imaging 39 (7), 2302–2315.

Zhao, S., Chen, B., Chang, H., Wu, X., Li, S., 2020. Discriminative dictionary-embedded network for comprehensive vertebrae tumor diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 691–701.

Zhao, S., Gao, Z., Zhang, H., Xie, Y., Luo, J., Ghista, D., Wei, Z., Bi, X., Xiong, H., Xu, C., et al., 2017. Robust segmentation of intima–media borders with different morphologies and dynamics during the cardiac cycle. IEEE J. Biomed. Health Inform. 22 (5), 1571–1582.

Zhao, S., Wu, X., Chen, B., Li, S., 2019. Automatic spondylolisthesis grading from MRIs across modalities using faster adversarial recognition network. Med. Image Anal. 58, 101533.

Zhao, S., Wu, X., Chen, B., Li, S., 2019. Automatic vertebrae recognition from arbitrary spine MRI images by a hierarchical self-calibration detection framework. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 316–325.

Zhao, S., Wu, X., Chen, B., Li, S., 2021. Automatic vertebrae recognition from arbitrary spine MRI images by a category-consistent self-calibration detection framework. Med. Image Anal. 67, 101826.