

Faint Features Tell: Automatic Vertebrae Fracture Screening Assisted by Contrastive Learning

Xin Wei^{*1,2}, Huaiwei Cong^{*1,2}, Zheng Zhang³, Junran Peng⁴, Guoping Chen^{†1} and Jinpeng Li^{†1,2}

¹HwaMei Hospital, University of Chinese Academy of Sciences (UCAS), Ningbo, China

²Ningbo Institute of Life and Health Industry, UCAS, Ningbo, China

³Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen

⁴National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

Abstract—Long-term vertebral fractures severely affect the life quality of patients, causing kyphotic, lumbar deformity and even paralysis. Computed tomography (CT) is a common clinical examination to screen for this disease at early stages. However, the faint radiological appearances and unspecific symptoms lead to a high risk of missed diagnosis, especially for the mild vertebral fractures. In this paper, we argue that reinforcing the faint fracture features to encourage the inter-class separability is the key to improving the accuracy. Motivated by this, we propose a supervised contrastive learning based model to estimate Genent’s Grade of vertebral fracture with CT scans. The supervised contrastive learning, as an auxiliary task, narrows the distance of features within the same class while pushing others away, enhancing the model’s capability of capturing subtle features of vertebral fractures. Our method has a specificity of 99% and a sensitivity of 85% in binary classification, and a macro-F1 of 77% in multi-class classification, indicating that contrastive learning significantly improves the accuracy of vertebrae fracture screening. Considering the lack of datasets in this field, we construct a database including 208 samples annotated by experienced radiologists. Our desensitized data and codes will be made publicly available for the community.

Index Terms—deep learning, vertebral fracture, contrastive learning, computer-aided diagnosis

I. INTRODUCTION

Vertebral fracture is a common disease that often occurs in aged people, which could severely affect patients’ living. Deformity and chronic pain are the major clinical manifestations of vertebral fracture, and according to Cauley et al. [1], long-term vertebral fractures can increase the mortality rate by eight times. However, such dangerous disease is often under-reported in clinical diagnosis. This is due to its less specific symptoms and less obvious radiological appearances, leading to radiologists’ neglect or misattribution[2]. Since early diagnosis and treatment are essential for alleviating vertebral fractures’ impact on human health, a computer-aided screening tool with high performance could be fairly helpful.

Currently, Genant semi-quantitative method[3] serves as the common standard in assessing vertebral fractures, but it

will be affected by radiologists’ experience and awareness of scrutinizing vertebrae radiography. On the other hand, deep learning based methods could be automatically inferred and thus excluding radiologists’ subjectivity, makes it an ideal method to screen vertebral fracture. Due to the slight difference among classes, the grading of vertebrae can be regarded as a fine-grained classification task. To address this, we propose a supervised contrastive learning based method to enhance the inconsistent feature among each grade. We validated our method on a dataset collected by us and a public dataset, both of them show improvements on classification by a large margin.

Our contribution can be summarized as follows:

- We design an end-to-end pipeline to segment, label and assess fracture on each vertebra of the given CT image. Such assessment procedure is fully automatic and without any human intervention, which excludes influence of radiologists’ subjectivity. We believe such approach could largely improve the under-diagnosed situation of vertebral fracture.
- We propose to utilize supervised contrastive learning in vertebrae fracture grading, and design series of studies to prove that our method has a better capability of detecting mild vertebral fracture. It can be concluded that forming feature space by contrastive learning further drives CNN to capture the information in the given images.
- To validate our method, we collected and arranged a novel vertebrae dataset that contains spine CT images of various fracture situations. Our dataset is collected from real clinical cases, which are well aligned and have suitable resolution for analysis. To support the research community of medical image analysis, **we will publicly share our desensitized dataset** shortly, together with the codes of this paper.

II. RELATED WORKS

A. Vertebrae Segmentation and Labeling

Segmentation and labeling of vertebrae are the fundamental tasks for further processing and analysis, for reliable segmentation and labeling algorithm could enable multiple automatic assessing tasks such as vertebral fracture grading or spine

This work was supported in part by National Natural Science Foundation of China (62106248) and Medical Health Science and Technology Project of Zhejiang Provincial Health Commission (2022KY1188)

^{*}Equal Contribution.

[†]Corresponding Authors: Guoping Chen(headoniones@aliyun.com) and Jinpeng Li(lijinpeng@ucas.ac.cn)

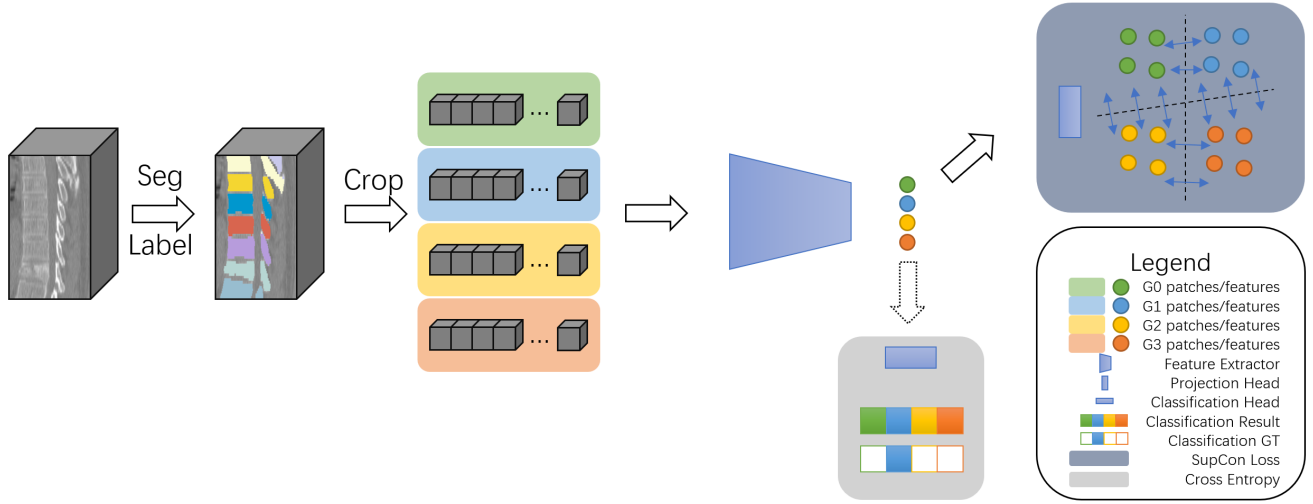


Fig. 1. Overview of our pipeline. Our pipeline is arranged in a two-stage manner. The vertebrae in the CT scans are segmented and labeled first to get the patches of vertebrae. It followed by the grading network which consists of a feature extractor, a projection head and a classification head. The features of patches are calculated by the feature extractor, and further clustered by fracture grade with the projection head via supervised contrastive learning. The grading results are given by the classification head, which won't propagate gradient to the feature extractor (marked as dotted arrow in the figure).

deformity detection. For this reason, vertebrae segmentation and labeling keep drawing research communities' attention. In 2019 and 2020, The Large Scale Vertebrae Segmentation Challenge (Verse) was held in conjunction with MICCAI, evaluating multiple algorithms of vertebrae segmentation and labeling. We highly suggest readers who are interesting in this topic referring to the report of the challenges[4]. In this paper, we utilize the algorithm of Payer et al. [5], which introduces a U-net[6] based structure and segments vertebrae in a coarse-to-fine manner.

B. Vertebrae Fracture Grading

On the other hand, studies about vertebrae fractures grading are relatively insufficient. Unlike segmentation and labeling tasks, grading of fractures is facing naturally imbalanced data for abnormal vertebrae only account for a small portion of overall vertebrae. This makes an enlarged data demand for fracture grading tasks. Since vertebral fracture is direct related to the deformity of vertebrae, conventional methods segment vertebrae to be assessed in the CT images, and calculate its shape statistic with the segmentation mask[7, 8, 9]. However, severe deformities like burst fractures could degrade the segmentation algorithm, limiting its major application to osteoporosis and compression fractures. Li et al. [10] assesses vertebrae fracture with neighboring CT slices, and Tomita et al. [11] aggregates feature in CT slices with a LSTM[12]. Murata et al. [13] also evaluates deep learning model's performance of vertebral fracture detection on plain spinal radiography. Similar to our idea, Nicolaes et al. [14] managed to detect vertebral fractures in CT volumes with a 3D CNN, and referring to metric learning, Hussein et al. [15] proposed a novel metric loss that could form a reasonable feature space. In this paper, we further demonstrate that supervised contrastive learning could reinforce the faint feature and form a better

clustered feature space, resulting in advanced performance for vertebral fracture grading.

C. Contrastive Learning

Contrastive learning aims at forming a clustered feature space to enhance feature extraction. Major contrastive learning methods focus on self-supervised learning, for clustering could be achieved only by appearances of given images, removing the necessity of manual annotation. The clustering is often made by narrowing the distance among positive samples while enlarging that among negative samples. In self-supervised scenario, the positive samples are two distinct views from the same item, while the negative samples are views from the others. MoCo[16] designed a memory mechanism to expand the negative samples from mini-batch to a dynamic memory bank, and SimCLR[17] carefully researched major factors of contrastive learning, finding the importance of projection head, data augmentation and batchsize. After that, they shared the idea mutually and came up with the updated version SimCLRv2[18] and MoCov2[19].

Contrastive learning could also improve fully-supervised learning. With fully annotated label, SupCon[20] expanding positive samples to same-class samples, showing a better performance as well as a more robust optimization comparing to cross entropy loss. In this paper, we follow the idea of SupCon[20] and further demonstrate its capability of fine-grained classification on medical images.

III. METHOD

A. Dataset

Generally, deep learning is a data-driven approach which requires massive annotated data to converge. However, annotating vertebral fracture of CT images is difficult and error-prone. To overcome this issue, we use a deep learning method

to aid the annotation procedure, alleviating the workload as well as improving the quality of the annotation. Specifically, we first adopt Payer et al. [5] to segment and label vertebrae in the CT volume. It could segment and label vertebrae with a Dice coefficient at 0.93, which is suitable for our application. By utilizing this automatic segmentation approach, radiologists could avoid manual segmentation of each vertebra in the CT scans. In practice, Payer et al. [5] could generally give precise segmentation masks and labels. Its major mistake is the occasionally mislabeling, which will be revised by our invited radiologists later.

Another issue of annotation is the faint radiological appearances of vertebral fractures. This may lead to inconsistent annotation, which brings ambiguous feature that dramatically hurt the clustering of contrastive learning. To address this, we first invited 3 junior radiologists to assess fractures of the vertebrae we segmented. The radiologists are entrusted to annotate each vertebra with its Genant's Grade, as well as revise the segmentation or label if the algorithm[5] gives incorrect output. For the disagreements in annotation, we invited a senior radiologist to verify the voting of initial annotation and give the conclusion. With the proposed annotation, we could arrange a high-quality vertebral fracture dataset with rather light workload.

Our method does not restrict to certain population, so we make no assumption about gender or age of participants we collected in the dataset. The generalization capability of deep learning model could promise the model's applicability for broad population. We balanced the ratio of each Genant's Grade in the adopted participants to relieve potential data imbalance. Genant's Grade classifies the fractures into 4 classes of G0, G1, G2 and G3, which can be regarded as normal, mild, moderate and severe. In practice, we adopt 208 CT scans in our dataset with the ratio of G0:G1:G2:G3=1:3:3:3, including 2,423 vertebrae in total.

B. Model

Our CNN architect is designed with a two-stage manner. Vertebrae in the CT volume are cropped to patches first with an existing segmentation and labeling model. Then the vertebrae patches will be fill into the grading network to estimate the Genant's Grade of each vertebra with a supervised contrastive learning manner. An overview of our pipeline is illustrated in Fig. 1, and the detailed design is stated as follows:

Segmentation and Labeling We adopt Payer et al. [5] as the segmentation and labeling model, which we also utilized in dataset annotation. The output segmentation masks are expanded to bounding box first to include nearby tissue as an additional clue for fracture grading, then the patches of vertebrae are cropped accordingly. we choose two windows to extract information from original CT patches: the bone window that sets window level with 1500 HU and window width with 400 HU, and the soft tissue window that sets window level with 200 HU and window width with 40 HU. To our knowledge, such windowing contains sufficient information for vertebrae fracture grading.

We also take the original segmentation mask and label into account, by concatenating the segmentation mask to the input patches of CNN. Label information is also helpful, for underlying fracture-related feature of different vertebrae are potentially varied. We modulate the original binary segmentation mask with its normalized label and concatenate it with the two windows of CT image at channel dimension. The combined three channel image will be used as input of the subsequent grading network, with an additional resampling to ensure the isotropic voxel spacing and uniform orientation.

Network Structure We follow the contrastive learning methods[17, 20] to design our grading network, which can be separated into three parts. First, a backbone network acts as the **feature extractor** to extracts and encodes the feature of radiograph. Then a **projection head** projects the feature to a low-dimensional space, where the optimization of contrastive learning applied. Additionally, a **classification head** takes the output of feature extractor and estimates the grading result, with a cross entropy loss to optimize. In practice, shallow networks like several linear layers could fulfill the intention of projection head and classification head. Also, as contrastive learning methods have proven that the feature space is separable without any information of classification head, we follow the advice of SupCon[20] to detach the classification head so that its gradient won't be back propagated to other part of the network. At inference time, the projection head will be discarded, leaving output of classification head as the grading result. A graphic demonstration of our network is illustrated in Fig. 1.

We adopt 3D-SEnet50[21] as the structure of feature extractor, which adds an attention mechanism to the conventional ResNet50[22] model. It shows a better feature extracting capability especially for the 3D fine-grained classification. The projection head and the classification head both consist of a single linear layer, with a 128-dimensional output for the projection head, and a 4-dimensional output for the classification head of 4-level grading. We also follow the convention to normalize the magnitude of the 128-dimensional vector to 1 for a spherical space is better for contrastive learning.

Loss Function Contrastive learning methods augment input sample to a pair of distinct views. For the self-supervised manner, the clustering is conducted by taking the pair of views as positive sample mutually while the others as negative samples. This manner can cluster appearance-similar features without additional labels, makes it a feasible and prevalent solution for self-supervised learning.

On the other hand, supervised contrastive learning keeps utilizing annotated labels to guide the clustering of feature space, with the main assumption that samples in the same class could vary in appearance. We keep using supervised manner for the faint disparities among fracture grades is much weaker than disparities in the overall appearance of vertebrae. Cole et al. [23] also gives an observation that fine-grained classification could degrade the self-supervised contrastive learning. However, we argue that with the guidance of class label, the contrastive learning method is encouraged to detect

TABLE I
QUANTITATIVE RESULT. FE, SPE, SEN ARE SHORT FOR FEATURE EXTRACTOR, SPECIFICITY AND SENSITIVITY.

FE	Loss	Optimizer	Dataset	Binary Classification			Multi-Class Classification		
				AUCROC	SPE	SEN	Macro-F1	Macro-Precision	Macro-Recall
ResNet50	Cross Entropy	Adam	Ours	0.95	0.96	0.72	0.53	0.52	0.56
ResNet50	SupCon	SGD	Ours	0.98	0.99	0.83	0.67	0.67	0.65
SENet50	Cross Entropy	Adam	Ours	0.97	0.99	0.79	0.59	0.61	0.57
SENet50	SupCon	Adam	Ours	0.98	0.97	0.87	0.64	0.62	0.67
SENet50	SupCon	SGD	Ours	0.98	0.99	0.85	0.77	0.78	0.77
SENet50	Cross Entropy	Adam	Verse	0.90	0.94	0.68	0.59	0.59	0.62
SENet50	SupCon	SGD	Verse	0.93	0.94	0.72	0.72	0.72	0.72
Husseini et al. [15]	Grading Loss	Adam	Verse w/o G1	-	0.99	0.77	-	-	-
SENet50	SupCon	SGD	Verse w/o G1	0.99	0.99	0.88	0.86	0.84	0.85

the fine-grained disparities which contribute most to the task.

We adopt SupCon loss[20] as the loss function of our method, with the equation in (1)

$$\mathcal{L}^{sup} = \sum_i \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_p) / \tau)}{\sum_k \mathbb{I}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k) / \tau)} \quad (1)$$

It is a enhanced version of NT-Xent loss[17] which expanded the positive sample set $P(i)$ to same-class samples, i.e. $P(i) \equiv \{p | \tilde{\mathbf{y}}_p = \tilde{\mathbf{y}}_i, p \neq i\}$. Here the \mathbf{z} is the feature vector of projection head, and the temperature parameter τ is used to control the intensity of loss. With SupCon loss, the disparities among classes become the major clues of classification, and in the radiographic diagnosis scenario, such disparities strongly hint regions of lesion. To prove this, we use Grad-CAM[24] to visualize some results to check whether features of lesion region attract high attention. The visualization can be found in the *Results* section.

Data augmentation For contrastive learning, data augmentation is used to generate multiple distinct views of original training data. SimCLR[17] carefully researched the combinations of different data augmentations and their impact on the contrastive learning. However, as Purushwalkam and Gupta [25] mentioned, the detailed configuration of data augmentation is often data-biased, especially for our task that is distinct from classification of natural images. Intuitively, data augmentation should try to avoid interfering the clue of classification, but regular data augmentation cannot promise this, for the vertebral fracture are related to global feature like posture and shape, as well as the local feature like minor fractures. To address this, as listed below, we design a set of data augmentations that are specialized for vertebral fracture grading:

- Random Padding.
- Pre-Rotation Mask of 2 boxes with side lengths of 1-20 voxels.
- Random Zoom in 0.9-1.1x.
- Gaussian Noise with $\mu = 0$ and $\sigma = 0.05$.
- Random Shift in ± 10 voxels.
- Random Rotation in ± 10 degrees.

- CT Value Jittering with the function

$$HU = (HU \times p1)^{p2}$$

, where p1 is within 0.9-1.1 and p2 is within ± 2 .

- Post-Rotation Mask of 2 boxes with side lengths of 1-20 voxels.

Random Padding takes an input patch and rescale its longest side length to 128, and further randomly padding it to the resolution of 3*128*128*128(3 are the channel dimension introduced in section *Segmentation and Labeling*). The other data augmentations are applied with possibilities of 70%. Multi-dimensional augmentations are applied to each axis with individual parameters. The result shows that it reaches a good trade-off between distinct views and fracture clue reserving.

As introduced in *Loss Function* section, input vertebra patches will be augmented twice individually to generate the pair of views. The pair will be accumulated to a mini-batch and fed into the subsequent grading network.

C. Training

We specially designed several training techniques to facilitate the contrastive learning. In this section, we will introduce the details in the training procedure.

Batch Sampler Ordinary batch sampler often traverses the dataset randomly in an epoch. However, since vertebral fracture is happened occasionally, its fracture grade is naturally facing data imbalance. And as for the supervised contrastive learning, this issue is non-negligible for the clusters may be optimized unevenly, causing a biased grading. To avoid this, we design a Per-Class Sampler, which randomly sample n patches in each class, forming a mini-batch of nC patches where C is the number of classes. The sampling procedure is without placement, and it resets every time when the least class traversed. Less rigorously, we still call it an 'epoch', and giving enough epochs, the dataset could be traversed with a data-adaptive sampling rate. This is a simple yet powerful batch sampler, which could avoid data imbalance as well as accelerate the converge of optimization.

Optimizer We adopt SGD as the optimizer with a weight decay of 1e-4 and a momentum of 0.9. For the supervised

contrastive learning, it converges to a better minimum than adaptive optimizers like Adam[26] in practical. The learning rate starts at $1e-3$, and decays by 0.1 at epoch 800 and 900.

Miscellaneous We implement our methods with Pytorch[27] on a workstation with two RTX A6000 graphic cards. The dataset is split into training set and test set with the ratio of 4:1, and vertebrae of individuals won't be split into different set. We sample 6 vertebrae patches per class at each iteration, forming a batchsize of 24 in total. Each patch is augmented twice to generate the pair of views. Due to the data limitation, the batchsize we set is relatively small comparing to that on natural images, but the result is still impressive, showing the potential of contrastive learning on medical images. On our workstation, the network takes about 18 hours to converge, with approximately 1000 epochs.

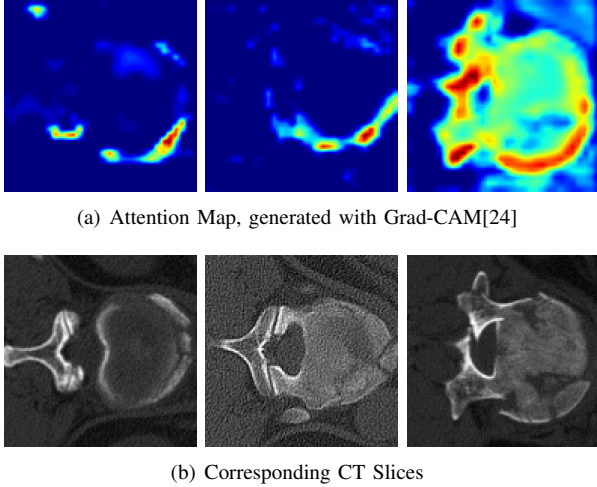


Fig. 2. Qualitative Result. The Genant's Grade of the CT slices are G1, G2 and G3 respectively.

IV. RESULTS

A. Quantitative Study

Metrics We mainly evaluate two aspects of our method, which are the **binary classification** of Benign(G0) vs Malignant(combination of G1,G2 and G3), as well as the **multi-class classification** of 4 grades. The intention of evaluating binary classification is that it's the most sensitive metrics for missed diagnosis.

To avoid the misleading of unbalanced test set, we use AU-CROC as the main metrics of binary classification, and macro-F1 score as the main metrics of multi-class classification. Note that we always take multi-class classification as the training target, while the evaluation of binary classification is only happened at inference time by combining the class G1, G2 and G3 in ground truth and prediction respectively.

ROC curve Firstly, we demonstrate the binary classification performance of our method with the ROC curve in Fig. 3. With a specificity of 99% and sensitivity of 85%, it could improve the diagnostic rate of vertebral fracture by a large margin.

Ablation study To prove the aforementioned benefits we claimed about our methods, we evaluate the basic ResNet50

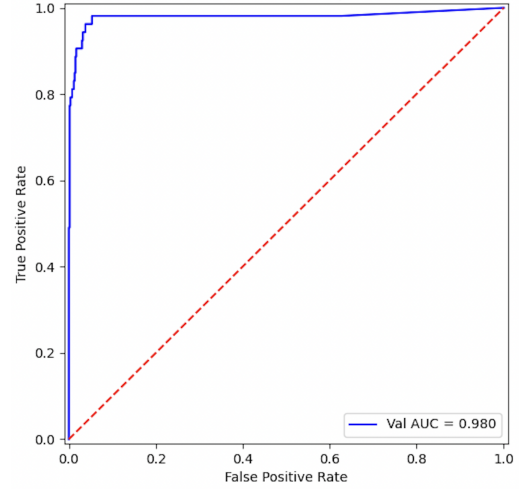


Fig. 3. The ROC curve of binary classification.

models with cross entropy loss, and gradually add it to the final version. The detailed experiments and results are listed in Table. I, and our full model as well as the best result are marked in bold. Note that SupCon loss largely improve the multi-class classification result comparing to cross entropy, which is the key contribution of our work.

Public Dataset Validation To avoid being data biased, we also evaluate our method on a public vertebral fracture dataset Verse[4]. As a challenge dataset, it contains more complicated situation than our clinical dataset, e.g., more varied resolution and orientation, as well as additional cervical vertebrae. These situations cause degraded metrics than our dataset, however the improvement of supervised contrastive learning is still remarkable. Also, our dataset does not contain any vertebrae with artificial implants, for they actually do not need screening. To keep a reasonable comparison, we remove vertebrae with artificial implants in Verse[28]. The result can be found in Table. I.

Comparative study We choose Hussein et al. [15] to conduct the comparative study, for it was validated on Verse[28] as well. With the auxiliary information from fracture grade, it managed to improve the binary classification with a novel grading loss. Due to the difficulty in distinguishing Grade 0 and Grade 1, Hussein et al. [15] didn't take Grade 1 fractures into account, while ours could detect the mild Grade 1 fractures and further enable multi-class classification. For reference, we also validate our method on Verse[28] without Grade 1 fractures. The result can be found in Table. I.

B. Qualitative study

As we mentioned, supervised contrastive learning picks the feature that strongly hints the region of lesion, and Grad-CAM[24] is a proper tool to visualize such region. The volumetric Grad-CAM[24] is generated with the implement of Gotkowski et al. [29]. As show in Fig. 2, with our method, the model could detect the regions of vertebral fractures in multiple situations.

V. CONCLUSION

We design a pipeline of vertebral fracture grading with supervised contrastive learning, which shows a great performance in both binary and multi-class classification. We believe our method could improve the diagnostic rate of vertebral fracture in real clinical scenario. Also, we arranged a high-quality vertebral fracture dataset with careful annotations of Genant's Grade, which may alleviate the data deficiency of related research.

REFERENCES

- [1] J. Cauley, D. Thompson, K. Ensrud, J. Scott, and D. Black, "Risk of mortality following clinical fractures," *Osteoporosis international*, vol. 11, no. 7, pp. 556–561, 2000.
- [2] A. Panda, C. J. Das, and U. Baruah, "Imaging of vertebral fractures," *Indian journal of endocrinology and metabolism*, vol. 18, no. 3, p. 295, 2014.
- [3] H. K. Genant, C. Y. Wu, C. Van Kuijk, and M. C. Nevitt, "Vertebral fracture assessment using a semiquantitative technique," *Journal of bone and mineral research*, vol. 8, no. 9, pp. 1137–1148, 1993.
- [4] A. Sekuboyina, M. E. Hussein, A. Bayat, M. Löffler, H. Liebl, H. Li, G. Tetteh, J. Kukačka, C. Payer, D. Štern *et al.*, "Verse: A vertebrae labelling and segmentation benchmark for multi-detector ct images," *Medical image analysis*, vol. 73, p. 102166, 2021.
- [5] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Coarse to fine vertebrae localization and segmentation with spatialconfiguration-net and u-net," in *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP*, vol. 5, 2020, pp. 124–133.
- [6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [7] S. Ghosh, A. Raja's, V. Chaudhary, and G. Dhillon, "Automatic lumbar vertebra segmentation from clinical ct for wedge compression fracture diagnosis," in *Medical Imaging 2011: Computer-Aided Diagnosis*, vol. 7963. SPIE, 2011, pp. 21–29.
- [8] J. E. Burns, J. Yao, and R. M. Summers, "Vertebral body compression fractures and bone density: automated detection and classification on ct images," *Radiology*, vol. 284, no. 3, p. 788, 2017.
- [9] A. Suri, B. C. Jones, G. Ng, N. Anabaraonye, P. Beyrer, A. Domi, G. Choi, S. Tang, A. Terry, T. Lechner *et al.*, "Vertebral deformity measurements at mri, ct, and radiography using deep learning," *Radiology: Artificial Intelligence*, vol. 4, no. 1, p. e210015, 2021.
- [10] Y. Li, Y. Zhang, E. Zhang, Y. Chen, Q. Wang, K. Liu, H. J. Yu, H. Yuan, N. Lang, and M.-Y. Su, "Differential diagnosis of benign and malignant vertebral fracture on ct using deep learning," *European Radiology*, vol. 31, no. 12, pp. 9612–9619, 2021.
- [11] N. Tomita, Y. Y. Cheung, and S. Hassanpour, "Deep neural networks for automatic detection of osteoporotic vertebral fractures on ct scans," *Computers in biology and medicine*, vol. 98, pp. 8–15, 2018.
- [12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.
- [13] K. Murata, K. Endo, T. Aihara, H. Suzuki, Y. Sawaji, Y. Matsuoka, H. Nishimura, T. Takamatsu, T. Konishi, A. Maekawa *et al.*, "Artificial intelligence for the detection of vertebral fractures on plain spinal radiography," *Scientific Reports*, vol. 10, no. 1, pp. 1–8, 2020.
- [14] J. Nicolaes, S. Raeymaeckers, D. Robben, G. Wilms, D. Vandermeulen, C. Libanati, and M. Debois, "Detection of vertebral fractures in ct using 3d convolutional neural networks," in *International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging*. Springer, 2019, pp. 3–14.
- [15] M. Hussein, A. Sekuboyina, M. Loeffler, F. Navarro, B. H. Menze, and J. S. Kirschke, "Grading loss: a fracture grade-based metric loss for vertebral fracture detection," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 733–742.
- [16] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.
- [17] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
- [18] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *Advances in neural information processing systems*, vol. 33, pp. 22 243–22 255, 2020.
- [19] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.
- [20] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 661–18 673, 2020.
- [21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [23] E. Cole, X. Yang, K. Wilber, O. Mac Aodha, and S. Belongie, "When does contrastive visual representation learning work?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 755–14 764.
- [24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
- [25] S. Purushwalkam and A. Gupta, "Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3407–3418, 2020.
- [26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.
- [28] M. T. Löffler, A. Sekuboyina, A. Jacob, A.-L. Grau, A. Scharr, M. El Hussein, M. Kallweit, C. Zimmer, T. Baum, and J. S. Kirschke, "A vertebral segmentation dataset with fracture grading," *Radiology: Artificial Intelligence*, vol. 2, no. 4, 2020.
- [29] K. Gotkowski, C. Gonzalez, A. Bucher, and A. Mukhopadhyay, "M3d-cam: A pytorch library to generate 3d data attention maps for medical deep learning," 2020.