2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-2923-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISBI52829.2022.9761435

MSRT: MULTI-SCALE SPATIAL REGULARIZATION TRANSFORMER FOR MULTI-LABEL CLASSIFICATION IN CALCANEUS RADIOGRAPH

Yuxuan Mu, He Zhao, Jia Guo, Huiqi Li

Beijing Institute of Technology, Beijing 100081, China

ABSTRACT

Calcaneus fracture is one of the most common fractures which affect daily life quality. However, calcaneus fracture subtype classification is a challenging task due to the nature of multi-label as well as limited annotated data. In this paper, an augmentation strategy called GridDropIn&Out (GDIO) is proposed to increase the uncertainty of the rough input mask and enlarge the dataset. A spatial regularization transformer (SRT) is designed to capture labels' spatial information, while a multi-scale attention SRT (MSRT) is built to synthesize spatial features from different levels. Our final proposal achieves an mAP of 87.54% in classifying six calcaneus fracture types.

***Keywords*—** Calcaneus Fracture, Multi-label Classification, Multi-scale Attention, Transformer

1. INTRODUCTION

Calcaneus is a piece of bone located in the foot's heel and is one of the most important stressed bones for daily activities. Calcaneus fractures, which account for 60% of tarsal fractures, are the most common type of tarsal fracture and usually occur during a high-energy event [1].

Plain X-ray is a suitable method for fracture screening. It is low cost and easy to use. However, the lack of information contained in plain radiographs is a challenge. The diagnosis based on radiographs requires the participation of experienced senior clinicians. In this paper, we design a computer-aided diagnosis system for calcaneal fracture based on X-ray film, which carries out multi-label classification for fracture types to provide necessary diagnostic reference for physicians.

In our previous work, a normalization and rough registration method for calcaneus was proposed [2]. The preliminary fracture region was predicted by a multi-task model trained with limited rough annotations. These methods only screen fracture, while more worthy information like fracture subtypes is not involved. The identification of fracture subtypes is important for proper treatment. Patients may suffer from multiple types of fractures, which means it is a multi-label classification task. Six subtypes of calcaneus fracture are investigated in this paper: Calcaneal Tuberosity, Primary Line, Collapse Fracture Block, Calcaneal Body, Posterior Articular Surface, and Tongue Fracture Block. These fracture types were related to fracture location and shape.

Classification of calcaneal fracture types is a multi-label classification problem, which means for a single instance, it is associated with multiple labels simultaneously. Since the label category in our task has a high correlation with the position of the lesion, the use of this correlation for feature selection is one of the keys to improve classification performance. From this aspect, the representative approach is Spatial Regularization Net (SRN) proposed by Zhu et al. [3]. SRN makes the model pay attention to the object's spatial position through a multistage training process. Chen and You et al. combined the spatial information and semantic information of labels and proposed a multi-label classification model using cross-modality features and attention mechanism [4, 5]. Complex training process in these methods also leads to an extensive training variance. Furthermore, the scale changes of the image are diverse, which means it is difficult to fit the multi-scale spatial location relationship of the label with the general convolution structure.

In recent years, lots of research has verified that the self-attention mechanism enables DNNs to learn more efficiently and effectively. Wang et al. proposed Non-local Neural Network which captures long-range dependencies [6]. Hu et al. designed the Squeeze-and-Excitation block synthesized the features from different perception fields [7]. Later, there are plenty of research combining multiple attention mechanisms to achieve improved performance [8]. Dosovitskiy et al. proposed the Vision Transformer (ViT) image classification network [9]. The multi-head attention module in Transformer, which is a multi-stream combination of scaled dot-product attention, intrinsically perceives information globally. But it acquires a large training dataset because of its shortage of inductive bias in the image domain.

The contribution of our work can be summarized as follows: 1) A Multi-scale Spatial Regularization Transformer (MSRT) is proposed to classify different fracture categories in the same image. 2) A spatial regularization transformer (SRT) is designed to capture labels' spatial semantic information and dependence by

perceive global correlation among patches. 3) Furthermore, an augmentation strategy called GridDropIn&Out is proposed to enhance the performance, which adapts the model to guided mask with uncertainty. The final proposal achieves the best results with mAP 87.54%, which is 5% higher than the state-of-the-art. Our work preliminarily demonstrates the potential of Transformer for medical image context without a large-scale dataset.

2. METHOD

The network structure of our proposed method is illustrated in Fig. 1, which contains multi-scale Transformer-style spatial regularization blocks and attention module. The proposed data augmentation method GDIO in Fig. 1c introduces the uncertainty to help model learn efficiently.

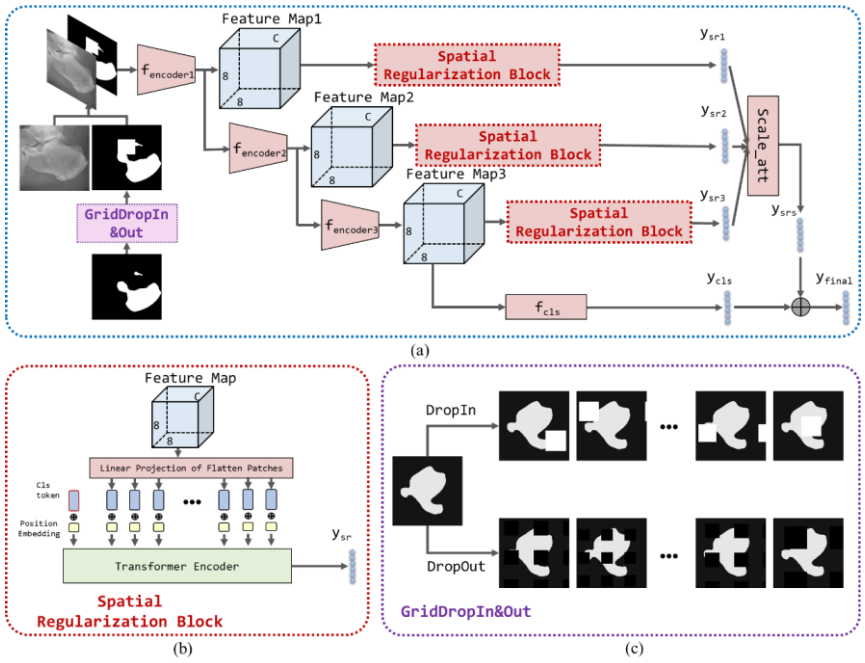


Fig. 1. Network architecture: (a) the overall design of our proposed Multi-Scale Spatial Regularization Transformer (MSRT); (b) the spatial regularization block built with Transformer; (c) the illustration of GDIO.

2.1 Spatial Regularization Transformer (SRT)

Our SRT model takes X-ray image and the corresponding fracture mask as inputs and gives the prediction of the fracture categories. The fracture mask can be generated by arbitrary segmentation method.

The overall design of SRT is based on SRN [5]. We apply the Transformer as the spatial regularization block (Fig. 1b), which is good at long-range dependencies perception with excellent efficiency. Firstly, the feature in latent space is mapped to $\mathbb{R}^{C\times 8\times 8}$ (C is the number of categories) considering the feature map to 64 spatial-wise patches. We also add class tokens to collect specific information from each patch. Finally, a trainable position feature embedding vector is added to the series to obtain the implicit spatial representation. Serialized features are fed into standard Transformer modules, including multi-layer multi-head attentional modules and fully connected modules, with a residual fashion. Transformer naturally separate spatial-wise feature fusion and channel-wise feature synthesis. The fully-connected layer is the equivalent for convolution with kernel size of 1×1, which deals with each patch independently. In this manner, the latent feature vectors can work as 64 isolated specialized classifiers and 1 comprehensive classifier. Moreover, the Dot-Product Attention (DPA) is inherently global for the matrix multiplication. The multi-head strategy which integrates several DPAs enables the attention net to fit multiple complex correlations. This design has significant advantages over convolution-based spatial attention modules which needs multi-step to build the global dependencies and would blur the operations of space and channel.

Furthermore, we explore multi-scale perception to SRT and design a scale attention block to synthesize the features of different levels to provide a scale adaptive spatial regular term. Its architecture is shown in Fig. 1(a).

The network uses EfficientNetB3 as the feature extraction encoder. The last three depths feature maps of the encoder are fed into three spatial regularization modules. Through a fully-connected layer, the dim of vector is adjusted to the specified higher dimension, which is 128, 64 and 32 respectively in this network. Only the class tokens

vector which contains significant information and drops the tangential features is directly used for the spatial regularization prediction. The output result y_{sri} of three spatial regularization modules of different scales will be fused into a spatial regularization prediction through a scale attention module in a Squeeze-and-excitation fashion, and then added to the prediction output in the backbone network to get the final multi-label prediction.

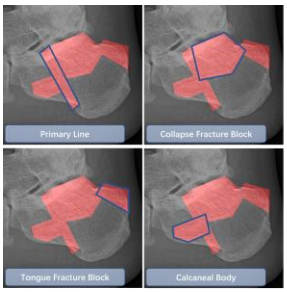
2.2 Uncertain Mask Guided Classification

The fracture region input mask can be generated by arbitrary segmentation model trained with rough annotation. Compared between the fine labeling ground-truth and the predicted mask, the prediction has zigzagged edge and it misses or adds partial regions, which introduces noise to the useful information. Therefore, we design an augmentation method inspired by GridDropout [10] to increase the uncertainty of the segmentation mask and stimulate the network to pay attention to the image texture information with the assistance of the mask information.

Two operations are adopted here called DropIn and DropOut. A random patch with fixed size is selected and is filled either with pixel value of 255 or 0 and the effect is to gain or drop the mask block randomly. An example is shown in Fig. 1(c). For an instance image, the model would see it in random but instrumental perspectives through masks in epochs, which teach the model to handle noise and make use of the supplementary information of fracture region.

3. EXPERIMENTS & RESULTS

3.1 Dataset



(a)

Fracture types	Label Count
Calcaneal Tuberosity	325
Primary Line	329
Collapse Fracture Block	248
Calcaneal Body	298
Posterior Articular Surface	145
Tongue Fracture Block	154

(b)

Fig. 2. Illustrations and statistics of the dataset: (a) Example of the calcaneal fracture types; (b) Amount of fracture-subtype labels

The dataset contains 905 lateral X-ray images of the calcaneus, which were collected from the clinical data of orthopedics departments of the First Affiliated Hospital of Jinzhou Medical University and were pre-processed by our previous rough registration work [2]. The image category label contains six fracture categories: Calcaneal Tuberosity, Primary Line, Collapse Fracture Block, Calcaneal Body, Posterior Articular Surface, and Tongue Fracture Block (see Fig. 2(b)). In this task, the Label Cardinality is 2.94, and the data set normalized Label Diversity is 0.061. Different fracture types can be presented in a single image of a calcaneal fracture, as shown in Fig. 2(a).

3.2 Experiments Settings

Since our dataset is small, 4-fold cross-validation was performed and the results were averaged. We adopted Adam as the optimizer and set the learning rate to 0.0001, momentum to 0.9, and maximum epochs of 70. Besides, we decreased the learning rate by 10% every 20 epochs and utilized early stopping to avoid over-fitting. For model comparison on our dataset, we adopted ResNet50, CNN-RNN [11], Efficientnetb3, SRN (auto_aug) [5], SRN, SRT, MSRN. SRN (auto_aug) was trained with common augmentation strategy while others were trained with GDIO. CNN-RNN used ResNet50 as backbone while others used Efficientnetb3.

3.3 Results

Our SRT model significantly improves the classification performance compared with the EfficientNet model (see Table 1). Moreover, the multi-scale version increases the score on most of the fracture types, except Primary Line. We suppose this is because Primary Line is very directional while our multi-scale operation is isotropic. The average performance over the six classes improves 3.8% on accuracy and 3.1% on F1 score.

Table 1. Statistical results between our approach and other methods by category classification.

Category	ACC/%			F1 score/%		
	Effib3	SRT	MSRT	Effib3	SRT	MSRT
Calcaneal Tuberosity	72.57	73.89	**76.11**	77.70	79.00	**79.55**
Primary Line	85.84	**87.61**	84.96	90.19	**91.03**	89.03
Collapse Fracture Block	84.07	84.51	**84.96**	84.87	85.48	**85.83**
Calcaneal Body	64.60	68.14	**71.68**	61.54	68.42	**71.17**
Posterior Articular Surface	76.99	79.65	**80.97**	76.36	77.45	**77.95**
Tongue Fracture Block	88.05	89.38	**91.59**	73.27	76.47	**81.19**
Overall	78.69	80.53	**81.71**	78.92	80.70	**81.38**

Further statistic results of comparison study are shown in Table 2 on average performance over six categories. Exact matching accuracy (Ex ACC), macro/micro precision (P-C/P-O), macro/micro recall (R-C/R-O), macro/micro F1-score (F1-C/F1-O), and mean average precision (mAP) were adopted for performance evaluation metrics. The proposed MSRT is superior to all state-of-the-art methods on most metrics. Since there is a trade-off between precision and recall depending on the threshold, mAP would be more representative for model performance assessment. And we can see that MSRT shows significant performance improvement. Considering the exact matching accuracy, the

results reveal that MSRT predicts more label combinations precisely, which demonstrates its strength in modeling labels correlations.

Table 2. Statistical results of comprehensive performance

Model	Ex Acc	mAP	F1-C	P-C	R-C	F1-O	P-O	R-O
ResNet50	26.99	81.37	70.96	73.16	73.66	75.64	73.02	78.45
CNN- RNN	21.24	74.50	71.52	67.56	77.62	74.17	69..16	79.97
Efficientnetb3	30.53	82.27	77.32	75.01	80.55	78.92	75.98	82.09
SRN (auto_aug)	32.74	83.87	78.36	74.51	**83.55**	79.58	74.24	**85.73**
SRN	33.63	84.21	79.99	77.32	82.91	80.79	77.89	83.92
MSRN	35.40	84.79	79.77	78.15	81.48	80.50	78.70	82.39
SRT	34.07	84.50	79.64	77.12	82.53	80.70	77.86	83.67
MSRT	**37.61**	**87.54**	**80.79**	**80.06**	81.60	**81.38**	**80.53**	82.25

3.4 Ablation Study

Table 3 shows that model trained with our proposed GDIO augmentation strategy gains super performance compared with the normal augmentation. The results of SRN (auto_aug) and SRN shown in Table 3 also identified the strength of GDIO on model's precision improvement.

For the effect of Transformer, while the statistical results of SRN are close to SRT (shown in Table 2), the training strategy of SRT is much simpler. Besides, when we expand them to MSRN and MSRT, respectively, MSRT performs much better since Transformer is easier to optimize under complex model architectures. And we can find obvious performance enhancement from the multi-scale attention architecture from the preferable results of MSRN and MSRT.

Table 3. Statistical results by GDIO and Auto_aug

Augment Strategy	ACC/%		F1-Score/%	
	GDIO	Auto_aug	GDIO	Auto_aug
Calcaneal Tuberosity	72.57	69.47	75.97	76.92
Primary Line	86.73	83.63	90.32	88.33
Collapse Fracture Block	84.51	83.63	85.48	85.82
Calcaneal Body	71.24	69.03	70.85	69.83
Posterior Articular Surface	77.43	76.99	74.63	76.15
Tongue Fracture Block	90.27	88.94	78.00	73.12
Overall	**80.46**	78.61	**80.12**	79.58

4. DISCUSSIONS

The Transformer module training on our dataset does not show a performance degradation compared with the convolution network. We assumed that rough registration of calcaneus almost eliminates affine transformation. The two-dimensional properties of serialized images are less affected. This avoids the disadvantage of Transformer's inductive bias for vision and takes full advantage of the attention mechanism of long semantic dependencies. In medical images, imaging results from modes such as MRI, CT, and plain X-ray are registered or registerable which is similar to the context in this paper. Therefore, Transformer has strong applicability in the field of medical image processing and has promising application prospects.

5. CONCLUSIONS

In this paper, we proposed a Multi-Scale Spatial Regularization Transformer for multi-label classification on calcaneus radiograph. Our experimental results indicate that our approach can recognize the fracture structure difference among the six types. The superior performance is achieved by multi-scale spatial regularization Transformer, while our DropIn&Out augmentation methods further improve the performance by introducing uncertainty of the input fracture mask. In addition, our work demonstrates the potential of Transformer for medical image context. In our future work, we will explore the spatial regularization on other data modality and address noisy labels for the model's stability and generalization.

6. REFERENCES

[1] Joseph R. Cass, "Calcaneus (Heel Bone) Fractures," 2016. OrthoInfo - AAOS.

[2] J. Guo et al., "Automatic analysis system of calcaneus radiograph: Rotation-invariant landmark detection for calcaneal angle measurement, fracture identification and fracture region segmentation," Comput. Methods Programs Biomed., vol. 206, p. 106124, 2021.

[3] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5513–5522.

[4] T. Chen, L. Lin, X. Hui, R. Chen, and H. Wu, "Knowledge-guided multi-label few-shot learning for general image recognition," IEEE Trans. Pattern Anal. Mach. Intell., 2020.

[5] R. You, Z. Guo, L. Cui, X. Long, Y. Bao, and S. Wen, "Cross-modality attention with semantic graph embedding for multi-label classification," in Proceedings of the AAAI Conference on Artificial Intelligence, 2020, vol. 34, no. 07, pp. 12709–12716.

[6] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.

[7] 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.

[8] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in Proceedings of the European conference on computer vision, 2018, pp. 3–19.

[9] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv Prepr. arXiv2010.11929, 2020.

[10] P. Chen, "GridMask data augmentation," arXiv Prepr. arXiv2001.04086, 2020.

[11] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A Unified Framework for Multi-label Image Classification."