# KALM: KEY AREA LOCALIZATION MECHANISM FOR ABNORMALITY DETECTION IN MUSCULOSKELETAL RADIOGRAPHS

*Wei Huang* *Zhitong Xiong* *Qi Wang*[*] *Xuelong Li*

School of Computer Science and Center for OPTical IMagery Analysis and Learning(OPTIMAL),
Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China

## ABSTRACT

Recently abnormality detection in musculoskeletal radiographs has attracted many attentions. For abnormality detection, it is crucial to locate the most important area in the musculoskeletal radiographs. To achieve this goal, we propose a key area localization mechanism (KALM) for abnormality detection for the first time in this paper. The proposed KALM explicitly defines the process of selecting the most important area from the whole image with using only image-level label. Based on KALM, we further present a joint global and local feature representation strategy for abnormality detection which takes as input both the entire image and the selected local area. The experimental results based on several classical convolutional neural network (CNN) architectures of MURA, the largest abnormality detection dataset of musculoskeletal radiographs, demonstrate the effectiveness of our KALM.

***Index Terms***— Key area localization mechanism, abnormality detection, musculoskeletal radiographs, CNN

## 1. INTRODUCTION

Recently, automated medical image analysis has achieved great development in the past few years benefiting from the breakthrough of deep learning techniques, including abnormality detection [1], lesion classification [2], object detection [3], organ segmentation [4] and the others. Musculoskeletal radiography is one of the most common medical imaging examination, and therefore abnormality detection in musculoskeletal has important potentiality in clinical application. Automated abnormality detection can adjust worklist prioritization, which is helpful for combating radiologist fatigue and improving work efficiency.

The methods used in abnormality detection can be roughly divided into two types: traditional methods and deep learning methods. Traditional methods [5] rely on hand-crafted features that are at the pixel level or the local-region level, while deep learning methods especially convolutional neural network (CNN) [6, 7] is an end-to-end trainable framework which can automatically learn the most important semantic features with the guidance of attribute label. Because abnormality detection in musculoskeletal radiographs is an image understanding task at the semantic level, the trainable end-to-end CNN is more suitable for it. For a CNN-based detection model, the key step is to find the most meaningful area from the whole image, which is the crucial reference for the detection result.

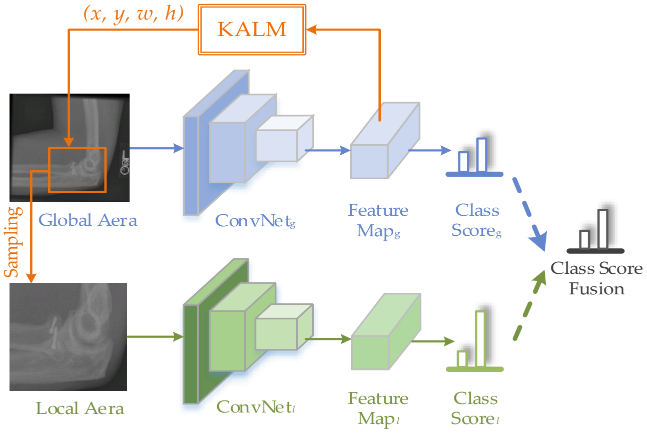

**Fig. 1**. The framework of joint global and local feature representation with KALM.

To solve this problem, we specially design a key area localization mechanism (KALM) for abnormality detection in musculoskeletal radiographs in this paper, which connects two deep feature extractors of different scales as shown in Fig. 1. In the framework, the global features and the key local features are extracted by $ConvNet_g$ and $ConvNet_l$ respectively. From the perspective of feature extraction, global and local features are complementary because the former provide the comprehensive information while the later emphasize the key information, which probably is the abstract of the abnormality in the image. Then the global and local class scores are fused as the final classification result. The proposed KALM builds a bridge between the two types of feature extractors. It can accurately locate the key area that contains the most

Qi Wang is the corresponding author. This work was supported by the National Key R&D Program of China under Grant 2017YFB1002202, National Natural Science Foundation of China under Grant U1864204, 61773316, U1801262, 61871470, and 61761130079.

important local features. More specifically, the KALM calculates the center coordinate[$x$, $y$], the width and height [$w$, $h$] of the most significant local area. Moveover, $w$ and $h$ are the optimizable parameters based on technique of differentiable image sampling.

The main contributions of this paper can be summarized as the following three aspects:

1. We rethink the abnormality detection in musculoskeletal radiographs from the point of global and local feature extraction and fusion, instead of only the global features.
2. In order to extract more significant local features from the area most likely to be abnormal, the key area localization mechanism (KALM) is specially designed for abonormlity detection in musculoskeletal radiographs in this paper.
3. The comparative experiments based on several classical CNNs including Resnet18 [8], VGG16 [9] and InceptionV4 [10] are conducted on MURA [1], which is the largest dataset of musculoskeletal radiographs. And the results demonstrate the effectiveness of our method.

## 2. RELATED WORK

### 2.1. Abnormality Detection in Medical Images

Many researchers have explored the solutions for abnormality detection in medical images. Mark Cicero *et al.* [11] train the GoogleNet to detect abnormalities in frontal chest radiographs. For pulmonary tuberculosis at chest radiography, Paras et.al [12] attempt some shadow convolutional networks and DCNNs. And several classical CNN architectures are applied in mammography abnormality detection [13]. What's more, Rajpurkar *et al.* [1] release MURA, which is the largest radiologist-level abnormality dataset in musucloskeletal radiographs even all kinds of radiograhs, and use DenseNet-121 to detect and show the abnormalities. In these literature, deep learning methods, especially CNN mothods, are widely used for abnormality detection and achieve excellent results. However, these CNN-based methods are just directly applied in the field of abnormality detection in medical images without taking into consideration the joint global and local feature extraction and fusion, which is beneficial for the improvement of classification performance.

### 2.2. Weakly Supervised Object Localization

In the field of natural image processing, weakly supervised object localization technique is the core operation for searching the most important area. Max *et al.* [14] propose differentiable image sampling for the first time, which allows gradient forward and backward propagation in the image sub-sampling. Attention maps are critical for object detection [15], and Diba *et al.* [16] utilize the attention maps to produce high-response region proposals. Recurrent attention convolutional neural network (RA-CNN), which can learn effective region attention and multi-scale region-based feature representation, is proposed in [17] for fine-tuning classification. Qian *et al.* [18] introduce metric learning into the domain of weakly supervised scene parsing. In musculoskeletal radiographs, the abnormalities mostly only occupy a small area in the whole image. Therefore it is meaningful to localize and crop the most important area for the more accurate judgement.

## 3. METHOD

As shown in Fig. 1, the whole abnormality detection framework can be roughly divided into four ordered steps: ***Step I:*** Global-Area Based Abnormality Detection. It is achieved by a convolutional neural network denoted as $CNN_g$, which takes the entire radiograph image as input and gives a preliminary detection result. ***Step II:*** Key Area Localization Mechanism (KALM). KALM is used to locate the coordinate center of the key area denoted as [$x$, $y$], and predict its width and height denoted as [$w$, $h$], based on the multi-layer feature map extracted in the last step. ***Step III:*** Differentiable Image Sampling. In order to ensure that the gradient can be propagated backward to [$w$, $h$], sampling in the global images needs to be differentiable. ***Step IV:*** Local-Area Based Abnormality Detection. It is realized by $CNN_l$, which takes as input the key area sampled from the whole image and provides another detection result.

### 3.1. Global-Area Based Abnormality Detection

The abnormality detection in musculoskeletal radiographs is essentially a binary classification task. In this paper, we use convolutional blocks $conv_g$ to extract the global features denoted as multi-layer feature map, $F_g \in \mathbb{R}^{C\times H\times W}$ ($C$, $H$ and $W$ correspond to channel numbers, height and width, respectively), from the entire radiograph image $I_g$. It is formulated as:

$$F_g = conv_g(I_g), \tag{1}$$

and the global-area classifcation score, $S_g \in \mathbb{R}^2$ (abnormality/normality), is calculated by:

$$V_g(c) = \frac{1}{HW}\sum_{i=0}^{H}\sum_{j=0}^{W} F_g(c,i,j), \tag{2}$$

$$S_g = softmax(FC_g(V_g))), \tag{3}$$

here (2) is the global average pooling (GAP). $V_g \in \mathbb{R}^C$ is the high-dimension feature vector derived from $F_g$ by GAP. (3) is the classifier of fully connected (FC) layer followed by a *softmax* operation. The cascade of (1), (2) and (3) is the aforementioned $CNN_g$.

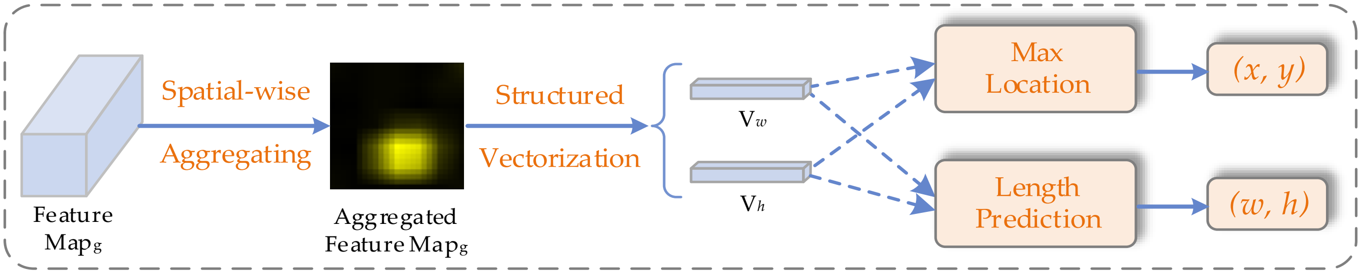

**Fig. 2**. The workflow of KALM.

### 3.2. KALM: Key Area Localization Mechanism

The workflow of key area localization mechanism (KALM) is shown in Fig. 2. KALM takes as input the global multi-layer feature map $F_g$ extracted by $conv_g$ in (1), and provides the bounding box of the key area for the next step: [*x*, *y*, *w*, *h*]. KALM can be splited into the following sub-steps.

**(1) Spatial-Wise Aggregating**. It is the prerequisite for area localization to quantitatively evaluate the importance of each patch in the $F_g$. Refering to [19], we sum the multi-layer feature map $F_g$ along the channel axis as:

$$F_{agr} = \sum_{i=0}^{C} F_g(i, H, W), \tag{4}$$

here $F_{agr} \in \mathbb{R}^{H \times W}$ is the aggregated feature map, which represents the spatial-wise response degree of features. $F_{agr}(i, j)$, the value at location (*i, j*), increases with the importance of the patch.

**(2) Structured Vectorization**. Motivated by GAP which brings CNNs translation invariance in space, we pool the 2-D $F_{agr}$ into two 1-D vectors of $V_w \in \mathbb{R}^W$ and $V_h \in \mathbb{R}^H$ along its spatial height and width respectively as:

$$\begin{cases} V_w = \sum\limits_{i=0}^{H} F_{agr}(i, W), \\ V_h = \sum\limits_{i=0}^{W} F_{agr}(H, i), \end{cases} \tag{5}$$

Then the elements of $V_w$ and $V_h$ are scaled to [0, 1] by min-max scaling as:

$$\begin{cases} V_w(i) = \frac{V_w(i) - min(V_w)}{max(V_w) - min(V_w)}, \\ V_h(i) = \frac{V_h(i) - min(V_h)}{max(V_h) - min(V_h)}, \end{cases} \tag{6}$$

here $min(V)$ and $max(V)$ represent the minimum and maximum value in the structured feature vetor $V$. $V_w$ and $V_h$ can be seen as the structured information of $F_g$ along the axis of spatial width and height.

**(3) Max Location and Length Preediction**. Next $V_w$ is used to locate the column center *x* and predict the width *w*, while the $V_h$ is used to locate the row center *y* and predict the height *h*. They are formulated as:

$$\begin{cases} x = \frac{index(max(V_w))}{W} \times 2 - 1, \\ y = \frac{index(max(V_h))}{H} \times 2 - 1, \\ w = sigmoid(FC_w(V_w)), \\ h = sigmoid(FC_h(V_h)), \end{cases} \tag{7}$$

here [*x*, *y*] are the center coordinate of the maximum response element in [$V_w$, $V_h$], and they are scaled to [-1, 1] for the next image sampling. [*w*, *h*] are predicted from [$V_w$, $V_h$] by FC layer, followed by *sigmoid* activation function, and therefore [*w*, *h*] are in the range of [0, 1]. It worth mentioning that the location models of [*x*, *y*] are non-parametric while the length prediction models of [*w*, *h*] are parametric. The FC layers of [*w*, *h*] need to be optimized by gradient backpropagation algorithm at the training stage.

### 3.3. Differentiable Image Sampling

When applying differentiable image sampling technology, the key local area $I_l \in \mathbb{R}^{3 \times H_i \times W_i}$ can be cropped from the global image, $I_g \in \mathbb{R}^{3 \times H_i \times W_i}$ ($H_i \times W_i$ are the input image zise), allowing the gradient backpropogation.

It is the prerequisite to establish the cropping projection from input coordinate to output coordinate, which is formulated as:

$$\begin{pmatrix} x_i^g \\ y_i^g \end{pmatrix} = \begin{bmatrix} w & 0 & x \\ 0 & h & y \end{bmatrix} \begin{pmatrix} x_i^l \\ y_i^l \\ 1 \end{pmatrix} \tag{8}$$

here $[x, y, w, h]$ is the scaled bounding box of the most important area in the global image $I_g$. $[x_i^l, y_i^l]$ is the pixel coordinate of the cropped local area $I_l^{'} \in \mathbb{R}^{3 \times H_i^{'} \times W_i^{'}}$, while $[x_i^g, y_i^g]$ is the pixel coordinate of global image $I_g$.

In order to make the input image size of local feature extractor consistent, it is necessary to sample $I_l^{'} \in \mathbb{R}^{3 \times H_i^{'} \times W_i^{'}}$ to $I_l \in \mathbb{R}^{3 \times H_i \times W_i}$. In this paper we apply bilinear sampling [14] to achieve it, which is denoted as:

$$I_l = bilinear(I_l^{'}) \tag{9}$$

### 3.4. Local-Area Based Abnormality Detection

Local-area based abnormality detection is realized by another convlutional network $CNN_l$. $CNN_l$ has the same architecture as $CNN_g$ but does not share the parameters with it because the scales of their input image are different. The local-area classification result $S_l \in \mathbb{R}^2$ is calculated as:

$$S_l = CNN_l(I_l) \tag{10}$$

## 4. EXPERIMENTS

### 4.1. Experiments Setup

***Dataset***: MURA is the largest abnormality detection dataset of multi-view musculoskeletal radiographs, containing 9,045 normal and 5,818 abnormal musculoskeletal radiographic studies of humerus, elbow, forearm, wrist, shoulder, hand and finger. There are 13,457 training studies (11,184 patients, 36,808 images) and 1,199 validation studies (783 patients, 3,197 images).

***Evulation Metric***: Classification accuracy, area under the Receiver Operating Characteristic curve (AUROC) and Cohen's Kappa Statistic are used in this paper. We train and save the models acccroding to the image-level accuracy, and evaluate the study-level performance by all the above metrics.

***Training Settings***: We apply two-stage trainging strategy to optimize the global-area and local-area based abnormality detection models respectively: At the first stage, only the parameters of $CNN_g$ are optimized by cross-entropy loss. At the second stage, the parameters of two length prediction layers ($FC_w$ and $FC_h$) and $CNN_l$ are optimized together by cross-entropy loss. Adam is used to optimize the models for 30 epochs with the batch size of 64. The learning rate is set as 1e-5 in the first 20 epochs and 1e-6 in the last 10 epochs. The input sizes of $CNN_g$ and $CNN_l$ are both 224 x 224, but it worth mentioning that the input of $CNN_l$ is cropped from the image of 448 × 448 and then is resized to 224 × 224.

### 4.2. Experimental results

**Table 1**. Comparative results of different CNNs on MURA.

| | *Accuracy* | *AUROC* | *Kappa* |
|---|---|---|---|
| Resnet18$_g$ | 82.1% | 0.875 | 0.633 |
| Resnet18$_l$ | 81.1% | 0.870 | 0.609 |
| Resnet18$_{KALM}$ | **83.2%** | **0.888** | **0.653** |
| InceptionV4$_g$ | 80.0% | 0.865 | 0.584 |
| InceptionV4$_l$ | 81.3% | 0.873 | 0.616 |
| InceptionV4$_{KALM}$ | **81.4%** | **0.881** | **0.617** |
| VGG16$_g$ | 83.8% | 0.892 | 0.668 |
| VGG16$_l$ | 82.7% | 0.895 | 0.649 |
| VGG16$_{KALM}$ | **84.2%** | **0.901** | **0.678** |

To verify the effectiveness of our KALM in abnormality detection, comparative experiments based on three types of classical CNN architectures including Resnet18, InceptionV4 and VGG16 are conducted on MURA. We compare the following three kinds of results by the aforementioned metrics: only global-area detection score (subscript *g*), only local-area detection score (subscript *l*) and their average score based on KALM (subscript *KALM*). The results are reported in Table 1. According to Table 1, it can be found that only the local-area is used for abnormality detection performs no better than the global-area. But when their results are averaged, the fused scores become much better that each of them. The results can strongly support the effectiveness of our KALM.

We also compare our KALM based multi-scale method with other literature in Table 2. Our method outperforms the MobileNet and Ensemble200 [20] in all the metrics, and is superior to DNN [21] in accuracy. In [22], Dense-169 is bigger and deeper than our VGG16$_{KALM}$, and there are 13,942 studies used for training, which are larger than 13,457 studies in this paper. However, our VGG16$_{KALM}$ just slightly falls behind Dense-169 in AUROC. There are some examples of KALM shown in Fig. 3.

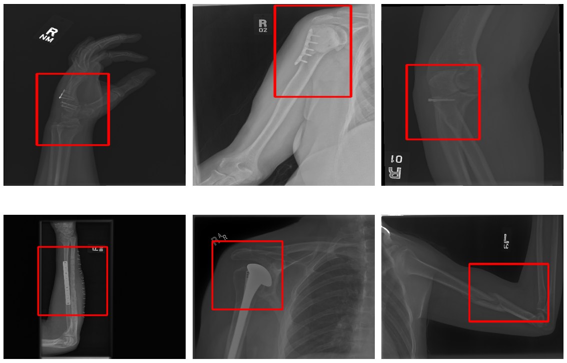

**Fig. 3**. Some samples of KALM based on VGG16.

**Table 2**. Comparison with other literature.

| | *Accuracy* | *AUROC* | *Kappa* |
|---|---|---|---|
| MobileNet [20] | 77.3% | 0.67 | 0.34 |
| Ensemble200 [20] | 79.7% | 0.82 | 0.66 |
| DNN [21] | 82.7% | – | – |
| DenseNet-169 [22] | – | **0.91** | – |
| VGG16$_{KALM}$ | **84.2%** | 0.901 | **0.678** |

## 5. CONCLUSION

In this paper, we propose a key area localization mechanism (KALM) for abnormality detection in musculoskeletal radiographs. And the KALM based multi-scale abnormality detection method is attempted for the first time. The comparative experiments on several classical CNNs are conducted on the largest dataset of MURA, and the excellent results demonstrate the effectiveness of the proposed KALM.

## 6. REFERENCES

[1] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H.l Mehta, B. Yang, K. Zhu, D. Laird, R. L Ball, et al., "Mura: Large dataset for abnormality detection in musculoskeletal radiographs," *arXiv preprint arXiv:1712.06957*, 2017.

[2] A. Mahbod, G. Schaefer, C. Wang, R. Ecker, and I. Ellinge, "Skin lesion classification using hybrid deep neural networks," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1229–1233.

[3] F. Li, H. Huang, Y. Wu, C. Cai, Y. Huang, and X. Ding, "Lung nodule detection with a 3d convnet via iou self-normalization and maxout unit," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1214–1218.

[4] 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.

[5] M. L. Giger, K. Doi, C. E. Metz, and F. Yin, "Automated method and system for the detection and classification of abnormal lesions and parenchymal distortions in digital medical images," 1992, US Patent 5,133,020.

[6] W. Huang, Q. Wang, and X. Li, "Feature sparsity in convolutional neural networks for scene classification of remote sensing image," in *IEEE International Geoscience and Remote Sensing Symposium*, 2019.

[7] Yuan Yuan, Zhitong Xiong, and Qi Wang, "Vssa-net: vertical spatial sequence attention network for traffic sign detection," *IEEE transactions on image processing*, vol. 28, no. 7, pp. 3423–3434, 2019.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[10] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[11] M. Cicero, A. Bilbily, E. Colak, T. Dowdell, B. Gray, K. Perampaladas, and J. Barfett, "Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs," *Investigative radiology*, vol. 52, no. 5, pp. 281–287, 2017.

[12] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, 2017.

[13] Pengcheng Xi, Chang Shu, and Rafik Goubran, "Abnormality detection in mammography using deep convolutional neural networks," in *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 2018, pp. 1–6.

[14] M. Jaderberg, K. Simonyan, A. Zisserman, et al., "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[15] Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Macro-micro adversarial network for human parsing," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 418–434.

[16] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool, "Weakly supervised cascaded convolutional networks," in *IEEE conference on computer vision and pattern recognition*, 2017, pp. 914–922.

[17] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *IEEE conference on computer vision and pattern recognition*, 2017, pp. 4438–4446.

[18] R. Qian, Y. Wei, H. Shi, J. Li, J. Liu, and T. Huang, "Weakly supervised scene parsing with point-based distance metric learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 8843–8850.

[19] Y. Yuan, Z. Xiong, and Q. Wang, "Acm: Adaptive cross-modal graph convolutional neural networks for rgb-d scene recognition," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[20] Dennis Banga and Peter Waiganjo, "Abnormality detection in musculoskeletal radiographs with convolutional neural networks (ensembles) and performance optimization," *arXiv preprint arXiv:1908.02170*, 2019.

[21] Sathiesh Kumar Kaliyugarasan, "Deep transfer learning in medical imaging," M.S. thesis, The University of Bergen, 2019.

[22] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," *arXiv preprint arXiv:1909.12475*, 2019.