Robust MSFM Learning Network for Classification and Weakly Supervised Localization

Komal Kumar[1], Balakrishna Pailla[2], Kalyan Tadepalli[2,3], Sudipta Roy[1,*]
[1]Artificial Intelligence & Data Science, Jio Institute, Navi Mumbai-410206, India
[2]Reliance Jio - Artificial Intelligence Centre of Excellence (AICoE), Hyderabad, India.
[3]Sir HN Reliance Foundation Hospital, Mumbai – 400004, India
Komal2.Kumar@jioinstitute.edu.in; Balakrishna.Pailla@ril.com;
Kalyan.Tadepalli@rfhospital.org; sudipta1.roy@jioinstitute.edu.in;

Abstract

Robust classification and localization of bone fractures are beneficial to avoid misdiagnosis or underdiagnosis. However, state-of-the-art classification methods aim to improve accuracy which lacks reliability, and tackled localization problems in a supervised manner with much-annotated data that leads to high costs. In this paper, we propose a multistage feature map (MSFM) learning network to predict the class of the image and the area of interest without annotated bounded box. MSFM consists of three stages to predict the representation with different objectives and aims to improve the accuracy and reliability of classification. The weakly supervised MSFM model localizes the region of interest (ROI) by taking representation from all the stages supervised by image-level labels only. We also introduced a feature augmentation technique to enforce the model to consider other discriminative regions. End-to-end training of MSFM is performed jointly at all stages. Based on the comprehensive experiments, our approach achieves state-of-the-art results on the standard MURA dataset, which includes the elbow, finger, forearm, humerus, shoulder, wrist, hand, and bone tumor dataset. Code: github.com/MAXNORM8650/MSFM.

1. Introduction

Medical image analysis has been an active research area in recent years, intending to develop automated systems that can assist radiologists in the diagnosis and treatment of various diseases. X-ray imaging is one of the most commonly used imaging modalities in clinical practice as it provides a quick and cost-effective way of detecting and monitoring various pulmonary diseases.

How to tell an X-ray abnormality? How to detect that abnormality? These are challenging tasks even for the average person and usually require domain expertise. Detecting musculoskeletal (MU) abnormalities are particularly important because 1.7 billion people globally suffer from musculoskeletal disorders [18].

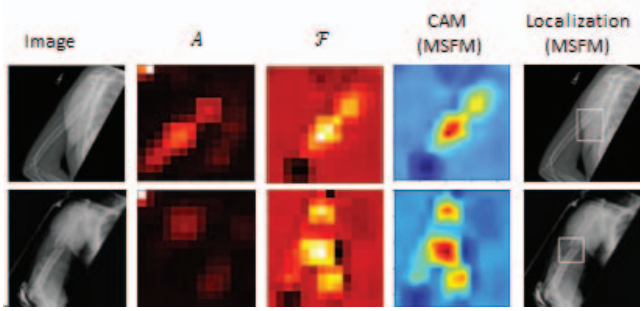

Figure 1: This image shows components of the Class activation map (CAM), activation map (A), and weighted feature map (F) from the first stage classifier of the MSFM model.

Disease localization in X-ray images is a challenging task due to the complexity and variability of the human anatomy and the presence of various disease patterns that can be difficult to identify. Traditional supervised learning methods require a large amount of labeled data [16, 19], which can be time-consuming and expensive to acquire, particularly for rare diseases or conditions that require specialized expertise. To address this issue, weakly supervised learning has emerged as a promising approach for disease localization in X-ray images. Weakly supervised learning leverages the availability of image-level labels, which indicate the presence or absence of a disease in an image, to learn a model that can automatically localize the disease regions within the image. Previous work by the authors of the paper [26] proposed a weakly supervised learning approach for thoracic disease classification and localization using X-ray. They used a convolutional neural network (CNN) [23]

with a localization layer to predict the disease regions in the images. A multi-instance deep learning framework [27] is used for body part recognition in medical images. They used a CNN with attention mechanisms to identify the discriminative local anatomy in the images and attention-based multiple instance learning [9] utilize the attention mechanism to highlight the disease regions in the images. [21] employed a CNN with a pooling layer for feature extraction. This approach, rooted in multi-instance learning, allowed the model to predict disease presence in individual patches and aggregate these predictions to achieve accurate disease localization results. Recent work on WSL by [13] used a CNN with an attention mechanism to learn the disease regions in the images and a segmentation network to generate the semantic segmentation priors in the medical image. These studies have shown the potential of weakly supervised learning for disease localization in medical images. Many works [24, 3, 17, 30] have shown the potential in increasing the robustness of the models using multi-stage learning. Class Activation Mapping (CAM) [32] is also gained a lot of interest in WSL for medical images [28], but CAM-based localization methods lack robust data variability, unclearness to the localization results, and may not capture complex disease patterns. Recent studies have proposed new weakly supervised learning methods that use attention mechanisms, multiple instance learning, or semantic segmentation priors to improve disease localization performance. In this work, we present a multi-stage feature map (MSFM), a novel weakly supervised learning approach for disease localization in X-ray images. MSFM is based on two main components in CAM, feature map (map before global average pooling in CNN network) based activation map (A) and fully connected (FC) layers' weight-based class feature map (F). A and F, as shown in Figure 1, consist of more information related to the object than trivial CAM as it not only learns to localize the most discriminating features but also contains full information about the object. MSFM consists of multiple stages which make it more robust to the variations in the image for disease localization and generalize CAM for the entire object as shown in Figure 1. Based on our experiments we observed that MSFM is biased towards artifacts if the training data contained a large number of artifacts or if the artifacts are highly correlated with the disease patterns. To solve this problem, we also proposed a feature augmentation method to remove the artifacts based on high intensity. Overall, our proposed approach has the potential to improve the accuracy and efficiency of disease localization in X-ray images, which can ultimately benefit patient care and outcomes. The main contributions of our work are as follows:

- Proposed a multistage model (MSFM-net) architecture consisting of three stages that aim to improve the accuracy and reliability of image classification.
- Opportunistic fracture localization in a weakly supervised manner which does not require a bounded box for training.
- An Augmentation technique that uses the feature map to calculate the magnitude of the high-intensity values and remove them.
- On top of that, we conducted a comparative analysis of loss functions to resolve class imbalances as well as a comprehensive evaluation of a large dataset of X-ray images with various abnormalities and a benchmark dataset for localization. Our proposed model achieves excellent classification results with exceptional detection visualization.

2. Methodology

2.1. Class activation map decomposition

If we have an image X of size $C \times H \times W$, we would like to have a representation that consists of approximately all the image's information to classify it. Typically, a neural network comprises convolutional layers followed by the average pooling, and a fully connected layer for classification is used to compute the CAM as follows:

$$CAM(X) = W_{cl}^{T} F(X). \tag{1}$$

Where $F : \mathbb{R}^{C\times H\times W} \implies \mathbb{R}^{n\times h\times w}$ represents features map before average pooling for (h, w) spatial dimension of the n channels. $W_{cl} \in \mathbb{R}$ are the weights of the FC layer corresponding to the target class cl. Authors from [10] try to bridge the gap between the classification and localization by decomposing CAM in terms of a cosine similarity map as follows:

$$CAM\,(X) = \|W_{cl}\|F(X)\|cos\theta \leq \|W_{cl}\|\|F(X)\| \tag{2}$$

Where $cos\theta$ is the cosine similarity between two vectors whose large value represents the less degree of alignment between the vectors. We define a weighted feature space (F) map which corresponds to every class of target as follows:

$$F = \|W \cdot F_A(X)\|, \tag{3}$$

Where $W \in \mathbb{R}^{nclass\times n} = [W_1, \ldots W_{cl}, \ldots W_{nclass}]$ and norm is taken for all the classes. Based on Figure 1, CAM alone cannot localize the full informative object corresponding to its class level as it learns the difference between the classes which leads to poor localization. However, normed feature map $\|F(X)\|$ and weighted norm feature map F contain more information to localize the object corresponding to its class level. The object can be localized based on F and activation map A as shown in Figure

1 where $A = \|F(X)\|$. If the position (i, j) in A is higher than the mean of activation map $(\bar{A})$ is part of the object which we need to localize for $\forall i \in [0, h]$, and $\forall j \in [0, w]$. Mathematically:

$$\bar{A} = \frac{\sum_{i=0}^{h-1} \sum_{j=0}^{w-1} A(i,j)}{h \times w} \tag{4}$$

$$\widehat{M_{(i,j)}} = \begin{cases} 1, & \text{if } A(i,j) \geq \bar{A} \\ 0, & \text{if } A(i,j) < \bar{A} \end{cases} \tag{5}$$

Where $\widehat{M_{(i,j)}}$ is the possible area of the object to localize. The final possible area is based on the intersection of the area obtained by the ensemble of F and A. When the informative region is localized, the cropped image contains additional information that can provide additional insights into the image by looking at closure using localized images. It does not require any additional parameters as it is based on the trained classifier model. By observing Figure 1, the area with the higher value of is the area where the key parts are located, most of the time CAM indicates the joints of the bone which may be incorrect for tumor identification. We use a technique that involves dividing the image into overlapping windows, and then classifying each window as a foreground (marked as 1) or background (marked as 0) using the equation 5. The overlapping windows are then moved across the image, allowing the classifier to process multiple regions of the image in a sliding manner.

2.2. MSFM learning network.

The architecture MSFM model is shown in Figure 2. We have used a Resnet [6] as the encoder (En) to encode the image into a feature representation. Our main model consists of main three stages that include: The main stage, the Object stage, and the Parts stage.

- **Main stage**: We encode the image using En to get the feature representation map. A coordinate of the bounding box is generated using the intersection of the informative region by feature map (FM) and Class feature map (CFM) which is shown in Figure 2. In this stage, the full feature is used for classification using the FC layer to produce CFM. Furthermore, the hypothetical object using the box generated by the second stage is cropped from the image and passed to the object stage.

- **Object stage**: Cropped image is used to get its feature representation map by shared parameter from En of the main stage (first stage). Then it is passed into the second stage to get the discriminative regions. Hypothetically that regions cover the informative part.

- **Parts and multi-scale stage**: Local proposed region from the object stage are cropped. Now different windows are proposed based on some scale to produce multiple parts. These parts (w_n's) are supervised by the label of corresponding classes. Finally, these parts are combined to produce one possible box for fracture.

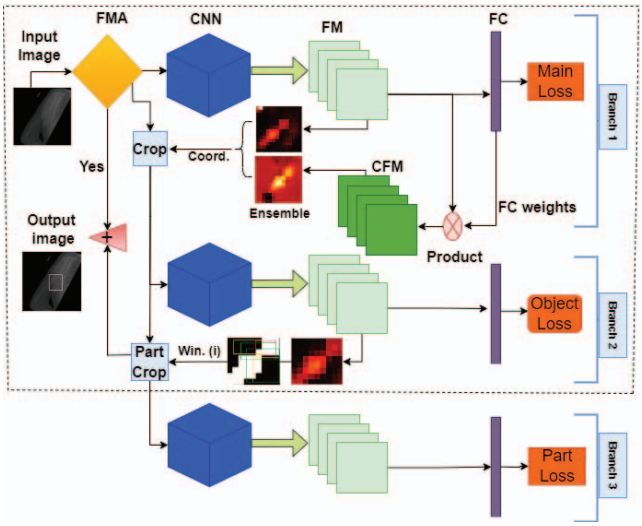

Figure 2: The full MSFM architecture of the multi-stage network that consists of three stages starting with CNN in the training phase and dotted black box is the structure in the test phase. The CFM represents the product of Fully Connected (FC) weights and FM. "Yes" represent the condition to the output for a specific class (fractured image in our case), and '+' in the output image is to paste the combined part in the original image. The CNN and FC layers of the same color represent parameter sharing.

For each image in the training set, we optimized the overall loss for all stages using three types of loss function by sharing the parameter shown by the same color in Figure 2. Due to the class imbalance problem, we have used weighted cross entropy that is defined in the equation 6.

$$CE_t^j(P(\theta), Q) = -\sum_{p_i \in P} \sum_{j \in T} w_+^j q_i \ln \mathbb{P}(q_i = 1|p_i) - \sum_{q_i \in Q} \sum_{j \in T} w_-^j q_i \ln \mathbb{P}(q_i = 0|p_i) \tag{6}$$

Where, $P(\theta) = \{p_1, p_2, \ldots, p_n\}$ are the predicted labels, $Q = \{q_1, q_2, \ldots, q_n\}$ are the ground truth labels for the corresponding instances in $P(\theta)$, $\mathbb{P}(q_i = 1|p_i)$ is the predictive probability $\forall q_i \in \{0, 1\}$ conditioned on$P(\theta)$, T is set of all the parts in the MURA data, w_+^j is the weight for all positive class of part type $j \in T$, and w_-^j is the weight for the negative class of part type $j \in T$. To focus on misclassification due to the class imbalance problem, we have also used the focal loss [14] of the second type which is defined in equation 7.

$$FL_t(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left(1 - e^{-CE_t^i(\theta)}\right)^{\beta} CE_t^i(\theta) \tag{7}$$

Where $CE_t(\theta)$ is cross entropy att^{th} step with θ shared parameter defined in equation 6, β is the focal loss hyper-parameter. It is clear from the equation 7, to get the loss for the t^{th} step we need to run in the entire mini-batch size to get the mean which increases the little computation. So instead of mean, we have used a scaler α. Rewriting equation 7 as,

$$FL_t(\theta) = \alpha\left(1 - e^{-CE_t(\theta)}\right)^{\beta} CE_t(\theta) \tag{8}$$

So, the total loss is the sum of the loss of all stages which is shown in the equation 8.

$$L_t^{total}(\theta) = FL_t^M(\theta) + FL_t^O(\theta) + \sum_{i=1}^{w_n} FL_t^{P_i}(\theta) \tag{9}$$

2.3. Class Feature Map

For c class classification problem from an image, we are interested in the c different feature map to get the image representation concerning output weights for the corresponding class level. Inspired by TS-CAM [5], we use CFM to get a better feature map of interest from the feature map of the last layer and the output attention weight.

$$CFM = W \times F \tag{10}$$

Where $F \in {}^{n\times h\times w}$ is feature and $W \in \mathbb{R}^{c\times h\times w} \times \mathbb{R}^{n\times h\times w}$ is a weights of output layer.

2.4. Feature Localization and Amplification

The class feature map (CFM in equation 10) of the full image is binarized based on the equation using mean thresholding value as shown in equation 5. The pixels are connected according to their neighboring values if they are equal in value when pixels are mounted in a binary map. In this case, it refers to how many orthogonal hops a pixel must undergo to be considered a neighbor that will return all connected regions that are assigned the same value. We also find the area based on the feature map and select the intersecting region to produce a bounding box if the interesting area is zero, we assign a default bounding box of w width and h height (same as the spatial dimension of each feature). Finally, we selected the region that covers the maximum activation area to look closer as shown in Figure 3(a) for the localization of informative region within X-ray with bilinear upsampling method.

2.5. Informative regions localization

Although the cropped local image contained the informative region with good probability, the idea is to localize the key part of the image. Based on the feature map, we search for the areas with higher activation (A), which indicates the location of key parts in the local image (cropped image). So, we extract the feature map of the cropped image

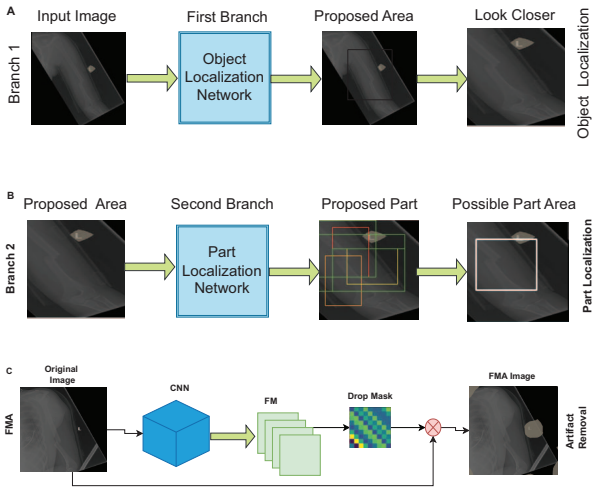

Figure 3: (a) This network trains to localize the key area in a weakly supervised manner. The informative area is represented by the bounded box. (b) This pipeline is based on the second stage of MSFM. Here red, orange, yellow, and green colors indicate the order of the proposed window and the white window indicates the final window by combining all the windows using discounted factor method in the same order. (c) This figure represents the feature map-based augmentation. CNN of different colors represents the different pre-train CNN for removing the high-intensity values based on the drop mask that produces FMA with the product of the original image.

by sharing the En's parameter to obtain the activation map (A) for the selected region (the region of the window with height h_w, and width w_w) and calculate the score by Average pooling with Kernal size of $(h_w,\ w_w)$. Window size $(h_w,\ w_w)$ is a hyperparameter to tune the different types of problems. The basic idea for selecting the window size is to cover as different parts as possible (see Figure 3(b)). Localization is based on the binary map obtained from the activation map of each window by mean thresholding $(\bar{A}_w)$ for the window as defined in equation 4. Non-Maximum Suppression (NMS) [7] is applied after scoring to select the fixed number of parts in images so that there are fewer redundant parts in each region.

2.6. Feature Map-based Augmentation

In our experiment, the model was initially biased toward the artifacts or some high-intensity values present in the training images due to its weakly supervised nature. Data augmentation helps to increase the variety of relevant training data and helps to improve the performance of the model but random erasing in the training data may remove the informative region instead of removing the non-informative region. Inspired by SWS-DAN [29], we augment the image based on the feature map pre-trained on imagenet21k which represents the high-intensity value. The presence and inten-

sity of artifacts (e.g., letters) in the Mura dataset are high, so we reduced the intensity of artifacts by percentage (p) before feeding into the network. As shown in Figure 3(c), all the highly active regions are then thresholded which makes the model no more biased toward artifact (See Figure 10).

2.7. Abnormality detection

From the informative regions localization section, the larger the value of $\bar{a}_w$, the larger the information the part contains. We combine w windows to detect the fracture by discounting the value window for decreasing the value of $\bar{a}_w$ after shorting. The final window value (U) is given by:

$$U\left(x,\ y,\ w_w,\ h_w\right) = \sum_{i=1}^{w_n} \gamma^i\, U\left(x_i,\ y_i,\ w_{wi},\ h_{wi}\right) \quad (11)$$

Where $\gamma \in \mathbb{R}^{(0,1]}$ are the weights assigmn to the windows, and for w_n number of the proposed windows, $U\left(x_i,\ y_i,\ x_i + w_{wi},\ y_i + h_{wi}\right)$ is sizes of the windows $\forall\, i \in [1, w_n]$.

3. Experiments

3.1. Experimental Setting

Dataset Description: For training and testing of the model, we have used MURA dataset [18], a large dataset of a bone X-ray image that consists of seven parts elbow, finger, forearm, hand, humerus, shoulder, and wrist. The task in the dataset is a binary class problem to find the positive and negative classes. The size of each image in the dataset is different and varies from 117×512 to 512×512. The presence of an imbalanced class and the inconsistency in the size of the image, make MURA dataset classification a challenging problem. The MURA dataset contains musculoskeletal imaging studies from $14,863$ patients and is tagged by certified radiologists as positive or negative. There are one or more views in each study that are labeled as 0 for positive or 1 for negative. The training dataset consists of $36,812$ images, and the testing dataset consists of a total of $3,197$ images. A summary of all parts in the MURA dataset is shown in Figure 4 for training and testing the models. The authors have made MURA dataset freely available at http://stanfordmlgroup.github.io/competitions/mura. Furthermore, Tumor detection can be a problem in X-ray images, especially if the tumor is small or located in an area with complex overlapping structures. To work on this, We have also collected bone Tumor $1,100$ X-ray images to check the performance of the model on it. We divide this dataset by $70\% - 30\%$ for training and testing of the model with $1,200$ images of normal bone inference data to make a binary classification problem.

Implementation details : We pre-processed the image to size 448×448 to get the image for augmentation, for the first stage, and second stage, as shown in Figure 2. After cropping the original image with the first stage, input in the second stage is also scaled to 448×448. All images are reshaped to 224×224 for the part stage. For the augmentation based on the feature map, the percentage reduction of the high pixel we have used is 0.15, and the threshold value we have used is 0.6 based on our experiments. As discussed, we select the window with broad categories of scale: $\{[6 \times 6,\ 7 \times 5],\ [8 \times 8,\ 6 \times 10,\ 7 \times 9],\ [10 \times 10,\ 9 \times 11,\ 8 \times 12]\}$. The $h \times w$ is 14×14 which is the size of activation map A, and the number of the image's part images is $w_n = 7$, where $w_{1n} = 2$, $w_{2n} = 2$, and $w_{3n} = 3$ are number of broad categories of scales. We have also used pre-trained Resnet-50 on the MURA dataset as the backbone to fire the feature map as shown in Figure 2 within the same-colored CNNs. We have also used Resnet-50 pre-trained on ImageNet for augmentation based on the feature map to remove the high-intensity pixel. During the training of the model, we do not use any type of annotation other than image class labels, but annotated data were used in testing to measure the performance. We optimize the loss using SGD [20], initial learning rate 1×10^{-4} and minibatch size of 5 on RTX A400 GPU. We used PyTorch as our codebase.

Baseline: We have used Res-Net as a baseline network for the feature map. First, we trained on the full Mura training dataset (size 33k) to use in the main network where Res-Net is also pre-trained on ImageNet-21k [4]. We used multiple architecture resnet18, resnet34, and resnet50 for ablation.

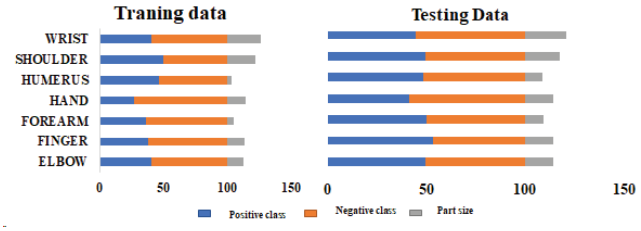

Figure 4: Part size represents the size of the corresponding part concerning the size of the Mura dataset. Positive and Negative classes represent the fracture and non-fracture image split respectively.

3.2. Performance of the models

We have evaluated our model on seven parts of the publicly available largest bone-fracture Mura dataset and compare its accuracy and Cohen's kappa statistic (κ) with the existing baseline classifiers.

Classification performance of the models: We have reported the accuracy of the models for each part in Table 1. Our MSFM models achieve good performance compared to baseline models based on CNNs. We achieved highest accu-

racy of 87%, 81%, 84%, 89%, 80%, 87%, and 79% for each part with MSFM model with resnet-50 backbone. The average accuracy for all parts is 85%. The highest 89% accuracy was achieved in the Humerus part, and the lowest accuracy was achieved in 79% in the Hand part using Resnet-50 as a backbone. More importantly, we have also reported the results of Tumor detection in Table 1. MSFMR50 is outperforming in the Tumor dataset as well as other models.

Comparison based on Cohen's kappa statistic (κ): The performance of classifier models is evaluated through the utilization of statics called Cohen's kappa statistic (κ). We compared all results with Cohen's kappa statistic (κ) [15] which agrees on each model with the gold standard. In [18], authors chose three random radiologists to give the comparison of their performance with the deep learning models based on the gold standard. We have reported the average performance of all three radiologists (RD), and the average performance of the existing models to compare with our models. This is done using Cohen's kappa score shown in Table 2.

Localization interpretation To check the localization performance from the first stage, we have reported the Percentage of Correctly Localized Parts (PCP) metric in Table 3. PCP is the percentage of predicted boxes that are correctly localized with more than 50% IOU with the ground-truth bounding box. The Attention Object Learning Module (AOLM) [30] initially achieved an impressive PCL of 85.1% using two ResNet-50 layers and a pre-trained ImageNet21k backbone. However, the PCL declined to 71.1% during training as the CNN-based network prioritized prominent regions. Our proposed model, combining attention and backbone modules, outperformed recent weakly supervised methods, achieving the highest accuracy of 82.6%. Although the PCL decreased to 74.4% as training progressed, it still surpassed the performance of the other weakly supervised object localization models.

Comparison with CAM : In Figure 5, we compare CAM from first stage of MSFM resnet50 backbone and CAM from training resnet50. CAM highlights the significant area in the image that helps in classification and confirms our model's smooth training. Drawing a bounding box from the CAM (MSFM) produces similar localization results as our proposed method, demonstrating the agreement between CAM (MSFM) and our method. However, CAMs cannot identify the regions responsible for errors in an image, making it challenging to determine the necessary improvements to increase accuracy, as evident in the first and second images of Figure 5. In contrast, our method provides a clear idea of the required bounded box The figure also ensures the robustness of the model because of multi-stage training compared to normal CAM. Our method carefully inspects the image in the first stage, followed by the second stage that decides on windows with a different confidence. These windows have a low probability of being incorrect as the first stage ensures accuracy, and we combine all windows to form the final bounded box.

Fracture result visualization: The image visualization

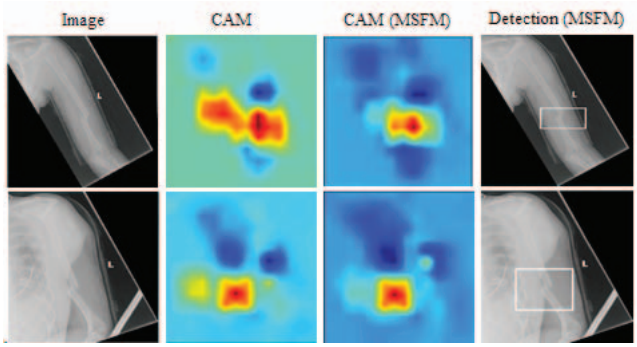

Figure 5: Visual comparison with weakly supervised CAM.

of the model in each step is shown in Figure 6. It consists of different columns namely original image, feature map-based augmentation (FMA), area localization and looking closer as discussed in Main-stage, and the final column. These results are based on the MSFM approach in each part of the Mura datasets. FMA removes the non-informative region shown in the second column. Looking closer is the output of the first stage which is the input of the first stage. Finally, for the fracture class, we have reported the results in the column of abnormality. These results show the potential of MSFM as weakly supervised learning (More results of this dataset are provided in Figure 2 and Figure 3 of supplementary material).

Tumor result visualization: Tumors can be difficult to distinguish from surrounding tissue in X-ray images, as they often appear as subtle changes in density or shape. Tumor detection is shown in Figure 7. These results can be interpreted based on the bouned box which is containing the Tumor. This description provides information about the location of the tumor (left femur), and its appearance on the X-ray images (well-defined radiolucent lesion with a sclerotic margin). This information can help doctors determine the appropriate treatment for the tumor and monitor its progression over time. Hence our approach can be used to detect the severity of abnormality which also shows the generalizability of the model (More results of this dataset are provided in Figure 1 of supplementary material).

3.3. Ablation Studies

Model variants: Comparison based only on accuracy is not the sole indicator, so we have used Joint prediction error (JPE) [11, 12] to check the combined effect of F1-score being fractured or not which is defined as,

$$JPE\ (\gamma_J) = \frac{\left(1 - F_1^A\right)^2}{2} + \frac{\left(1 - F_1^B\right)^2}{2} \qquad (12)$$

Table 1: Comparison based on classification results to baseline models. The highest accuracy of each part is highlighted. MSFM-R-n is our proposed method with the backbone of R=Resnet for n= [18, 34, 50].

Part Name	Elbow	Finger	Forearm	Hand	Humerus	Shoulder	Wrist	Av.	Tumor
Res Net[6]	0.79	0.65	0.73	0.72	0.61	0.75	0.78	0.72	0.94
Dense Net [8]	0.85	0.73	0.58	0.75	0.76	0.75	0.76	0.74	0.76
Inception [22]	0.69	0.61	0.70	0.64	0.71	0.75	0.76	0.70	0.90
APGA [1]	0.83	0.77	0.83	**0.79**	0.86	0.76	0.84	0.81	-
MSFMR50	**0.87**	**0.81**	**0.84**	**0.79**	**0.89**	**0.80**	**0.87**	**0.85**	**0.97**
MSFMR18	0.84	0.73	0.68	0.72	0.85	0.71	0.81	0.77	0.89
MSFMR34	0.80	0.78	0.77	0.79	0.82	0.75	0.82	0.79	0.94

Table 2: Results based on Cohen's kappa statistic (κ) score and comparison to the baseline technique and three unbiased radiologists. RD1, RD2, and RD3 represent three radiologists. The highest Cohen's kappa score in all RDs/models and the models marked as bold. Av. is the average kappa score of all parts.

RDs/models ↓ Parts Name →	Elbow	Finger	Forearm	Humerus	Shoulder	Wrist	Hand	Av.
RD1 [18]	**0.85**	0.30	0.79	0.86	**0.86**	0.79	0.66	0.74
RD2 [18]	0.71	0.40	**0.80**	0.73	0.79	**0.93**	**0.92**	0.72
RD3 [18]	0.71	0.41	0.79	**0.93**	**0.86**	**0.93**	0.78	**0.77**
Dense [18], [8]	0.71	0.38	0.50	0.60	0.72	0.62	0.55	0.58
Res Net [6]	0.59	0.34	0.46	0.22	0.50	0.54	0.39	0.44
Inception [22]	0.39	0.26	0.41	0.43	0.50	0.53	0.17	0.42
MSFMR34	0.61	0.57	0.55	0.65	0.50	0.64	0.54	0.58
MSFMR18	0.68	0.47	0.37	0.71	0.43	0.61	0.38	0.54
MSFMR50	**0.72**	**0.62**	**0.64**	**0.77**	**0.73**	**0.67**	**0.56**	**0.70**

Table 3: Localization performance of the first stage of MSFM on CUB. "Yes" represents the training from scratch.

Methods	Training from scratch	PCL (%)
ACOL [31]	No	46.0
ADL [2]	No	62.3
SCDA [25]	No	76.8
MMAL [30]	No	71.1
MSFM (ours)	Yes	77.4

Where F_1 is the F1 score, γ_J is '0' for the best-performing model, and '1' for the worst-performing model. We compare our models for different Resnet-based feature extractors as a backbone to our model across all the parts of the Mura dataset, both in terms of joint prediction error (JPE) and accuracy. The accuracy for the resnet-18, resnet-34, and resnet-50 as the backbone in MSFM is reported in Table 1 whereas the value of γ_J is reported in Figure 8 through the table and the γ_J graph. MSFM-R50 gives the best results for most of the parts, except Elbow where MSFM-R18 is performing best. The worst performing model is in part Forearm with MSFMR18. MSFMR34 is performing better in the Hand part, than MSFM-R50 and MSFMR18. MSFMR34 and MSFMR50 are outperforming the same in Elbow, Shoulder, and Wrist and MSFMR50 is better than MSFMR34 in another part.

Effect of Loss functions variety : Mura training dataset has the problem of imbalance classes that can be seen in Figure 4. The performance in some parts is not up to the mark due to the high-class imbalance, especially in the hand part where 27.77% is the '+' class and 73.22% for the '−' class. This is one of the reasons models perform worst in the Shoulder and Hand parts of the Mura dataset. To overcome this problem, three types of loss function as described in equations 6 7 8 based on cross-entropy is performed. The weighted cross-entropy (WCE) loss function is shown in equation 6 and CE is the special case by putting equal weightage. Similarly, MFL and $\alpha - FL$, represent the mean focal loss and α focal loss, shown in the equation 7 and equation 8 respectively. From Figure (B), the highest accuracy is achieved using WMFL which is approximately the same for α-FL and α-WFL whereas the lowest accuracy is achieved by CE. Based on Cohen's kappa score, α-WFL is the best-performing loss function for our models. α-WFL is used in all our experiments because of the highest Cohen's kappa score and approximately the same accuracy as the highest performing loss function based on accuracy.

Effect of feature map-based augmentation parameter: FMA has two hyperparameters, one is the thresholding value, and the other is the reduced weight of the pixel value (r). The r $= 0.70$ indicate that the pixel value was reduced

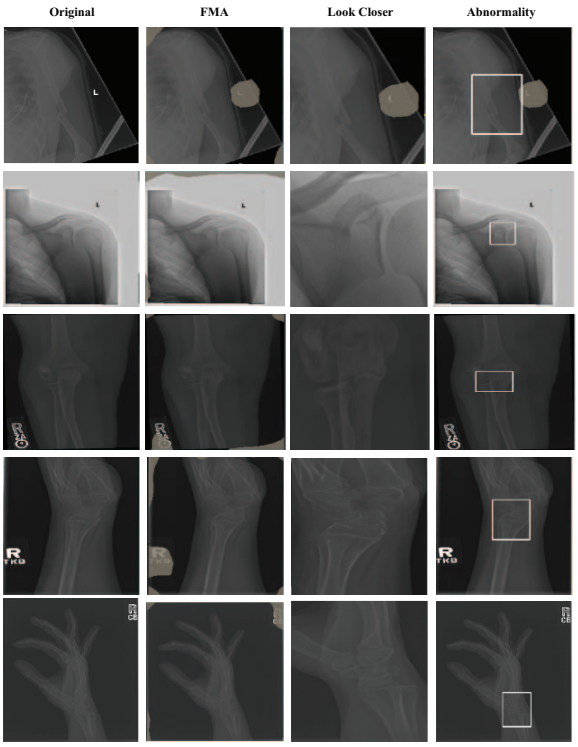

Figure 6: Output from MSFM at each step of the model. The first column consists of an original image from each part followed by the columns of FMA, Look Closer in the original image, and abnormality detection in the original image.

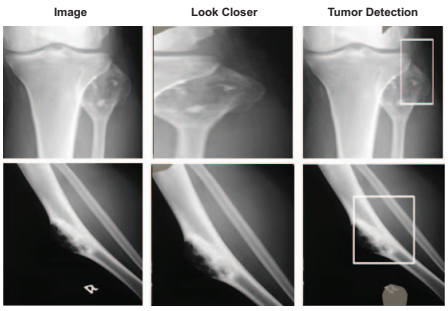

Figure 7: Tumor detection output visulization

by 70% because of the elimination of unnecessary region-like artifacts. Figure 10 (A) shows that when the intensity of the artifact (assuming high-intensity value) is reduced by a reduction rate (r=0.15), the bounding box is free from the artifact, while if the reduction rate (r) is 0.7, the bounding box contains the artifact with the informative region. There is one region for the highest intensity value, but many images are consisting of metal and some more artifacts. One image is shown in Figure 10 (B) and consists of multiple high-intensity regions and the effect of r is also shown for the feature augmented image. As the value of r decreases,

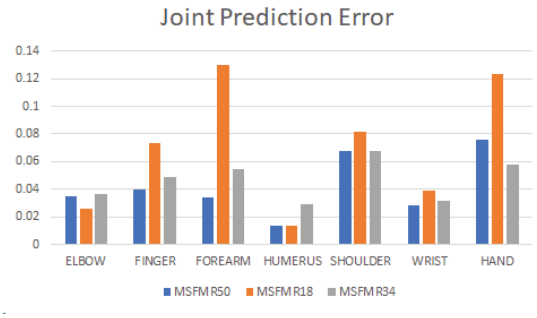

Figure 8: Comparison of MSFM on backbone models using JPE.

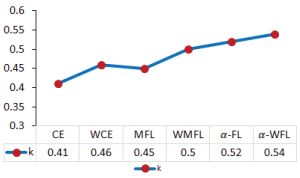

Figure 9: Effect of loss function on Cohen's kappa statistic.

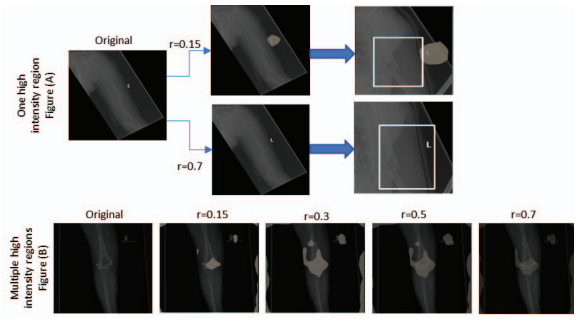

Figure 10: (A) The effect of r in the output bounded box due to artifact. (B) Effect of r on FMA image visualization.

the FMA starts to remove the region from taking away highlighters, increasing the localization accuracy based on our experiments.

4. Conclusion

Our study presents a robust classification and localization approach for informative regions, eliminating the need for costly bounding box annotation. Extensive experiments demonstrate superior fracture classification and localization accuracy compared to baselines. The multistage structure efficiently leverages MSFM images from various stages, contributing to its excellent performance. Future work includes applying the MSFM model to diverse medical and natural image datasets, expanding its scope beyond pre-trained models.

References

[1] Kaiyang Cheng, Claudia Iriondo, Francesco Calivá, Justin Krogue, Sharmila Majumdar, and Valentina Pedoia. Adversarial policy gradient for deep learning image augmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 450–458. Springer, 2019. 7

[2] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2219–2228, 2019. 7

[3] Marcos V Conde and Kerem Turgutlu. Exploring vision transformers for fine-grained classification. *arXiv preprint arXiv:2106.10587*, 2021. 2

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[5] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2886–2895, 2021. 4

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 7

[7] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. Learning non-maximum suppression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4507–4515, 2017. 4

[8] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 7

[9] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 2

[10] Eunji Kim, Siwon Kim, Jungbeom Lee, Hyunwoo Kim, and Sungroh Yoon. Bridging the gap between classification and localization for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14258–14267, 2022. 2

[11] Deepak Kumar, Suraj S Meghwani, and Manoj Thakur. Proximal support vector machine based hybrid prediction models for trend forecasting in financial markets. *Journal of Computational Science*, 17:1–13, 2016. 6

[12] Komal Kumar, Hement Kumar, and Pratishtha Wadhwa. Encoder–decoder (lstm-lstm) network-based prediction model for trend forecasting in currency market. In *Soft Computing for Problem Solving: Proceedings of the SocProS 2022*, pages 211–223. Springer, 2023. 6

[13] Xi Li, Huimin Ma, Sheng Yi, Yanxian Chen, and Hongbing Ma. Single annotated pixel based weakly supervised semantic segmentation under driving scenes. *Pattern Recognition*, 116:107979, 2021. 2

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[15] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012. 6

[16] Tanushree Meena and Sudipta Roy. Bone fracture detection using deep supervised learning from radiological images: A paradigm shift. *Diagnostics*, 12(10):2420, 2022. 1

[17] Qi Qian, Rong Jin, Shenghuo Zhu, and Yuanqing Lin. Fine-grained visual categorization via multi-stage metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

[18] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957*, 2017. 1, 5, 6, 7

[19] Sudipta Roy, Tanushree Meena, and Se-Jung Lim. Demystifying supervised learning in healthcare 4.0: A new reality of transforming diagnostic medicine. *Diagnostics*, 12(10):2549, 2022. 1

[20] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 5

[21] Evan Schwab, André Gooßen, Hrishikesh Deshpande, and Axel Saalbach. Localization of critical findings in chest x-ray without local annotations using multi-instance learning. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1879–1882. IEEE, 2020. 2

[22] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 7

[23] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016. 1

[24] Jiabao Wang, Yang Li, Zhuang Miao, Xun Zhao, and Zhang Rui. Multi-level metric learning network for fine-grained classification. *IEEE Access*, 7:166390–166397, 2019. 2

[25] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE transactions on image processing*, 26(6):2868–2881, 2017. 7

[26] Chaochao Yan, Jiawen Yao, Ruoyu Li, Zheng Xu, and Junzhou Huang. Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 103–110, 2018. 1

[27] Zhennan Yan, Yiqiang Zhan, Shaoting Zhang, Dimitris Metaxas, and Xiang Sean Zhou. Multi-instance multi-stage

deep learning for medical image recognition. In *Deep Learning for Medical Image Analysis*, pages 83–104. Elsevier, 2017. 2

[28] Seunghan Yang, Yoonhyung Kim, Youngeun Kim, and Changick Kim. Combinational class activation maps for weakly supervised object localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2941–2949, 2020. 2

[29] Zhen Yang, Zhipeng Wang, Lingkun Luo, Hongping Gan, and Tao Zhang. Sws-dan: Subtler ws-dan for fine-grained image classification. *Journal of Visual Communication and Image Representation*, 79:103245, 2021. 4

[30] Fan Zhang, Meng Li, Guisheng Zhai, and Yizhao Liu. Multi-branch and multi-scale attention learning for fine-grained visual categorization. In *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part I 27*, pages 136–147. Springer, 2021. 2, 6, 7

[31] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2018. 7

[32] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2