Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19716-5

Weakly supervised learning based bone abnormality detection from musculoskeletal x-rays

Komal Kumar[1] · **Snehashis Chakraborty**[1] · **Kalyan Tadepalli**[1,2] · **Sudipta Roy**[1]

Received: 16 November 2023 / Revised: 12 June 2024 / Accepted: 19 June 2024

Abstract
Accurate localization of abnormalities within X-ray images is of the utmost importance for arriving at the correct diagnosis. Weakly supervised learning (WSL) aims to train deep learning models for object detection and localization using only image-level labels (without using localized annotation). Most existing WSL methods use a class activation map (CAM) to generate a localization map. However, CAM-based methods have been criticized for their lack of robustness. In this work, we present a novel weakly supervised multi-stage (WSMS) learning network for accurate and efficient classification and detection of abnormalities in X-ray images. WSMS trains to localize informative regions in the image with image-level supervision. In the first stage, the WSMS network encodes the image into feature representations and localizes activated regions that contain the detailed structure of the image. The second stage proposes informative regions based on attention maps at different scales, which are used for detecting abnormalities without requiring part annotations. The final stage uses a shared weight encoder to determine if the detected region contains an object of interest. WSMS combines the objective from all stages that potentially increase the robustness. WSMS method achieves an accuracy of 97.9%, Kappa scores of 92.8%, Matthew's correlation coefficient (MCC) of 92.8%, and AUC of 96.7% for classification on the benchmark datasets and outperforms the state-of-the-art results. WSMS was tested on multiple different datasets to ensure the generalizability and reproducibility of the model. This shows the potential usability of WSMS to significantly advance medical image analysis and improve patient care in healthcare. This method achieve SOTA results without using any localized annotated data. The proposed method also removes the need of highly tedious target abnormality annotation.

Keywords X-Ray · Bone abnormality · Detection and classification · Weakly supervised learning

Extended author information available on the last page of the article

Published online: 28 June 2024

1 Introduction

Interpreting musculoskeletal X-rays is a critical aspect of orthopedic care, as bone and joint problems affect a significant portion of the global population, estimated to be around 1.7 billion people [1]. However, accurately interpreting musculoskeletal X-rays is a time-consuming and resource-intensive process [2]. In particular, the accident and emergency (A&E) department requires timely and accurate clinical observations [3]. Prior approaches based on fully supervised learning [4, 5], have shown good performances in classifying and detecting abnormalities. However, these methods rely on expert-annotated precise labels, which can be a significant limiting factor because in many real-world scenarios, obtaining such precise labels can be difficult, expensive, or time-consuming [6]. Moreover, supervised methods are limited by the black-box nature of models, which hinders their widespread clinical adoption due to the difficulty of interpreting and explaining their decision-making process [7]. Providing visual evidence of the decision-making process by localizing informative regions in images corresponding to the target can be a way to test the black-box nature of a supervised model [7, 8].

Existing methods for unsupervised feature localization [9] rely on substantial amounts of unannotated data. Weakly supervised learning offers an alternative approach by providing image-level labels to train classification and localization models [10, 11]. Class activation map (CAM) [12] is indeed one of the most used approaches for weakly supervised object localization (WSOL) [13–15]. To create a CAM, Convolutional Neural Network (CNN) feature maps are globally average pooled, passed through an FC layer with softmax activation to generate class probabilities, and then weighted to generate the final localization map [12] CAM is used to identify the most discriminative regions of an input image that are associated with a specific object class [16]. However, CAM-based methods have been criticized for their lack of robustness due to sensitivity to the background [17], difficulty in localizing co-occurring object classes [18], limited expressiveness, and susceptibility to variations in object appearance [19]. Additionally, global average pooling (GAP) introduces bias by assigning a higher weight to features with less activated areas, which further contributes to CAMs' shortcomings.

In this work, we propose a novel weakly supervised multi-stage (WSMS) learning network to address the challenges of object localization. Our approach is based on a weakly supervised object detection method [12], which has previously shown improved performance using data labeled with predefined classes. WSMS consists of three stages which aim to develop a more effective and robust model for classification and weakly supervised object localization. In each stage, we encode the image by the shared weight CNN into feature representation then We use an attention squeeze-and-excitation (SE) [20] module to calculate attention maps from the output of CNN followed by a fully connected layer (FC). WSMS localizes the informative region or object in a bounded box by the ensembles of the norm feature map and weighted attention map using binarization with an average. These two maps contain the structure of the informative region or object as shown in Fig. 1.

In the context of abnormality localization, the informative region or object in an input image is used to estimate the different discriminating regions based on scalar values obtained from the attention map. These discriminating regions are then used to localize the abnormality within a bounding box, as illustrated in Fig. 1. This approach enables the accurate and efficient detection of abnormal regions in medical images. Our approach is illustrated by the clear localization results in Fig. 1. A multi-stage neural network with shared parameters increases the robustness of object classification,

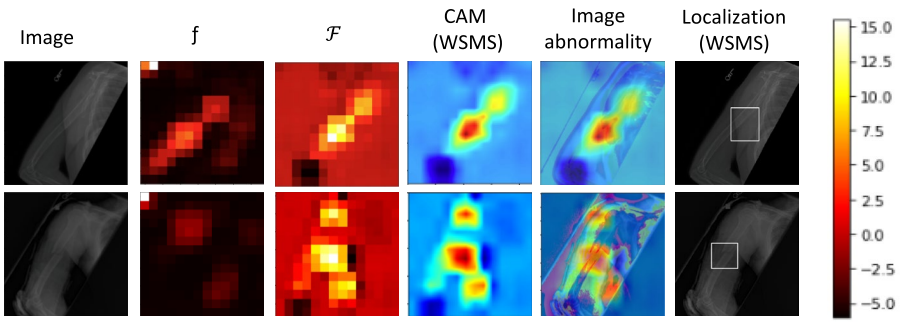

Fig. 1 This image shows components of the Class activation map (CAM), normed feature map (f) and weighted attention map ($\mathcal{F}$) from WSMS

improves feature map generalization, and enhances the clarity of CAM and localization results. This approach can also be used to check the black-box nature of the model, by providing visual evidence of the decision-making process and localizing informative regions in images corresponding to the target. Overall, multi-stage neural networks with shared parameters hold significant promise for improving the performance of object localization and classification, while also providing valuable insights into the black-box nature of the model. The potential impact of our approach on orthopedic emergency room diagnosis is significant, promising increased accuracy and reduced turnaround times.

To the best of our knowledge, this paper represents the first attempt to propose weakly supervised learning for abnormality detection from X-ray images. The key contribution of this research is summarized as follows:

1. Proposing a novel weakly supervised approach to address abnormality detection in X-ray images using a multistage attention map learning approach that not only classifies the dataset but also provides a bounding box region of interest.
2. An attention module, SE-based attention map, is introduced for feature map learning to localize the informative region in different branches. Each branch is only supervised by the image label, thus avoiding the need for costly pixel-level annotations.
3. A comprehensive evaluation is conducted on a large dataset of X-ray images with four different abnormalities. The proposed model achieves promising classification results with excellent detection visualization. Furthermore, it outperforms SOTA models on two publicly available benchmarks for classification and localization by a significant margin, demonstrating the effectiveness of the proposed approach.

The rest of this paper is structured to guide the reader through our research process step-by-step. Section 2 reviews the relevant literature, setting the stage for the discussion of our methodology. In Section 3, we explain the details of our proposed WSMS framework, including its underlying concepts and how it works. Section 4 presents our findings and provides a critical analysis of the results. Section 5 discusses our findings in the broader context of X-ray image analysis. Finally, Section 6 summarizes the key contributions of our research and suggests potential future directions.

2 Related work

The research on bone X-ray image classification and abnormality detection can be broadly divided into two areas: weakly supervised object localization and sophisticated abnormality localization. Early approaches for detecting abnormalities involved calculating neck-shift angles, Gabor analysis, and gradient-based intensity methods [21]. Later, classifiers were combined using majority voting schemes, Bayesian theory [22], AdaBoost with classifier weighting based on performance [23], and probabilistic combination [24]. With the emergence of CNNs in machine learning, researchers have applied deep learning to medical image analysis, including musculoskeletal X-ray images. Previous studies have demonstrated that deep learning (DL) models can effectively detect abnormalities in medical images using supervised learning methods for classification and bounding box localization [3]. However, it has been shown that pre-training deep CNN on non-medical images can be a viable alternative for abnormality detection [25]. Previous research on bone abnormality detection has placed less emphasis on ensuring the model's generalizability, with a greater focus on specific bone regions [26].

Researchers have used various techniques to extract features from pelvic CT and X-ray images for abnormality detection [27–29]. Adaptive filtering [30], boundary tracing [31], and wavelet transform [32] were applied to pelvic CT images, and an active shape model was developed for abnormality detection. Authors from [26] used stacked random forests based on feature fusion to detect abnormalities in X-ray images. Mathematical morphology has also been widely used for bone abnormality detection. Previous methods have considered the entire image to determine whether it is abnormal [33, 34], but they cannot localize the abnormality to a specific bone region. To address this challenge, researchers have started working with annotated bounding boxes of the abnormality as ground truth and segmenting the region of interest using techniques like entropy-based segmentation [35, 36]. Models such as ResNet [37] and Faster R-CNN [38] have been developed for abnormality detection and medical image segmentation, using annotations and bounding boxes in training. The field of weakly supervised object localization (WSOL) is a crucial area of deep learning as it reduces the time needed for user annotations by training methods to localize objects using trained classifiers. Many approaches have been developed in the past for learning object detectors with weak supervision in various problems. OXnet [39] shows promise as a feasible and general solution for real-world applications by leveraging as much available supervision as possible.

For instance, Zho et al. [12] introduced a CAM to localize the region of interest to the target level via GAP, which failed to localize due to the bias towards small activation area, as put higher weight. Recent methods like POSL [40] and SPOL [41] use two network localization and classification. MMAL-net [42] localizes without adding any additional perimeter parameters and [16] bridges the gap between classification and localization by adding new parameters to the objective. Weakly supervised is widely applied in X-ray images [39, 43]. Most recent work has shown the potential of full object localization in a weakly supervised manner [44, 45]. Our method aims to address the model's robustness by sharing parameters through the stages and localizing based on the feature map before GAP.

In Table 1 we provide a succinct overview of different weakly supervised object localization (WSOL) methods, detailing their operational principles and respective drawbacks. It contrasts approaches such as Class Activation Mapping (CAM) which targets highly discriminative features, with techniques like Divergent Activation for WSOL (DANet) that also consider less discriminative regions, aiming to enhance localization. The table also highlights the

Table 1 This table compares various WSOL methods discussed in literature, highlighting their unique approaches and inherent limitations, especially in achieving precise object localization without extensive supervision

Method	Description	Limitations
Class Activation Mapping (CAM)	Uses intermediate classifier activations to focus on the most discriminative parts of objects for localization	Focusing only on the most discriminative parts of objects is a limitation, as it does not cover the full extent of the object. Additionally, the reliance on full localization supervision for hyperparameter validation and model selection is prohibited under the WSOL setup, indicating a need for alternative evaluation protocols
Heatmap-based eXplanaible AI (XAI)	Employs heatmaps for WSOL, emphasizing explainability	XAI methods have not significantly improved beyond the CAM baseline, and their performance in terms of MaxBoxAcc scores is sub-standard, indicating a lack of accuracy in object localization
Neural Backed Decision Tree (NBDT)	Enhances models with decision tree learning to improve WSOL performance	While NBDT training can lead to better performance, it's not class-agnostic, limiting its applicability in scenarios where class-agnostic localization is desired
Divergent Activation for WSOL (DANet)	Introduces divergent activation to improve WSOL by focusing on both the most discriminative and least discriminative parts of objects	The specific limitations of DANet are not detailed in the provided sources, but it's implied that like other methods, it may face challenges in accurately localizing objects without full supervision
Adversarial Complementary Learning (ACL)	Uses adversarial training to improve WSOL by learning complementary features	The specific limitations of ACL are not detailed in the provided sources, but it's implied that like other methods, it may face challenges in accurately localizing objects without full supervision
Self-produced Guidance for WSOL	Proposes a method where the model generates its own guidance for localization, reducing the need for full supervision	The specific limitations of this method are not detailed in the provided sources, but it's implied that like other methods, it may face challenges in accurately localizing objects without full supervision

limitations each method faces, particularly in achieving accurate object localization without relying on extensive manual supervision.

3 Methodology

3.1 Overview of CAM 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 of convolutional layers followed by the average pooling and fully connected (FC) layer for classification is used to compute the CAM as follows:

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

where $F : R^{C\times H\times W} \Rightarrow R^{n\times h\times w}$ represents features map before average pooling for (h, w) spatial dimension of the n channels. $W_{cl} \in R^{n}$ are the weights of FC layer corresponding to the target class cl. Authors from [16] try to bridge the gap between the classification and localization by decomposing CAM in terms of 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. A larger value of cos θ indicates a higher degree of alignment, while a smaller value suggests a lesser degree of alignment.. We define weighted feature space ($\mathcal{F}$) map that corresponds to every class of target as follows:

$$\mathcal{F} = \|W \cdot F_{A}(X)\|, \tag{3}$$

where$W \in R^{n_class\times n} = [W1, \ldots W_{cl}, \ldots W_{n_class}]$ and norm is taken for all the classes. Based on Fig. 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 $\mathcal{F}$ contains more information to localize the object corresponding to its class-level. Object can be localized based on $\mathcal{F}$ and f where $\{ = \|F(X)||$. If the position (I, j) in f is higher than the mean of normed feature map $\bar{f}$ is part of object which we need to localize for $\forall$ I $\epsilon[0, h]$, *and* j $[0, w]$. Mathematically:

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

$$\widehat{M_{(i,j)}} = \begin{cases} 1 & iff(i,j) > \bar{f} \\ 0 & else \end{cases} \tag{5}$$

where $\widehat{M_{(i,j)}}$ is the possible area of the object to localize. The final possible area is based on the area obtained by the ensemble of $\mathcal{F}$ and f. 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 the localized images. It does not require any additional parameters as it is based on the trained classifier model. By observing Fig. 1, the area with the higher value of f are 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 Eq. 4(b). The overlapping windows are then moved across the image, allowing the classifier to process multiple regions of the image in a sliding manner.

3.2 WSMS network

Our proposed model is inspired by fine-grained (FG) visual classification models. The architecture of Weakly supervised learning-based abnormality detection is shown in Fig. 2A.

The Proposed network consists of three stages, where each stage has an encoder (En) to encode the image, which is composed of convolutional neural networks for feature map (F(X)) followed by nonlinear activation Rectified Linear Units (ReLU). Squeeze-and excitation (SE) block is then connected to the activation function to provide an attention map (AM) for the following stage. Each stage is supervised at the image-label, so a fully connected layer is added after the SE module and a revised focal loss based on binary cross entropy is used as an objective function. A further explanation of all the stages of the proposed architecture is mentioned below in detail.

3.2.1 Main stage

In this branch, we encode the image using En to get the feature representation map (f&$\mathcal{F}$). A coordinate of the bounding box is generated using the intersection of the informative region by f and $\mathcal{F}$ which is shown in Fig. 2B. In this branch, full feature AM is used for classification

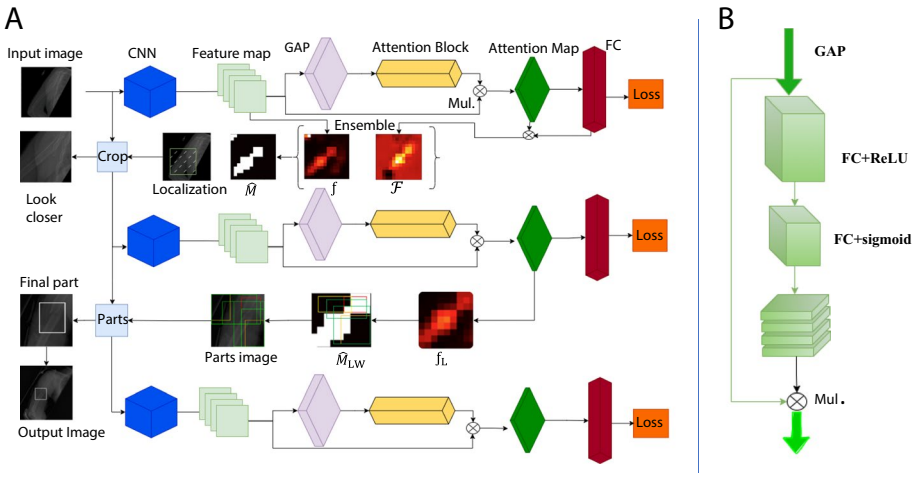

Fig. 2 **A** The full architecture of our proposed network consists of three branches, and in each branch, the same shape and same color represent the parameters shared. Attention block is used to calculate the attention map (AM) from the global average pooling by calculating the attention weights and multiplying (Mul) with feature mapping. B1, B2, and B3 represent the loss in each branch from the fully connected layer. **B** Attention block consists of two fully connected layers (FC) activated by ReLU and sigmoid respectively

using the FC layer. Furthermore, the hypothetical object by the box is cropped from the image and passed to the object branch.

3.2.2 Object stage

Cropped image (local image) is used get its feature representation by shared parameter of encoder from the main branch for classification. Then, based on local normed map (f_L) obtained from the En of second stage by passing the local image, we divide image into overlapping windows, and then classifying each window as a foreground or background. The overlapping windows are then moved across the image and cropped regions are shared to the third branch. Those regions cover the most informative part.

3.2.3 Parts classification and muti-scale stage

Object stage provides the discriminative regions in the local image by aggregating each window's normed map f_W with respect to channel and get its mean value $\overline{f_W}$ as follows:

$$\overline{\mathcal{F}_W} = \frac{\sum_{i=0}^{h_W-1}\sum_{j=0}^{w_W-1} f_W(i,j)}{h_W \times w_W} \tag{6}$$

$h_W \times w_W$ is spatial dimension of the window's normed map. We rank the windows by $\overline{f_W}$ as higher value the higher information, as shown in Fig. 2A in red (for higher f_W), orange, yellow, and green-colored windows. In this stage, cropped regions from the object branch is used for classification by shared parameters of the encoder shown in Fig. 2A by the same-colored CNN which also improves the robustness of the overall model. These cropped parts are combined to produce one possible box for abnormality.

We optimized the overall loss of the images in the training set for all stages by sharing the parameter shown by the same color CNN in Fig. 2A using three types of loss function. Due to the class imbalance problem, we have used weighted cross entropy defined as, in the Eq. (6).

$$CE_t^j(P(\theta), Q) = -\sum_{p_i \in P}\sum_{j \in T} w_+^j q_i \ln \mathbb{P}\left(q_i = 1 | p_i\right) - \sum_{p_i \in P}\sum_{j \in T} w_-^j q_i \ln \mathbb{P}\left(q_i = 0 | p_i\right) \tag{7}$$

where.

- $P(\theta) = p_1, p_2, \ldots, p_n$ are the predicted labels.
- $Q = \{q_1, q_2, \text{I.}, q_n\}$ are the labels for the corresponding instances in $P(\theta)$
- $\mathbb{P}\left(q_i = 1 | p_i\right)$ is the predictive probability $\forall q_i \in 0{,}1$ conditioned on $P(\theta)$
- T is a set of all the abnormality parts.
- w_+^j is the weight for all abnormal classes of part type $\mathrm{j} \in \mathrm{T}$
- w_-^j is the weight for the normal class of part type $\mathrm{j} \in \mathrm{T}$.

To focus on misclassification due to the class imbalance problem, we have used focal loss which is defined in Eq. (7),

$$FL_t(\theta) = \frac{1}{N}\sum_{\mathrm{i}=1}^{\mathrm{N}} \left(1 - \mathrm{e}^{-CE_t^i(\theta)}\right)^{\beta} CE_t^i(\theta) \tag{8}$$

where $CE_t(\theta)$ is cross entropy at t^{th} step with θ shared parameter defined in Eq. (6), β are the focal loss hyperparameter. It is clear from the Eq. (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 computational cost. So instead of mean, we have used a scaler α. Rewriting Eq. (7) as,

$$FL_t(\theta) = \alpha\left(1 - \mathrm{e}^{-CE_t(\theta)}\right)^{\beta} CE_t(\theta) \tag{9}$$

So, we sum the losses of all stages to get the total loss:

$$L_t^{total}(\theta) = FL_t^{B1}(\theta) + FL_t^{B2}(\theta) + FL_t^{B3}(\theta) \tag{10}$$

where $B1$, $B2$, and $B3$ in $FL_{t(\theta)}$ represent the corresponding loss of the first, second, and third branches respectively. For N number of the part images, the third branch loss is defined as the sum of the focal loss of each part:

$$FL_t^{P}(\theta) = \sum_{i=0}^{N-1} FL_t^{pi}(\theta) \tag{11}$$

3.3 Attention map module

CNN is used to encode the image into a feature map where we extract high-level features from images. Next, the output of CNN activated by ReLU is passed through the Attention map module shown in Fig. 2B, which computes attention weight for each F(X). Since each F(X) in F is predicted by a separate kernel, we assume that CNN may generate activation maps with unnecessary values across feature maps F(X). Our goal is to use F(X) relationships by scaling each channel according to the quality of representations produced by CNNs. To solve the problem, we employ a SE block. It aims to improve the overall network's performance by emphasizing the importance of certain features that result in better localization. First, we apply global weighting to the F(X) of a given feature map to highlight the most essential features and suppress the less important ones. As a result, the encoder can focus on the prominent features of the image, resulting in improved performance. Additionally, attention is intended to increase WSMS's modeling capacity and enable non-linearity between F(X) and attention map (AM) outputs. For $F \in \mathbb{R}^{C\times h\times w}$, attention map module reduces to $s \in \mathbb{R}^{\mathrm{C}}$ through global average weighting on each F(X). Thereafter, it is passed through fully connected layers activated by ReLU which reduces the number of feature maps C by $\frac{C}{r}$ where r the reduction ratio is followed by fully connected layers to compute C important features corresponding to the image label. The steps are as follows:

$$S\prime = ReLU\left(W_1 \cdot s + B_1\right) \tag{12}$$

$$S = \sigma\left(W_2 \cdot S\prime + B_2\right) \tag{13}$$

where $W_1 \in \mathbb{R}^{\frac{\mathrm{C}}{\mathrm{r}}\times C}$ & $B_1 \in \mathbb{R}^{\frac{\mathrm{C}}{\mathrm{r}}}$ are the weights & biases for first fully connected layer (FC) activated by ReLU followed by the second fully connected layer in which $W_2 \in \mathbb{R}^{C\times\frac{\mathrm{C}}{\mathrm{r}}}$ & $B_2 \in \mathbb{R}^{C}$ are the weights & biases. The first FC layer consists of $\frac{C}{r}$ nodes where information is squeezed by a reduction ratio of r. Finally, a tensor Z of the same spatial dimension as of F is generated as follows:

$$Z = S \cdot F(X) \tag{14}$$

where • element-wise multiplication and $Z \in \mathbb{R}^{\mathrm{C}\times\mathrm{h}\times\mathrm{w}}$ is the final tensor containing recalibrated attention map.

3.4 Feature localization and amplification

The normed feature map f of the full image is binarized based on equation using mean thresholding value as shown in Eq. 4(b). 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 select the region that covers the maximum activation area to look closer for the localization of informative region within X-ray with the bilinear up sampling method.

3.5 Informative regions localization

Although the cropped local image contained the informative region with good probability, but idea is to localize the key part of the image. We use a feature map to search for the areas with higher activation, which indicates the location of critical parts in the local image (cropped image). So, we extract the feature map of the cropped image by sharing the En's parameter to obtain the normed feature map f_{L} 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 a kernel size of $\left(h_w, w_w\right)$. 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 the many distinct parts as possible. Localization is based on the binary map obtained from the activation map of each window by mean thresholding f_{W} the window defined in Eq. (5). Non-Maximum Suppression (NMS) [38] is applied after scoring to select the fixed number of parts in images so fewer redundant parts are in each region.

3.6 Possible abnormality detection

From the informative region's localization section, the larger the value of f_{W}, larger is the information that part contains. We combine w windows to detect the abnormality by discounting the value window for decreasing the value of $\bar{f}_W$ after shorting.

Then final window value $(U(x, y, x + h, y + h))$ is given by:

$$U\left(x, y, x + w_w, y + h_w\right) = \sum_{i=1}^{w_n} \gamma^i U\left(x_i, y_i, x_i + w_{wi}, y_i + h_{wi}\right) \tag{15}$$

where,$\gamma \in R^{(0,1]}$ 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 window$\forall i \in \left[1, w_n\right]$. But $U(x, y, x + w, y + h)$ is a part localized in the local image cropped from the original image and resized in the size of the original image, so to get the reflection on the original image, we proposed a simple flowchart shown in Fig. 3.

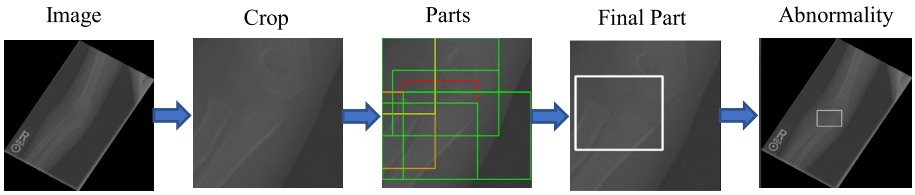

Fig. 3 Visualization of the output from our proposed weakly supervised fine-grained detection model at each step. Crop is the first localized informative region and Parts are the proposed region in Crop which produces the Final Part. The last images are abnormal in white box, which reflects the Final Part on input Image

4 Experiments results and analysis

4.1 Dataset description

Abnormality dataset To evaluate our models, a dataset comprising of 5 types of abnormalities, namely, fracture, tumor, dislocations, metal implant, and arthritis along with normal bone images was curated. The images were collected from various regions of the human body, with the majority of the fractured images being sourced from the MURA dataset [1]. Other classes of this dataset are collected by the radiologist. In this work, we have conducted experiments based on two classes i.e., normal, and abnormal classes for training. The size of the training dataset is 38.7k in which 16.7k are the abnormal images from five different abnormalities and 16k are the normal bone images and the test dataset consists of 6k images with 1k abnormal images collected by the same radiologist. This entire dataset has X-ray images in both normal and abnormal cases.

HAND Fracture The class imbalance problem is the most important problem in medical images for weakly supervised learning especially in our method. So, to check the effectiveness of our model in a class imbalance class environment, we have used a Hand dataset from MURA [1]. This dataset consists of 5.5k images in trainset with 26.7% fractured images and 460 images in test set with 41% fractured images.

CUB 200 2011 For FG classification and localization of objects, we use the CUB dataset to test the SOTA performance of our proposed model. The CUB-200–2011 dataset contains 11,788 images across 200 bird species categories, with annotations for part locations, attributes, and bounding boxes, suitable for supervised learning tasks involving fine-grained visual categorization..

Stanford Cars The Stanford Cars dataset is a collection of images of vehicles, along with their associated labels, that was collected by researchers at Stanford University. It contains more than 16,000 images of 196 cars, consisting of models made by various manufacturers. To check the SOTA performance of the proposed model on classification, we use Stanford Cars (CAR) dataset.

4.2 Experimental setup

4.2.1 Implementation details

The input images were pre-processed to a size of 448×448 to obtain the augmented images for the first and second branches, as illustrated in Fig. 2A. The original image was cropped based on the coordinates from the first branch, and the input in the second branch was also scaled to a height and width of 448. For the part branch, all images were reshaped to a height and width of 224. We selected a window with a broad range of scale categories: $\{[6 \times 6, 7 \times 5], [8 \times 8, 6 \times 10, 7 \times 9], [10 \times 10, 9 \times 11, 8 \times 12]\}$ and used 14×14 as the size of f in Eq. 4(a). The number of part images w_n was set to 7, where $w_{1n} = 2$, $w_{2n} = 2$, and $w_{3n} = 3$ were the number of wide varieties of scales. Different pre-trained baselines on ImageNet were used as the backbone within the same-colored CNNs shown in Fig. 2A. During model training, no other annotations were used except for the images' class labels. We optimized the loss using SGD, with an initial learning rate of 1×10^{-4} and a minibatch size of 5 on RTX A4000 GPU. PyTorch was used as the codebase for implementation.

We use various performance measures including (Area Under the Curve) (AUC) [46], Cohen's Kappa (KAPPA) [47], Matthews Correlation Coefficient (MCC) [48], and Accuracy. AUC measures the degree of separability between the positive and negative classes. AUC is calculated by plotting the Receiver Operating Characteristic (ROC) curve, which is a graph of the True Positive Rate ($\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$) against the False Positive Rate ($\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$) at various threshold settings, and then computing the area under this curve. Here we use the notation from the confusion matrix and represent TP (True Positive) as cases that are both Actual Positive and Predicted Positive, FN (False Negative) as cases that are Actual Positive but Predicted Negative, FP (False Positive) as cases that are Actual Negative but Predicted Positive, and TN (True Negative) as cases that are both Actual Negative and Predicted Negative. A higher AUC value indicates a better model performance. In discrete terms, AUC can also be approximated by summing the areas of trapezoids formed by the points on the ROC curve:

$$\text{AUC} \approx \sum_{i=1}^{n-1} \left(\text{FPR}_{i+1} - \text{FPR}_i\right) \left(\frac{\text{TPR}_{i+1} + \text{TPR}_i}{2}\right),$$

where i represents each point on the ROC curve and n is the total number of points. KAPPA is a statistical measure used to assess the agreement between two raters who each classify items into mutually exclusive categories. It accounts for the possibility of the agreement occurring by chance. The KAPPA is calculated using the observed agreement (P_o) and the expected agreement (P_e):

$$\text{KAPPA} = \frac{P_o - P_e}{1 - P_e}$$

Here, P_o represents the relative observed agreement among raters, while P_e represents the hypothetical probability of chance agreement and calculated as follows:

$$P_o = \frac{\text{TP+TN}}{n}$$
$$P_e = \frac{(\text{TP+FN})(\text{TP+FP}) + (\text{FP+TN})(\text{FN+TN})}{n^2}$$

MCC is a balanced measure of the quality of binary classifications, taking into account true and false positives and negatives. It is especially useful when the classes are of different sizes. MCC is calculated with the formula:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$
$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$$

And accuracy measures the proportion of correct predictions:

4.2.2 Baseline

We have used Resnet18, Resnet34, Resnet50, Resnet101, and inception [41] as a backbone pretrained on Imagenet21k for the feature map.

4.3 Comparison with baseline network

An experimental evaluation was conducted on the largest abnormality dataset to compare with well-established baseline models for weakly supervised fine-grained detection and classification. These classification results are shown in Table 2 for the performance measures discussed. With a good margin of more than 8.8% AUC, 12.6% KAPPA, 12.6% MCC, and 3.9% Accuracy, our proposed model exceeds the baseline models.

4.4 Comparison SOTA

To compare the model's performance against the state-of-the-art (SOTA), we conducted comprehensive experiments on publicly available datasets for fine-grained (FG) classification. The results are reported in Table 3. The table presents a comprehensive comparison

Table 2 Comparison results on Bone Abnormality dataset with baseline models

Methods	Measure	Abnormality
Proposed model	AUC	**96.7%**
	KAPPA	**92.8%**
	MCC	**92.8%**
	Accuracy	**97.9%**
Resnet-50 [37]	AUC	87.6%
	KAPPA	78.8%
	MCC	79.0%
	Accuracy	94.0%
Inception Net [41]	AUC	87.9%
	KAPPA	80.2%
	MCC	80.2%
	Accuracy	90.0%
Dense Net [49]	AUC	74.8%
	KAPPA	70.5%
	MCC	69.8%
	Accuracy	76.0%

Table 3 Comparison with SOTA models on their backbone. SOTA performance is highlighted by bold

Methods	Backbone	CUB (%)	CAR (%)
Bilinear-CNN [50]	VGG	84.1	91.3
RA-CNN [9]	VGG19	85.3	92.5
KP [51]	Resnet50	86.2	92.4
MAMC [52]	Resnet50	86.3	93.0
PC [53]	DenseNet-161	86.9	92.9
HBP [54]	VGG-16	87.1	93.7
Mask CNN [55]	Resnet50	87.3	-
DFL-CNN [56]	Resnet50	87.4	93.8
NTS-NET [57]	Resnet50	87.5	91.4
TASN [58]	Resnet50	87.9	93.8
LIO [59]	Resnet50	88.0	94.5
BNT [60]	Resnet50	88.1	94.6
ASD [61]	Resnet50	88.6	**94.9**
API-Net [62]	Resnet101	88.6	**94.9**
P2P-Net [63]	Resnet34	**89.5**	**94.9**
WSMS-Net (Proposed)	Resnet50	89.2	**95.0**

of various models' performance on the CUB and CAR datasets for fine-grained image classification. Several key observations can be made from the results. First, the choice of backbone architecture plays a significant role in the model's performance, with models using ResNet50 as the backbone consistently achieving high accuracy on both datasets. For instance, models like MAMC, DFL-CNN, TASN, LIO, BNT, ASD, and API-Net, all based on ResNet50, achieved accuracy rates above 86% on CUB and above 92% on CAR. Second, the proposed WSMS-Net, also based on ResNet50, achieved competitive accuracy rates of 89.2% on CUB and 95.0% on CAR, outperforming several SOTA models such as RA-CNN, KP, NTS-NET, and HBP. These results suggest that the WSMS-Net model effectively leverages the full feature map, leading to improved localization and classification performance.

Our proposed net outperforms the SOTA models on the CAR dataset by more than 0.1%. Specifically, our model achieved an improvement of over 0.4% compared to existing methods. However, when compared with the very recent P2P-Net [63] on the CUB200 dataset, our model lagged by just 0.3%. Finally, it is noteworthy that the performance on the CUB dataset is generally higher compared to the CAR dataset, indicating the difficulty of the CAR dataset due to its complex nature and diverse car models. The discriminative parts found in these datasets (See Fig. 4) can be valuable for further improving the performance of classification models.

4.5 Localization interpretation

The Percentage of Correctly Localized (PCL) region is used to evaluate the accuracy of a localization method in determining the location of an object or informative region with an IOU of over 50%. Using an ensemble of two ResNet-50 layers, the Attention Object Learning Module achieved a PCL of 85.1% after one epoch with a pre-trained ImageNet21k backbone. However, as training progressed, the PCL decreased to 71.1% because the CNN-based network focused more on the most prominent regions. Our proposed model, which

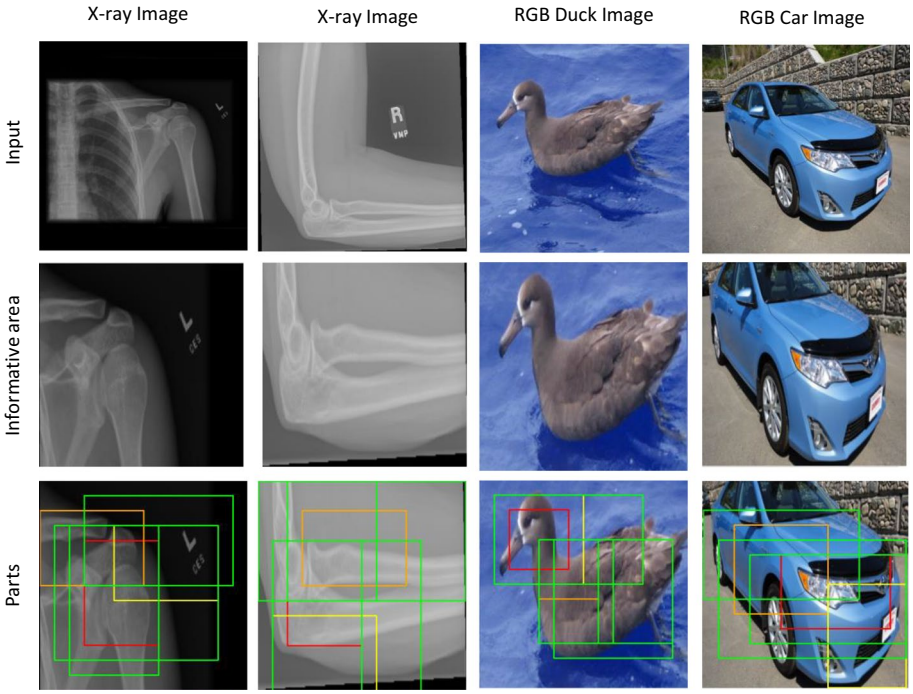

Fig. 4 Localization of crops and discriminative part localization. We use red, orange, yellow, and green colors to indicate the order of the windows by $\overline{f_W}$

combines the attention and backbone modules, achieved the highest accuracy of 77.2%, surpassing recent weakly supervised methods (See Table 4). However, the PCL also decreased to 73.4% as training progressed, which is still better than the Attention Object Localization Module (AOLM) (see Table 3) [42].

4.6 Visualizations

4.6.1 Localization

The second column of Fig. 4 shows the localization of informative parts for the X-ray abnormality, CUB, and CAR datasets based on the test images. The visualization

Table 4 Localization performance on CUB. "Yes" represents the training from scratch of the attention module (AM)

Methods	Training from scratch	PCL (%)
ACOL	No	46.0
ADL	No	62.3
SCDA	No	76.8
MMAL [42]	No	85.1 (drop to 71.1)
Our (F(X)& AM)	Yes	77.2 (drop to 73.4)

demonstrates that our model accurately focuses on the informative region of the image without any information loss in the crop part.

4.6.2 Discriminative part

Fig. 4 Part column visualizes the location of the discriminative part by our net. This figure displays regions with the highest average activation values on different scales using red, orange, yellow, and green boxes, with the red box representing the highest average activation value. The most discriminative features, which are similar to human perception, are found in the joint of bone abnormality dataset, the head and beak of birds, and the headlight front-side of cars

4.7 Abnormality location

The results of the model on bone abnormality detection in X-ray images from the test dataset are depicted in Fig. 5. The white box in the figure indicates bone abnormalities such as bone tumors, joint dislocations, arthritis, and metal inside the bone. In the third column of the figure, a unique bone tumor abnormality identified by the radiologist is shown in the white box. The radiologist verified all abnormality detections from the test dataset. The model has demonstrated its capability to perform in complex cases that even a human expert found difficult to diagnose.

4.8 Comparison with weakly supervised CAM

In Fig. 6, we compare CAM from the first branch of WSMS 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 (WSMS) produces similar localization results as our proposed method, demonstrating the agreement between CAM (WSMS) 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 Fig. 6. In contrast, our method provides a clear idea of the required bounded box, as depicted in Fig. 6. The figure also ensures the robustness of the model because of multistage training compared to normal CAM. Our method carefully inspects the image in the first branch, followed by the second branch that decides on windows with different confidence. These windows have a low probability of being incorrect as the first branch ensures accuracy, and we combine all windows to form the final bounded box.

4.9 Ablation study

4.9.1 Effect of Attention module

In Fig. 6, the second and third columns display localization using the feature map and attention map assembled with the feature map of the image in the first column of Fig. 6, respectively. The fourth and fifth columns of Fig. 7 show abnormality detection in the white box corresponding to the image. The impact of attention maps (AM) on classification performance is presented in Fig. 8. Although the results of localization are similar

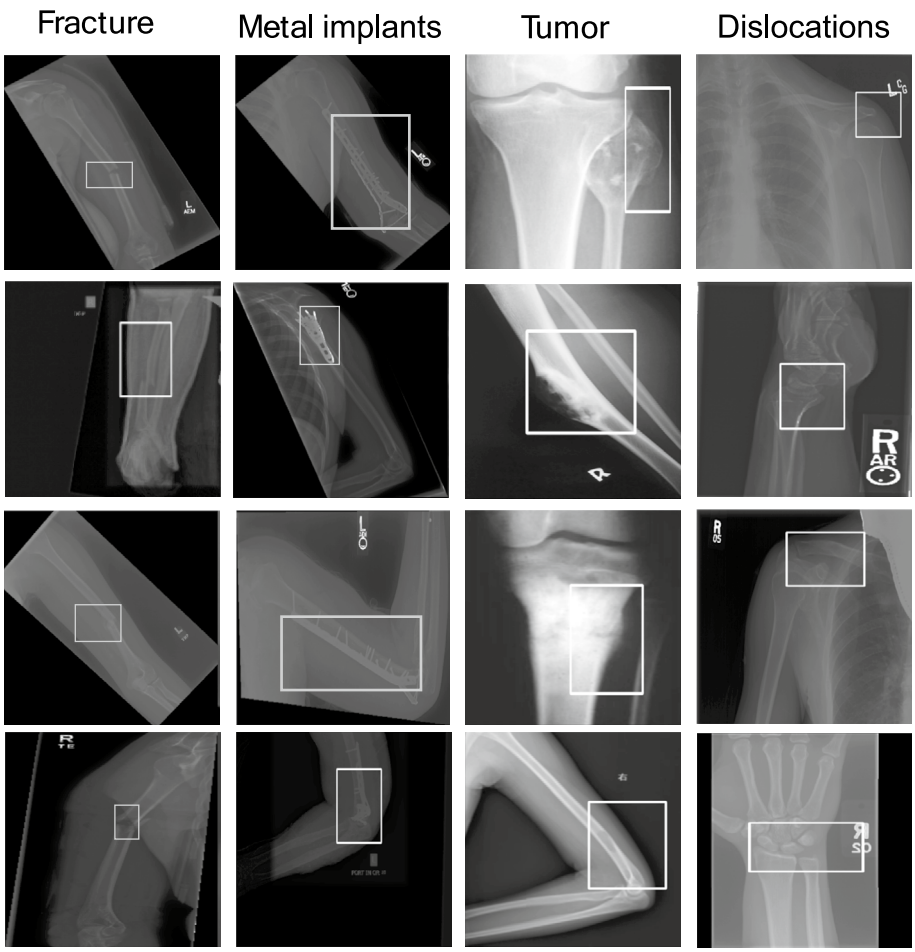

Fig. 5 The abnormality detected region is in the bounding box. The area in the fourth column indicates Bone Tumor

in both cases, the AM-based localization is highly focused on the weighted feature map, resulting in more accurate results, as seen in the fifth column verified by the radiologist. The use of AM not only improves the localization and abnormality detection accuracy but also enhances the classification performance, which is reported in Fig. 8. Compared to F(X)-based classification performance, AM outperforms it by a significant margin (17.8% of AUC, 23% of KAPPA, 23% of MCC, and 6.3% of accuracy) due to its effectiveness in improving the network's discriminative power by emphasizing important features and suppressing less important ones.

4.10 Class imbalance problem in weakly supervised learning

The hand training dataset has an issue of class imbalance, with only 27.77% of data belonging to the '+' class and 73.22% belonging to the '-' class, which may be one of the reasons

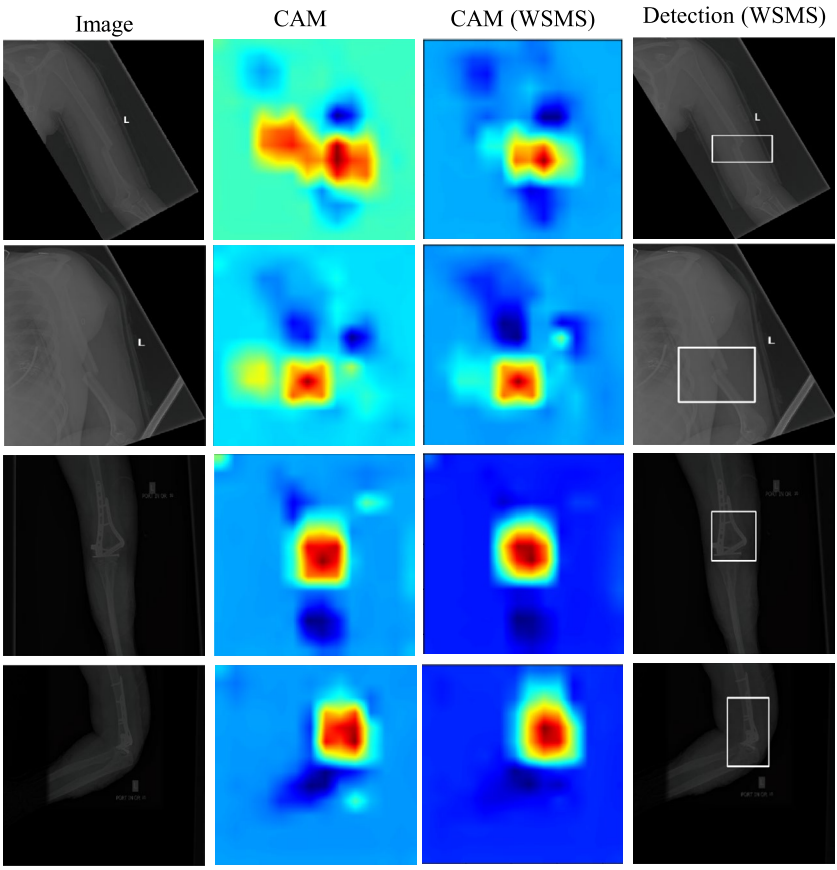

Fig. 6 Visual comparison with weakly supervised CAM

for models performing worse than expected. To address this problem, various loss functions based on cross-entropy have been employed, including the weighted cross-entropy (WCE) loss function (Eq. (6)), the mean focal loss (MFL) (Eq. (7)), and the α focal loss (Eq. (8)). In our experiments, the α-WFL loss function has shown the highest MCC score and approximately the same accuracy as the highest-performing loss function based on accuracy. As a result, we have used α-WFL for all our experiments. Results of our models with different loss functions are presented in Table 5.

4.11 Model baselines

Figure 9 shows the results of our Net on the CUB200 dataset with different baseline models. As the baseline model complexity increases (from resnet18 to resnet101 and xception), accuracy generally improves, except for resnet101 which improves accuracy compared to the xception net. This is because resnet101 is a deeper and broader version of resnet50, allowing it to learn more complex and nuanced features from the data, potentially improving accuracy on some tasks. However, this also means that training resnet101 may require more computation resources and may be more prone to overfitting if the training data is

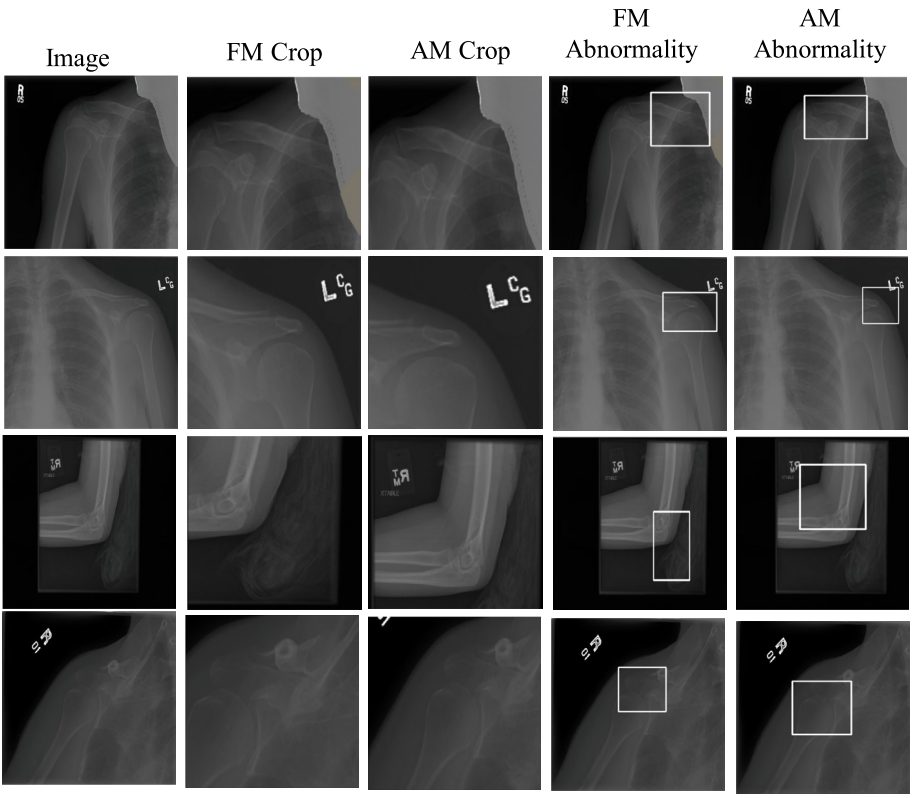

Fig. 7 Effect of attention module (AM) on localization (crop) and abnormality detection compared to feature maps (F(X)s)

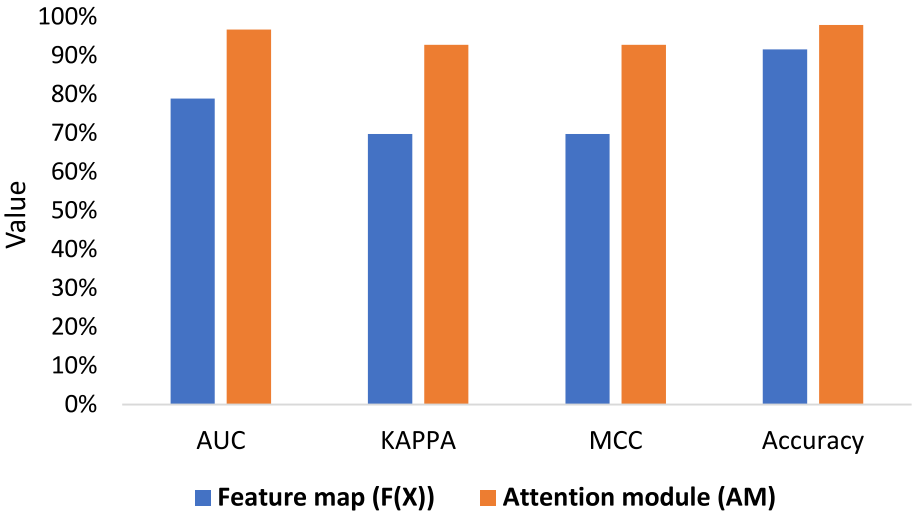

Fig. 8 Effect of attention module (AM) on feature map (F(X)) on Bone Abnormality datasets

Table 5 Effect of loss function for class imbalance in weakly supervised learning

Loss Functions	MCC	Accuracy
CE	0.41	0.71
WCE	0.46	0.75
MFL	0.45	0.75
WMFL	0.50	0.79
α-FL	0.52	0.78
α-WFL	0.54	0.78

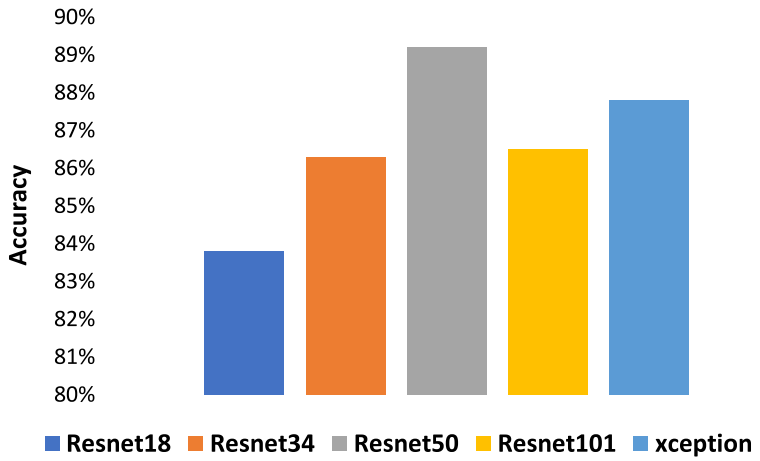

Fig. 9 Variety of baseline for WSMS net on CUB dataset

insufficient or the regularization is not strong enough. Across all the baseline models, our WSMS net is outperformed by the resnet50 baseline (see Fig. 9).

5 Discussion

Weakly supervised learning is a promising approach for medical image analysis, offering potential benefits in terms of scalability, cost-effectiveness, and accuracy. With further research and development, weakly supervised learning methods have the potential to greatly improve the field of medical image analysis.

In the development of weakly supervised learning for medical image analysis, we propose a WSMS learning network. The proposed weakly supervised learning approach in this research paper has shown promising results in addressing the challenge of abnormality detection in X-ray images. The WSMS approach, which leverages the available supervision to the maximum extent possible, not only classifies the abnormality dataset but also provides the region of interest in the form of a bounding box. This approach enables the model to learn discriminative features and focus on relevant regions in the X-ray images for better classification and localization performance.

Using CAM and attention maps improves the model's interpretability, providing visual evidence of the decision-making process and facilitating verification mitigating some of the black-box aspects of the classifier. The multi-stage architecture with shared parameters increases the robustness and generalization performance of the model by enabling it to learn more complex and abstract features. The experimental results demonstrate the effectiveness of the proposed approach, outperforming several state-of-the-art methods in terms of classification accuracy and achieved clear visualization of the area of interest.

However, we recognize that there is room for improvement in the proposed approach. More advanced attention mechanisms and architectures can be explored to further improve the interpretability and localization performance of the model. Additionally, the use of more diverse and larger datasets can be investigated to evaluate the generalization performance of the proposed approach in real-world scenarios.

A notable limitation of our weakly supervised method is its suboptimal performance on datasets with class imbalances. Despite incorporating objective functions designed to mitigate this issue, the model's reliance on positive samples during training remains a challenge, particularly for representing features of less prevalent classes. This limitation suggests the necessity for continued exploration of strategies to counter class imbalance in weakly supervised learning, especially within the context of medical imaging. Also, in future wanted to explore the explainability of the method and outcome [64, 65]. Overall, the proposed weakly supervised learning approach shows substantial potential for addressing the challenges of abnormality detection in X-ray images, and further research in this direction can lead to significant advancements in medical imaging applications.

6 Conclusion

The proposed approach in this work addresses the challenge of abnormality detection in X-ray images through a weakly supervised multistage attention map learning approach. The use of a multistage neural network with shared parameters increases the robustness of the model in classification and feature map generalization, resulting in clearer attention maps and better object localization. Comprehensive experiments demonstrate that the model outperforms SOTA baseline models in terms of classification and abnormality localization quality. The proposed model achieves SOTA results in the classification and localization of images, as demonstrated through the analysis of two public benchmarks. This approach is promising for real-world applications as it leverages as much available supervision as possible, but still requires some level of supervision. The ability to provide a region of interest in the form of a bounding box in addition to classification provides a useful tool for medical professionals in identifying abnormalities and their locations in X-ray images. Overall, this work has the potential to significantly impact the field of medical image analysis and improve diagnostic accuracy in clinical settings.

Future work could explore the use of additional sources of supervision or a more comprehensive weakly supervised approach to further improve the accuracy and robustness of the proposed method. Additionally, further evaluation and comparison with state-of-the-art methods could provide insight into the effectiveness and practicality of this approach.

Acknowledgements This research work was supported by the RFIER-Jio Institute "CVMI-Computer Vision in Medical Imaging" research project (RFIER-Jio Institute, Grant # 2022/33185004) fund under the "AI for ALL" research center.

Data availability All publicly available datasets used in this study and description are included in "Dataset Description" subsection in this manuscript. The source links are as follows:

Abnormality dataset: https://stanfordmlgroup.github.io/competitions/mura/
CUB 200 2011: https://www.kaggle.com/datasets/coolerextreme/cub-200-2011
Stanford Cars: http://ai.stanford.edu/~jkrause/cars/car_dataset.html

Code availability The source code is available at https://github.com/MAXNORM8650/WSMS.

Declarations

Conflict of Interest The authors have no conflict of interest to declare.

References

1. Rajpurkar P et al. (2017) MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs, https://doi.org/10.48550/ARXIV.1712.06957
2. Sharma S (2023) Artificial intelligence for fracture diagnosis in orthopedic X-rays: current developments and future potential. SICOT J 9:21. https://doi.org/10.1051/sicotj/2023018
3. Yancey CC, O'Rourke MC (2024) Emergency Department Triage, in StatPearls, Treasure Island (FL): StatPearls Publishing. Available: http://www.ncbi.nlm.nih.gov/books/NBK557583/. Accessed 27 Mar 2024
4. Meena T, Roy S (2022) Bone Fracture Detection Using Deep Supervised Learning from Radiological Images: A Paradigm Shift. Diagnostics 12(10):2420. https://doi.org/10.3390/diagnostics12102420
5. Roy S, Meena T, Lim S-J (2022) Demystifying Supervised Learning in Healthcare 4.0: A New Reality of Transforming Diagnostic Medicine. Diagnostics 12(10):2549. https://doi.org/10.3390/diagnostics12102549
6. Kumar K, Chakraborty S, Roy S (2023) Self-supervised Diffusion Model for Anomaly Segmentation in Medical Imaging, in Pattern Recognition and Machine Intelligence, vol. 14301, P. Maji, T. Huang, N. R. Pal, S. Chaudhury, and R. K. De, Eds., in Lecture Notes in Computer Science, vol. 14301. , Cham: Springer Nature Switzerland, pp. 359–368. https://doi.org/10.1007/978-3-031-45170-6_37
7. Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A (2017) Practical Black-Box Attacks against Machine Learning, in Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, Abu Dhabi United Arab Emirates: ACM, pp. 506–519. https://doi.org/10.1145/3052973.3053009
8. Yang G, Ye Q, Xia J (2022) Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Information Fusion 77:29–52. https://doi.org/10.1016/j.inffus.2021.07.016
9. Fu J, Zheng H, Mei T (2017) Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI: IEEE, pp. 4476–4484. https://doi.org/10.1109/CVPR.2017.476
10. Kumar K, Pailla B, Tadepalli K, Roy S (2023) Robust MSFM Learning Network for Classification and Weakly Supervised Localization, in 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Paris, France: IEEE, pp. 2434–2443. https://doi.org/10.1109/ICCVW60793.2023.00258
11. Galleguillos C, Babenko B, Rabinovich A, Belongie S (2008) Weakly Supervised Object Localization with Stable Segmentations, in Computer Vision – ECCV 2008, vol. 5302, D. Forsyth, P. Torr, and A. Zisserman, Eds., in Lecture Notes in Computer Science, vol. 5302. , Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 193–207. https://doi.org/10.1007/978-3-540-88682-2_16
12. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2015) Learning Deep Features for Discriminative Localization, https://doi.org/10.48550/ARXIV.1512.04150
13. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2020) Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis 128(2):336–359. https://doi.org/10.1007/s11263-019-01228-7
14. Choe J, Shim H (2019) Attention-Based Dropout Layer for Weakly Supervised Object Localization, in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA: IEEE, pp. 2214–2223. https://doi.org/10.1109/CVPR.2019.00232

15. Xue H, Liu C, Wan F, Jiao J, Ji X, Ye Q (2019) DANet: Divergent Activation for Weakly Supervised Object Localization, in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South): IEEE, pp. 6588–6597. https://doi.org/10.1109/ICCV.2019.00669
16. Kim E, Kim S, Lee J, Kim H, Yoon S (2022) Bridging the Gap between Classification and Localization for Weakly Supervised Object Localization, in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA: IEEE, pp. 14238–14247. https://doi.org/10.1109/CVPR52688.2022.01386
17. Zhu L et al (2023) Background-Aware Classification Activation Map for Weakly Supervised Object Localization. IEEE Trans Pattern Anal Mach Intell 45(12):14175–14191. https://doi.org/10.1109/TPAMI.2023.3309621
18. Shao F et al. (2024) Counterfactual Co-occurring Learning for Bias Mitigation in Weakly-supervised Object Localization. arXiv, Mar. 09. Accessed: Mar. 27, 2024. [Online]. Available: http://arxiv.org/abs/2305.15354
19. Wei J, Wang S, Zhou SK, Cui S, Li Z (2022) Weakly Supervised Object Localization Through Inter-class Feature Similarity and Intra-class Appearance Consistency," in Computer Vision – ECCV 2022, vol. 13690, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., in Lecture Notes in Computer Science, vol. 13690. , Cham: Springer Nature Switzerland, pp. 195–210. https://doi.org/10.1007/978-3-031-20056-4_12
20. Hu J, Shen L, Albanie S, Sun G, Wu E (2017) Squeeze-and-Excitation Networks, https://doi.org/10.48550/ARXIV.1709.01507
21. Tian TP, Chen Y, Leow WK, Hsu W, Howe TS, Png MA (2003) Computing Neck-Shaft Angle of Femur for X-Ray Fracture Detection, in Computer Analysis of Images and Patterns, vol. 2756, N. Petkov and M. A. Westenberg, Eds., in Lecture Notes in Computer Science, vol. 2756. , Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 82–89. https://doi.org/10.1007/978-3-540-45179-2_11
22. Dennis Wen-Hsiang Yap, Ying Chen, Wee Kheng Leow, Tet Sen Howe, and Meng Ai Png (2004) Detecting femur fractures by texture analysis of trabeculae, in Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., Cambridge, UK: IEEE, pp. 730–733 Vol.3. https://doi.org/10.1109/ICPR.2004.1334632
23. Kuncheva LI (2003) 'Fuzzy' versus 'nonfuzzy' in combining classifiers designed by boosting. IEEE Trans Fuzzy Syst 11(6):729–741. https://doi.org/10.1109/TFUZZ.2003.819842
24. Vineta Lai Fun Lum, Wee Kheng Leow, Ying Chen, Tet Sen Howe, Meng Ai Png (2005) Combining classifiers for bone fracture detection in X-ray images, in IEEE International Conference on Image Processing 2005, Genova, Italy: IEEE, p. I–1149. https://doi.org/10.1109/ICIP.2005.1529959
25. Hwang S, Kim H-E (2016) Self-Transfer Learning for Weakly Supervised Lesion Localization, in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, vol. 9901, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds., in Lecture Notes in Computer Science, vol. 9901. , Cham: Springer International Publishing, pp. 239–246. https://doi.org/10.1007/978-3-319-46723-8_28
26. Cao Y, Wang H, Moradi M, Prasanna P, Syeda-Mahmood TF (2015) Fracture detection in x-ray images through stacked random forests feature fusion, in 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), Brooklyn, NY, USA: IEEE, pp. 801–805. https://doi.org/10.1109/ISBI.2015.7163993
27. Sharma N et al (2010) Automated medical image segmentation techniques. J Med Phys 35(1):3. https://doi.org/10.4103/0971-6203.58777
28. Albadr MAA, Tiun S, Ayob M, AL-Dhief FT, Omar K, Hamzah FA (2020) Optimised genetic algorithm-extreme learning machine approach for automatic COVID-19 detection. PLoS ONE 15(12):e0242899. https://doi.org/10.1371/journal.pone.0242899
29. Chakraborty S, Kumar K, Tadepalli K, Pailla BR, Roy S (2023) Unleashing the power of explainable AI: sepsis sentinel's clinical assistant for early sepsis identification. Multimed Tools Appl. https://doi.org/10.1007/s11042-023-17828-y
30. Kavitha P, Prabakaran S (2019) A Novel Hybrid Segmentation Method with Particle Swarm Optimization and Fuzzy C-Mean Based On Partitioning the Image for Detecting Lung Cancer https://doi.org/10.20944/preprints201906.0195.v1
31. Wu J, Davuluri P, Ward KR, Cockrell C, Hobson R, Najarian K (2012) Fracture Detection in Traumatic Pelvic CT Images. Int J Biomed Imaging 2012:1–10. https://doi.org/10.1155/2012/327198
32. Arzhaeva Y, Prokop M, Tax DMJ, De Jong PA, Schaefer-Prokop CM, Van Ginneken B (2007) Computer-aided detection of interstitial abnormalities in chest radiographs using a reference standard based on computed tomography. Med Phys 34(12):4798–4809. https://doi.org/10.1118/1.2795672

33. Bandyopadhyay O, Biswas A, Bhattacharya BB (2016) Long-bone fracture detection in digital X-ray images based on digital-geometric techniques. Comput Methods Programs Biomed 123:2–14. https://doi.org/10.1016/j.cmpb.2015.09.013
34. Guggenberger R et al (2012) Diagnostic Performance of Dual-Energy CT for the Detection of Traumatic Bone Marrow Lesions in the Ankle: Comparison with MR Imaging. Radiology 264(1):164–173. https://doi.org/10.1148/radiol.12112217
35. Bandyopadhyay O, Chanda B, Bhattacharya BB (2011) Entropy-Based Automatic Segmentation of Bones in Digital X-ray Images, in Pattern Recognition and Machine Intelligence, vol. 6744, S. O. Kuznetsov, D. P. Mandal, M. K. Kundu, and S. K. Pal, Eds., in Lecture Notes in Computer Science, vol. 6744. , Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 122–129. https://doi.org/10.1007/978-3-642-21786-9_22
36. Chakraborty S, Kumar K, Reddy BP, Meena T, Roy S (2023) An Explainable AI based Clinical Assistance Model for Identifying Patients with the Onset of Sepsis, in 2023 IEEE 24th International Conference on Information Reuse and Integration for Data Science (IRI), Bellevue, WA, USA: IEEE, pp. 297–302. https://doi.org/10.1109/IRI58017.2023.00059
37. He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90
38. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
39. Luo L, Chen H, Zhou Y, Lin H, Heng P-A (2021) OXnet: Deep Omni-Supervised Thoracic Disease Detection from Chest X-Rays, in Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, vol. 12902, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds., in Lecture Notes in Computer Science, vol. 12902. , Cham: Springer International Publishing, pp. 537–548. https://doi.org/10.1007/978-3-030-87196-3_50
40. Wei J, Wang Q, Li Z, Wang S, Zhou SK, Cui S (2021) Shallow Feature Matters for Weakly Supervised Object Localization, in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA: IEEE, pp. 5989–5997. https://doi.org/10.1109/CVPR46437.2021.00593
41. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the Inception Architecture for Computer Vision, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, pp. 2818–2826. https://doi.org/10.1109/CVPR.2016.308
42. Zhang F, Li M, Zhai G, Liu Y (2020) Multi-branch and Multi-scale Attention Learning for Fine-Grained Visual Categorization, https://doi.org/10.48550/ARXIV.2003.09150
43. Ouyang X et al (2021) Learning Hierarchical Attention for Weakly-Supervised Chest X-Ray Abnormality Localization and Diagnosis. IEEE Trans Med Imaging 40(10):2698–2710. https://doi.org/10.1109/TMI.2020.3042773
44. Xie J, Xiang J, Chen J, Hou X, Zhao X, Shen L (2022) C ^{2} AM: Contrastive learning of Class-agnostic Activation Map for Weakly Supervised Object Localization and Semantic Segmentation, in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA: IEEE, pp. 979–988. https://doi.org/10.1109/CVPR52688.2022.00106
45. Murtaza S, Belharbi S, Pedersoli M, Sarraf A, Granger E (2023) Discriminative Sampling of Proposals in Self-Supervised Transformers for Weakly Supervised Object Localization, in 2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA: IEEE, pp. 1–11. https://doi.org/10.1109/WACVW58289.2023.00021
46. Hajian-Tilaki K (2013) Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian J Intern Med 4(2):627–635
47. McHugh ML (2012) Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 22(3):276–282
48. Powers DMW (2020) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, https://doi.org/10.48550/ARXIV.2010.16061
49. Huang G, Liu Z, van der Maaten L, Weinberger KQ (2016) Densely Connected Convolutional Networks, https://doi.org/10.48550/ARXIV.1608.06993
50. Lin T-Y, RoyChowdhury A, Maji S (2015) Bilinear CNN models for fine-grained visual recognition. In: IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, pp 1449–1457. https://doi.org/10.1109/ICCV.2015.170
51. Cui Y, Zhou F, Wang J, Liu X, Lin Y, Belongie S (2017) Kernel pooling for convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp 3049–3058. https://doi.org/10.1109/CVPR.2017.325
52. Sun M, Yuan Y, Zhou F, Ding E (2018) Multi-Attention Multi-Class Constraint for Fine-grained Image Recognition, https://doi.org/10.48550/ARXIV.1806.05372

53. Dubey A, Gupta O, Guo P, Raskar R, Farrell R, Naik N (2017) Pairwise Confusion for Fine-Grained Visual Classification, https://doi.org/10.48550/ARXIV.1705.08016
54. Yu C, Zhao X, Zheng Q, Zhang P, You X (2018) Hierarchical Bilinear Pooling for Fine-Grained Visual Recognition, https://doi.org/10.48550/ARXIV.1807.09915
55. Wei, XS, Xie, CW, Wu, J (2016) Mask-CNN: localizing parts and selecting descriptors for fine-grained image recognition. arXiv preprint arXiv:1605.06878, [Online]. Available: http://arxiv.org/abs/1605.06878. Accessed 12 Sept 2022
56. Wang Y, Morariu VI, Davis LS (2016) Learning a Discriminative Filter Bank within a CNN for Fine-grained Recognition, https://doi.org/10.48550/ARXIV.1611.09932
57. Yang Z, Luo T, Wang D, Hu Z, Gao J, Wang L (2018) Learning to Navigate for Fine-Grained Classification, in Computer Vision – ECCV 2018, vol. 11218, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., in Lecture Notes in Computer Science, vol. 11218. , Cham: Springer International Publishing, pp. 438–454. https://doi.org/10.1007/978-3-030-01264-9_26
58. Zheng H, Fu J, Zha Z-J, Luo J (2019) Looking for the Devil in the Details: Learning Trilinear Attention Sampling Network for Fine-Grained Image Recognition," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA: IEEE, pp. 5007–5016. https://doi.org/10.1109/CVPR.2019.00515
59. Zhou M, Bai Y, Zhang W, Zhao T, Mei T (2020) Look-Into-Object: Self-Supervised Structure Modeling for Object Recognition, in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: IEEE, pp. 11771–11780. https://doi.org/10.1109/CVPR42600.2020.01179
60. Zheng H, Fu J, Mei T, Luo J (2017) Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition, in 2017 IEEE International Conference on Computer Vision (ICCV), Venice: IEEE, pp. 5219–5227. https://doi.org/10.1109/ICCV.2017.557
61. Sun G, Cholakkal H, Khan S, Khan F, Shao L (2020) Fine-Grained Recognition: Accounting for Subtle Differences between Similar Classes. AAAI 34(07):12047–12054. https://doi.org/10.1609/aaai.v34i07.6882
62. Zhuang P, Wang Y, Qiao Y (2020) Learning Attentive Pairwise Interaction for Fine-Grained Classification. AAAI 34(07):13130–13137. https://doi.org/10.1609/aaai.v34i07.7016
63. Yang X, Wang Y, Chen K, Xu Y, Tian Y (2022) Fine-Grained Object Classification via Self-Supervised Pose Alignment, https://doi.org/10.48550/ARXIV.2203.15987
64. Roy S, Jain PK, Tadepalli K et al (2024) Forward attention-based deep network for classification of breast histopathology image. Multimed Tools Appl. https://doi.org/10.1007/s11042-024-18947-w
65. Roy S, Pal D, Meena T (2024) Explainable artificial intelligence to increase transparency for revolutionizing healthcare ecosystem and the road ahead. Netw Model Anal Health Inform Bioinforma 13:4. https://doi.org/10.1007/s13721-023-00437-y

Authors and Affiliations

Komal Kumar[1] · Snehashis Chakraborty[1] · Kalyan Tadepalli[1,2] · Sudipta Roy[1]

✉ Sudipta Roy
sudipta1.roy@jioinstitute.edu.in

Komal Kumar
komal2.Kumar@jioinstitute.edu.in

Snehashis Chakraborty
snehashis1.C@jioinstitute.edu.in

Kalyan Tadepalli
kalyan.tadepalli@rfhospital.org

[1] Artificial Intelligence & Data Science, Jio Institute, 410206, Navi Mumbai, India

[2] Orthopaedics Department, Sir HN Reliance Foundation Hospital, Girgaon400004, Mumbai, India