Integration of Multi-scale CAM and Attention for Weakly Supervised Defects Localization on Surface Defective Apple

Nguyen Bui Ngoc Han, Ju Hwan Lee, Jin Young Kim

Abstract

Weakly supervised object localization (WSOL) is a task of localizing an object in an image using only image-level labels. Previous studies have followed the conventional class activation mapping (CAM) pipeline. However, we reveal the current CAM approach suffers from problems which cause original CAM could not capture the complete defects features. This work utilizes a convolutional neural network (CNN) pretrained on image-level labels to generate class activation maps in a multi-scale manner to highlight discriminative regions. Additionally, a vision transformer (ViT) pretrained was treated to produce multi-head attention maps as an auxiliary detector. By integrating the CNN-based CAMs and attention maps, our approach localizes defective regions without requiring bounding box or pixel-level supervision during training. We evaluate our approach on a dataset of apple images with only image-level labels of defect categories. Experiments demonstrate our proposed method aligns with several Object Detection models performance, hold a promise for improving localization.

Keywords : WSOL|multi-scale CAM|attention|Surface defective apple

I. INTRODUCTION

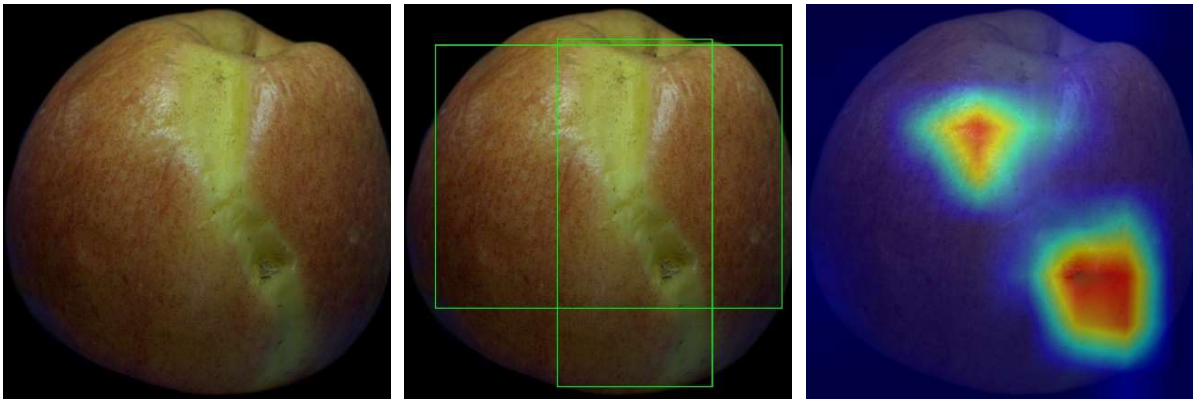
Weakly Supervised Object Localization (WSOL) [1] has revolutionized object detection by learning from image-level labels, bypassing the need for bounding box annotations. Despite its cost-effectiveness and usage of abundant unlabeled data, WSOL has limitations such as reduced accuracy, struggles with complex backgrounds and indistinct objects, and localization uncertainty.

To address these, Class Activation Mapping (CAM) [2] has been integrated with WSOL. For instance, Zhou et al. modified classification architectures like AlexNet [3] and VGG-16 [4] by substituting fully connected layers with a global average pooling layer to aggregate features from the final convolution layer and generate class-discriminative activation maps for localization.

However, there are inherent problems with the CAMs used in WSOL. Since the receptive fields of neural networks are fixed [5], as input scales increase, the activation area ratio for the same neurons decreases. This causes the network to only recognize localized parts of larger objects, resulting in vastly different CAM characteristics across varying input scales [6]. In Figure 1, class activation maps differ according to various input sizes and their localization abilities also differ.

* This work was supported by the Technological Innovation R&D Program(S3294129) funded by the Ministry of SMEs and Startups(MSS, Korea)
* Nguyen Bui Ngoc Han and Ju Hwan Lee contributed equally to this work

Manuscript : 2023.08.22
Revised : 2023.09.06, 2023.09.21
Confirmation of Publication:2023.10.11
Corresponding Author: Jin Young Kim, e-mail: beyondi@jnu.ac.kr



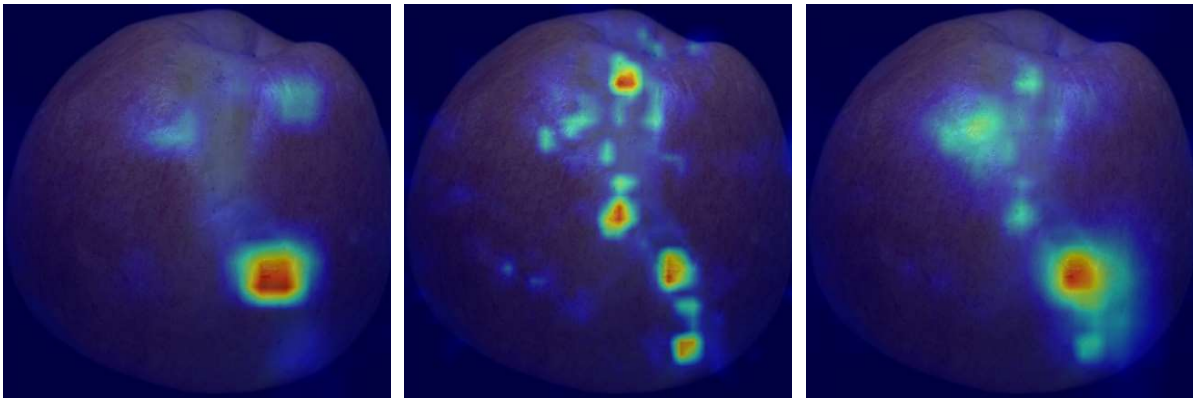


Fig 1. Class-relevant features localization on CAMs image at varying resolutions

For small-scale input images, the general location of objects can be found but the particles are coarse-grained, while CAM maps generated with large-scale inputs contain fine-grained details but have considerable noise. Attempts have been made to complementarily fuse these in a multi-scale framework, but it is difficult to expect fully satisfactory results from such an approach [7].

We attempted to boost localization performance by concatenating attention map insights rooted in the Vision Transformer (ViT) [8] model as an auxiliary branch within Multiscale-CAMs. The use of Multi-Head Attention (MHA) and Multi-Layer Perceptron (MLP) modules serves as a crucial tool for refinement, and self-attention has been proven as a key mechanism in effectively capturing global information [8]. By incorporating the attention map branch, we can minimize the noise prevalent in multiscale CAM. To summarize, the multiscale CAM provided by CNN and the attention map derived from ViT, both generated from a singular input image, offer complementary benefits. This combined approach allows for a more intuitive understanding of both local and global patterns, thereby enhancing the precision in localizing discriminative regions.

In this work we validate our proposed method using images of apples cultivated in an orchard. This dataset, being substantially different from the benchmark datasets [9][10][11], also presents a

higher degree of difficulty. In fact, the performance of CAM-based WSOL is not notably high in datasets demanding such precise localization [6] [7] [14]. However, despite these constraints, our proposed method has proven to outperform traditional CAM-based WSOL. This could be used as a beneficial tool in agriculture, helping to alleviate the need for expert efforts in generating extensive datasets. Further details about the dataset are provided in Section 3.1.

The localization of defects is crucial to classify them according to their severity, among other factors, on the surface area of the affected apple. Working in a weakly-supervised manner minimizes the work of image labeling since we only need image-label annotations. In this way, we avoid creating bounding box-level labels that require more time, effort, and human expertise.

In short, our contributions are summarized as follows:

1) By incorporating the attention map as an auxiliary branch, we have added supplementary information that enables the multi-scale CAM to detect discriminative features with increased precision and interpretability of CNN-based models.

2) Enables benefits of multi-scale attention while retaining the use of pretrained CNN and visual saliency models.

3) Through our proposed method, we have enhanced the localization performance within the apple dataset, achieving more refined results compared to traditional CAM-based WSOL.

The paper is organized as follows: Section 2 reviews related work in the field; Section 3 offers an in-depth description of our proposed algorithm; Section 4 outlines the experiments conducted under different WSOL conditions and presents a comparative analysis involving various methods at the multi-scale level. Section 5 discusses the limitations encountered during our experiments. Finally, Section 6 concludes the paper, summarizing the key insights derived from our research. The main abbreviations and notations used throughout the paper are defined along with their corresponding full terms in Table 1 and Table 2.

Table 1. List of abbreviations

Abbreviations	Definition
MS-C	Multi-scale CAMs
MS-CA	Multi-scale CAMs + Attention
MS-CW	Multi-scale CAMs with Watershed
MS-CS	Multi-scale CAMs with SLIC
MS-CSS	Multi-scale CAMs with Selective search

Table 2. List of notations

Symbol	Description
F_k	Sum of feature map $f^k(x,y)$
S_c	Classification score for class c
w_c^k	Weight corresponding to class c for feature map k
α_k	Average gradient value for channel k
F_{ij}^k	Value of the i-th row and j-th column in the feature map k
$L_{Grad-CAM}$	Grad-CAM heatmap highlighting important areas for class c
$L_{i,c}^j$	Grad-CAM for category c at scale j of image i
$A_{c\_multiscale}$	Fused CAM for category c across multiple scales of image i
A_h	Attention map from the h-th head of the ViT model
$\mathcal{E}_\ell$	The last transformer layer in the ViT model
L_{Att}	Aggregated attention map
S_{final}	Final aggregated attention map
S_{mask}	Segmented mask for precise object localization
Threshold	Threshold value for binarizing the mask

II. RELATED WORK

The primary strategy for addressing the WSOL framework is through the utilization of deep learning methods. Several reasons support this choice. Firstly, feature learning plays a critical role in enhancing the weakly supervised learning process. Secondly, deep CNN models can infer discriminative spatial locations when trained with image-level supervision. Thirdly, pre-training deep models on large-scale training data serves as a simple yet highly effective way to encode valuable cues for the weakly supervised learning process [15].

One approach to weakly supervised object localization using deep learning is to apply CAM techniques. Since CAM approach was initially introduced, most of previous studies on WSOL have followed its convention e.g., Grad-CAM [16], HaS [17], ACoL [1], Score-CAM [18], SPG [19], ADL [20], DANet [21], Ablation-CAM [22]. Regardless of their differences, they all follow a similar pipeline to generate CAMs, they modify the network structure and use the global pooling layer instead of the fully connected layer for feature fusion. However, these methods differ in the way each feature map's weights are generated. CAM obtains weights from fully connected layers, Grad-CAM weight the 2D activations by the average gradient, Ablation-CAM zero out activations and measure how the output drops, Score-CAM perturbate the image by the scaled activations and measure how the output drops. HaS, ACoL add convolutional classification layers on top of the backbone to generate CAMs directly.

In several works [6] [7] [23] [24], the authors analyzed the potential bottlenecks of CAM-based approaches for WSOL framework:

1) Insufficient data and similar appearance features of some categories causes confusion and discards low discriminative features.

2) The network cannot learn the complete object features since the perception range of neurons is limited, even the top neurons can only perceive a part of the image [25].

3) Objects have discriminative features at different scales, so using a single network scale only activates limited features.

III. PROPOSED METHOD

In this section, we present the details of the custom dataset used in our study, specifically designed to address the challenges of weakly supervised object localization in the context of skin apple disease diagnosis. Then we describe the proposed method with a detailed description of each phase.

3.1. Differences from the benchmark

WSOL benchmark datasets, such as ImageNet [9], COCO [10], and OpenImages [11], are characterized by their vast volume, class diversity, comprehensive annotation, and compact resolution. Our custom dataset, however, differs significantly. Challenges with this dataset arise from the complexity and variability of the defective labels. The boundaries between the labels are often

blurred, making precise classification and localization more difficult.

Despite the challenges, the specialized nature of our dataset allows us to concentrate on fine-grained features and specific challenges related to apple defects, potentially leading to more precise localization models.

3.2. Surface defective apple dataset

The Surface Defective Apple (SDA) dataset comprises apples from six distinct categories: normal, physiological, scratch, malformation, blight, and others. These categories were meticulously defined based on the presence of surface defects such as irregular patterns, pest infestations, and morphological or physiological anomalies.

The data collection process involved taking high-resolution images of apples with an industrial-grade camera. In total, we collected 12,000 images from 2,000 Fuji apples harvested from an orchard in Jangseong-gun, South Korea, capturing six images per apple. The resolution of the images was 2448x2048 pixels.

To prepare the dataset for model training and testing, experts manually annotated all images in two different ways. Initially, each image was individually categorized into one of the six different classes. As depicted in Table 3, the instance-label dataset contained 254 normal, 30 physiological, 694 scratch, 5 malformations, 1010 blight, and 125 images classified as 'others.' It is noteworthy that the last five labels are considered defective cases, which will be employed in subsequent analysis.

Next, class-label annotation was performed, wherein only two primary labels – defective and normal – were considered. Due to the substantial quantity of images to be classified, it should be noted that no information about the defect location (bounding box annotation) was utilized, which simplified the task for the annotating experts.

The primary purpose of the first dataset was to evaluate the localization process, while the second was utilized solely for the training phase. The second dataset was further divided, with 90% of the apple images used for model training and validation, and the remaining 10% designated for testing. Concurrently, the first dataset was employed to validate the inference process.

Table 3. Data on number of Class label and Instance label

	Class-label	Instance-label
Normal	6,201	255
Defective	3,799	823
Total	10,000	1,078

Background noise refers to irrelevant or distracting elements in the image that are not part of the target object, in this case, the apple. Therefore, we utilize the apple segmentation algorithm. This algorithm effectively identifies and extracts the apple region from the rest of the image, eliminating irrelevant background elements. By focusing solely on the apple area, we ensure that the CNN-based classifier receives clean and relevant input data, free from any potential confounding factors [31] [32]. Following image segmentation, we cropped the isolated apple region to a standardized size of 1024x1024 pixels and got the final image.

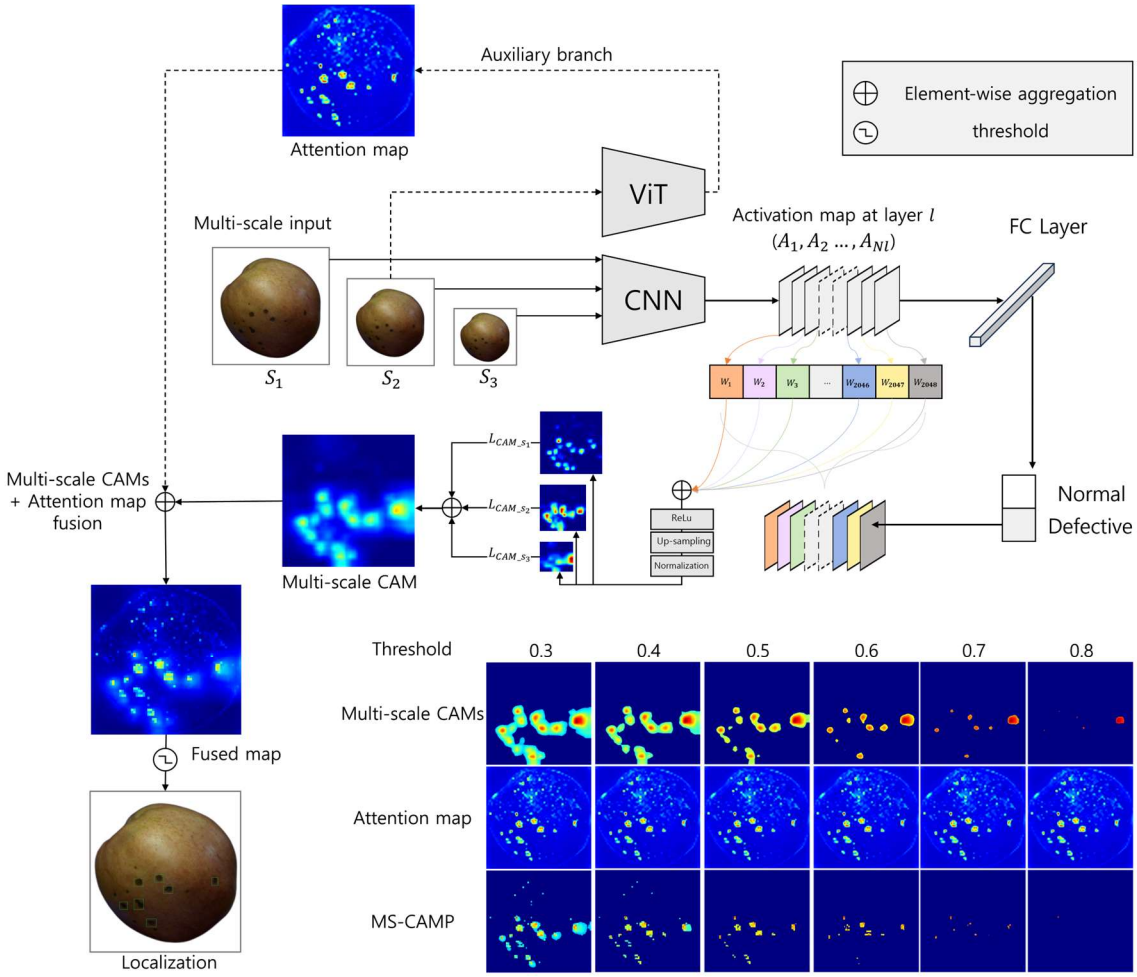


Fig 2. Proposed method: fused map (MS−CA) obtained by fusing a multiscale CAM (MS−C) with an attention map obtained by crossing one of the multi−scale inputs through an auxiliary branch.

3.3. Multi−scale CAM based localization with Attention.

In this section, we present an MS−CA for localizing discriminative features from apple images in the SDA dataset. Our entire process is depicted in Figure 2. Our approach consists of three main phases:

1. Extracting samples from each input image at multiple different scales.

2. Training CNN to classify each scale into two classes. This localizes defects, highlighting key image regions guiding CNN's decision−making. At this stage, a fusion method is introduced.

3. Utilizing fusion maps from the previous phase to attain localization results.

3.3.1. Localization with Multi−scale CAM

To extract the most discriminative target areas using a well−trained classification network [33] [34] [35], we employ Grad−CAM, an evolution of the CAM [2] method. To achieve weakly supervised localization, we modify the CNN backbone model by replacing the last dense layer with a Global Average Pooling (GAP) layer. This GAP layer performs average pooling on k feature maps denoted as $f^k(c, y)$ from the last convolutional layer, where $F^k = \sum_{x,y} f^k(x, y)$. The resultant spatially pooled values are then fully connected to output

classification score S_c and the weight w_c^k corresponding to class c. The weight α_k for each channel k is calculated as the average gradient value:

$$S_c = \sum_k w_c^k \sum_{x,y} f^k(x,y) = \sum_{x,y} \sum_k w_c^k f^k(x,y) \quad (1)$$

$$\alpha_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial S_c}{\partial F_{ij}^k} \quad (2)$$

In Equation (1,2), y is the prediction score of the network, F_{ij}^k signifies the value of the $i-$th row and $j-$th in the feature map of channel k. And Z denotes the product of the feature map's width and height.

Subsequently, a weighted summation operation is performed using weight α_k and feature map A^k for each channel. The ReLU activation is then applied to filter out negative values, resulting in the final Grad-CAM map $L_{Grad-CAM}$:

$$L_{Grad-CAM} = \mathrm{ReLU}\left(\sum_k \alpha_k F^k\right) \quad (3)$$

In Equation (3), F_k represents the $k-$th channel within the feature layer F. w_c^k represents the weight associated with the $k-$th channel in the feature layer. This resulting heatmap $L_{Grad-CAM}$ highlights the important areas for class c, providing valuable insights into how the model makes its decisions.

Small defects like decayed areas, blemishes, or scratches are common in apple images. But a key issue is that single CAM only captures parts relevant to certain categories, often differing from the actual truth. Applying rescaling transformations to input images does not guarantee the same changes in generated CAMs, due to a gap between full and weak supervision.

To address this, inspired by [6] [7] [25] [26] [27], we merge CAMs from multi-scale images to gather varied and complementary objects. This approach is motivated by the question of whether the discriminative parts the network relies on for recognition remain consistent for objects of varying scales within the same category in the dataset.

For a given input image i, we sample it n times by setting different sampling rates result in S_1, S_2, S_3 image. Thus, $L_{i,c}^j$ represents the Grad-CAM of category c corresponding to the scale j of image i. MS-C are then averaged after resizing, resulting in $A_{c\_multiscale}$. According to equation 4, The fused CAM $A_{c\_multiscale}$ is obtained through:

$$A_{c\_multiscale} = \sum_j^n \frac{L_{i,c}^j}{n} \quad (4)$$

3.3.2. Attention map for refinement in Multi-scale CAM.

In contrast to prior attention-based CNN methods that integrate attention mechanisms directly into the backbone architecture, our approach generates CAMs and attention maps separately outside the CNN backbone without modifying the pretrained architecture.

CAMs are extracted at multiple sampling scales to provide class-specific localization information at varying

resolutions. Meanwhile, the attention map gives a measure of visual saliency. The key is merging both in a late fusion manner, which aims to distinguish class-relevant region from fused map and provide more accurate defects. Additionally, a key advantage of this approach is the flexibility to extract CAMs and attention map from any existing pretrained CNN classifier without needing to retrain the model to include attention.

To generate attention maps, we utilize a ViT backbone as a visual saliency model. Specifically, multi-head self-attention maps are extracted from input images using the pretrained ViT model. The model comprises transformer layers with self-attention heads that focus on relevant features. To visualize map, the last layer was chosen to extract maps from multiple heads, each captures distinct image patterns. We aggregated information across heads into comprehensive attention maps L_{Att}:

$$L_{Att} = \frac{1}{N}\sum_{h=1}^{H}\left(A_h\left(\mathscr{E}_{\ell}(I)\right)\right) \quad (5)$$

In this formula, L_{Att} denotes the aggregated attention map, where A_h stands for the attention map from h-th head, $\mathscr{E}_{\ell}$ refers to the last transformer layer, and I represents the image.

We normalized the MS-C to a 0-1 range to ensure consistent activation values. The MS-C in Equation (4) were then combined with auxiliary branch attention map L_{Att} from Equation (5) using the mean aggregation to obtain the aggregated attention map S_{final}:

$$S_{final} = \frac{L_{Att} + A_{c\_multiscale}}{2} \quad (6)$$

We applied a threshold to S_{final} to obtain a segmented mask, removing regions with activation below the threshold. The final S_{mask} provides precise localization of the objects of interest to generate bounding boxes on connected areas, with each box corresponding to an instance. the thresholding process is defined as:

$$S_{mask}(x,y) = \begin{cases} 1 \text{ if } S_{final}(x,y) \geq \text{threshold} \\ 0, \text{otherwise} \end{cases} \quad (7)$$

IV. EXPERIMENTS

This section describes the datasets, experiments results obtained in each phase of the proposed method and their comparison with other methods.

We implemented the proposed method in Python using Pytorch framework on an Ubuntu 20.04 machine with Intel Xeon CPU, Nvidia Tesla V100 GPU, and 32GB memory.

We evaluate the weakly supervised object classification and localization performance of the proposed method on our specific dataset.

A binary-class test set containing 1078 apple images with visually labeled defects was tested. After being collected and pre-processed as described in Section 3-1-3, the labeled images are classified as the training set (60%), validation set (30%), and testing set (10%). We trained ResNet50 for classification on the training set. The architecture was evaluated on the test set to obtain localization results. We assess performance at both image and instance level.

Performance at image level. To evaluate apple surface classification, we used precision, recall and F1-score. Precision measures positive predictions that are correct. Recall measures positive cases correctly detected. F1-score combines precision and recall into a single metric.

Performance at instance level. For localization evaluation, we used mean Intersection over Union (mIoU). mIoU measures spatial overlap between predicted and ground-truth boxes, averaged over all instances. Higher mIoU indicates more accurate localization. The metrics expressed as follows:

$$IoU_c = \frac{TP_c}{TP_c + FP_c + FN_c} \quad (8)$$

$$mIoU = \frac{1}{C}\sum_{c} IoU_c \quad (9)$$

To assess our method's performance, we incorporate additional techniques to enhance object localization. We combine MS-C with three refinement methods: SLIC (Simple Linear Iterative Clustering) [36] (MS-CS), Selective Search [37] (MS-CSS), and Watershed [38] (MS-CW). These serve as post-processing steps to refine localization accuracy and delineate object boundaries. Across all approaches, we apply Non-Maximum Suppression with a 0.5 threshold to eliminate redundant bounding boxes, retaining the most confident predictions.

4.1. Experimenting with CNNs classifier selection

We fine-tuned ResNet50 [33], EfficientNet [34], and GoogLeNet [35] on 512x512 images. Input images were standardized by subtracting the mean and dividing by standard deviation. Data augmentation consisted of flipping and rotation. The Adam optimizer was used for training the models with a base learning rate of 0.001 and weight decay factor of 0.0001 to regularize the model. The learning rate was scheduled to decrease by 10% after 10 epochs. The network was trained with batch size of 16 and the loss function used for training all models was cross-entropy.

Table 4. CNN classifier performance (%)

CNN Classifier	Recall	Precision	F1-Score
ResNet50	94.50	94.44	94.47
EfficientNet-V2	94.62	84.32	89.14
Inception-V4	94.13	88.22	91.15

ResNet50 achieved the best average F1-score of 94.47%, compared to 91.15% and 89.14% for EfficientNet and GoogLeNet. ResNet50 highlighted discriminative features for localization while maintaining good classification accuracy.

We adopted ResNet50 as the CNN backbone given its class-specific features and high F1 performance. Results are shown in Table 4.

4.2. Experimenting with units of scale in CAM and Attention map

After training, we generated CAMs to highlight important object regions. To evaluate localization, CAMs were thresholded to create segmented heatmaps. By varying the threshold percentage of the max CAM value, we derived pseudo masks that localized defects.

We validated performance across different thresholds. Table 5 shows mIoU scores corresponding to each threshold choice. A threshold of 0.8 gave optimal mIoU for multi-scale CAMs.

Table 5. Localization accuracy measured by mIoU (%) for scale and corresponding threshold values.

Scale \ Threshold	0.4	0.5	0.6	0.7	0.8
One-scale	11.21	12.53	15.00	17.26	17.50
Two-scale	11.05	12.95	16.27	16.28	13.83
Multi-scale	12.78	16.52	21.65	27.60	31.46

For multi-scale CAMs, we sampled the image at rates of 0.25, 0.5 and 1.0. Since the CNN classifier was trained on 512x512 images, using these sampling rates allows evaluating CAMs on the same scale (0.5x), a smaller scale (0.25x), and the full original scale (1x) of the images.

To generate the saliency maps, we utilized a DINO ViT [34] small pre-trained on ImageNet dataset. This Vision Transformer architecture contained a backbone with the following configuration: image patches of size 8x8, 12 transformer blocks in depth, 6 attention heads and layer normalization applied after each block.

As shown in Fig 1, original single-scale CAMs only gave a general defects area, differing in shape and size from the ground truth. In contrast, refined multi-scale CAMs were more consistent with actual defects, enabling improved detection of multiple areas. As seen in Table 6, multi-scale CAMs outperformed single scale for localization, achieving higher mIoU across thresholds. This shows fusing multiple scales better captures target objects with enhanced accuracy.

4.3. Quantitative and Qualitative Comparisons

We found that the MS-C performed better when three scales were utilized in terms of localization performance. We compared the performance of our proposed MS-CA, which introduced an auxiliary branch for refinement, both quantitatively and qualitatively, against other refinement methods. Both quantitative and qualitative analysis were able to evaluate model ability in capturing precise defected localization.

4.3.1. Quantitative comparisons

Among the refinements we compared with, the MS-CS resulted in varied segmentations depending on the number of the super-pixel. We thus divided its results based on the number of super-pixel values of 100, 200, 400, and 800. As depicted in Table 6, the mIOU scores varied according to the threshold, and within the same threshold, we observed that different refinement methods yielded different scores. Our proposed method, MS-CA, achieved the highest score of 39.40% at a threshold of 0.5. Most methods showed a lower mIOU at this threshold. At another threshold setting of 0.8, MS-CW recorded the highest mIOU at 32.29%, which was still 7.11% lower than our proposed method.

4.3.2. Qualitative comparisons

The combination of MS-C and selective search had a limited impact on the results in Table 6, so it was excluded from this section. Referring to the previously presented experimental outcomes,

Table 6. Quantitative comparison of localization accuracy across different threshold values measured by mIoU (%)

Refinement \ Threshold	0.4	0.5	0.6	0.7	0.8
MS-CS (100)	11.59	06.89	04.33	1.92	0.57
MS-CS (200)	11.39	11.45	11.06	8.00	0.76
MS-CS (400)	10.23	13.08	14.86	14.56	02.28
MS-CS (800)	09.29	12.69	16.14	19.27	18.08
MS-CSS	01.81	00.74	00.74	0.22	0.11
MS-CW	12.80	16.70	22.12	28.39	32.29
Proposed method (MS-CA)	31.82	**39.40**	16.66	1.08	0

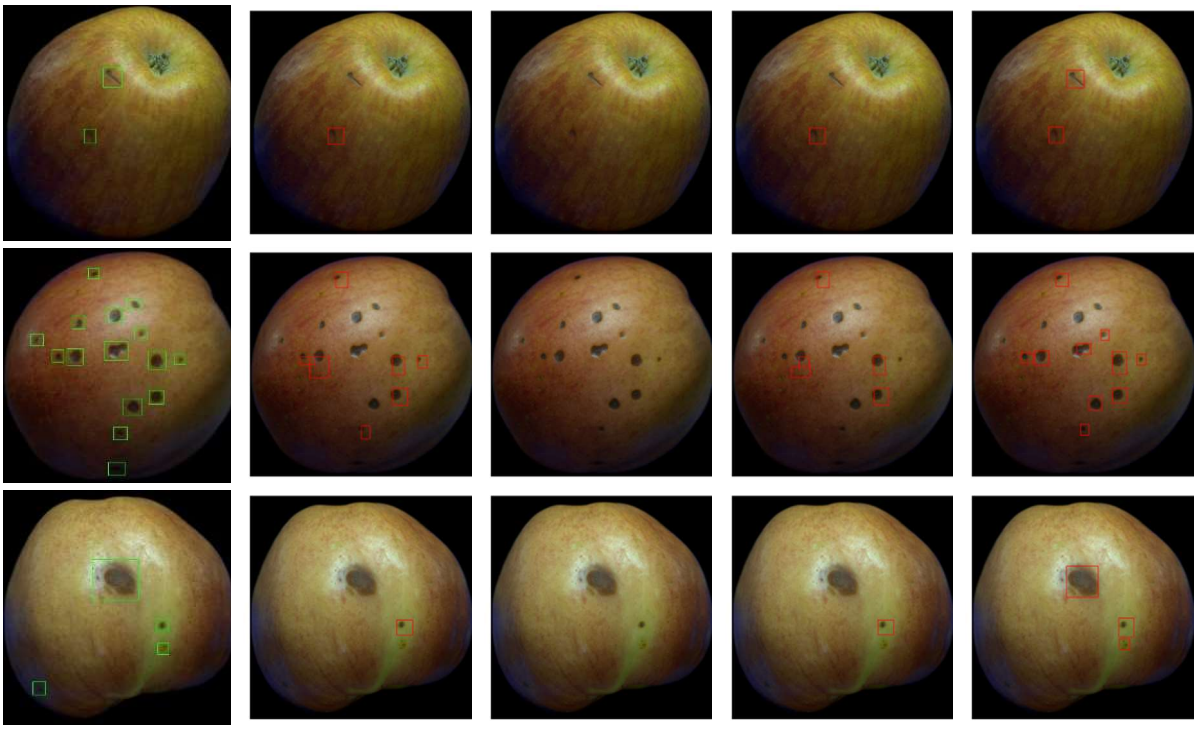


Fig 3. Qualitative comparisons on Surface defective apple dataset

integrating various multi-scale techniques, and selecting thresholds resulted in corresponding mIoU achievements for each method. Figure 3 clearly demonstrated distinct differences in the accuracy and efficiency of each method.

The first row of Figure 3 allowed us to evaluate how accurately each method localized minor defects. While some methods either detected just one or none, our proposed approach could identify both. The second row of Figure 3 evaluated the ability of each method to detect many similar defective areas. It assessed how many of these defects could be accurately identified by each method. Relatively, our method localized accurately, excluding a few areas. However, when compared to ground-truth, there were cases where localization was imprecise. The last row of Figure 3 assessed the ability to correctly localize when presented with both large and minor areas simultaneously. Our proposed method proved superior in accurate localization. Yet, identifying very fine areas obscured by shadows was

challenging, a challenge that was consistent across all refinement methods. In conclusion, based on the quantitative evaluation, it was evident that our proposed method demonstrated better performance in accurately localizing various types of defective areas.

V. Discussion

5.1. Object detection

We thoroughly evaluated our proposed apple defect detection architecture against established object detection models [31] [32], [33]. Compared with Faster R-CNN [31] and RetinaNet [32], both employing ResNet as their feature extraction backbone, our method showed competitive performance.

Table 7. Object Detection performance in terms of apple defects localization

Model	Faster R-CNN	RetinaNet	SSD
mAP (%)	36.00	39.47	36.14

On the test dataset with instance-level annotations, as shown in Table 7, Faster R-CNN achieved a mAP score of 36%, RetinaNet scored 39.47%, while our method achieved a mIoU score of 39.40%. These results demonstrate the effectiveness of our method in defects localization, even without an additional bounding box refinement module. Notably, the exclusion of the 'normal' class during both training and evaluation aligns the mAP score with the mIoU score in our context. Additionally, when extending our analysis to include the SSD [33] model using VGG as its backbone, we found a mAP score of 36.14%, which highlighted the SSD model's effective defect localization capabilities. Our approach performed comparably to detection models, highlighting its capability for accurate defect localization using only image-level labels.

5.2. Reasons for low IoU.

Our achieved mIoU of 39.40% highlights challenges posed by our complex custom dataset. The presence of diverse surface defects posed unique challenges in localization under weak supervision with only class labels. The industrial image capture introduced potential domain shifts in lighting, image quality and viewpoints affecting generalization.

Compared to fully supervised object detection model, we see similar results arising from annotation limits and dataset complexity. Despite the moderate mIoU, our method aligns with object detection models, reflecting task difficulty given the data complexity and weak supervision.
While not perfect, our approach performs comparably on this challenging dataset using only class labels. Results highlight inherent trade-offs between supervision and localization precision.

Moving forward, we expect that our proposed method can work well on benchmark datasets, which have relatively large quantities and more easily recognizable objects.

VI. Conclusion

In this study, we present a weakly supervised deep learning that integrates the image classification and localization of defected apples. Our proposed weakly

supervised object localization method can be used in an industrial application due to its high efficiency and speed. Besides, it only requires an image-level labeled dataset, which is less time-consuming, reduces the reliance on fully annotated data and improves efficiency. Future directions for improvement involve exploring data augmentation techniques, domain adaptation, or fine-tuning strategies to enhance localization performance on our specific dataset.

REFERENCES

[1] X. Zhang, Y. Wei, J. Feng, Y. Yang and T. Huang, "Adversarial Complementary Learning for Weakly Supervised Object Localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019.

[2] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[3] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in neural information processing systems 25*, 2012.

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

[5] J. Long, N. Zhang and T. Darrell, "Do convnets learn correspondence?," in *Advances in neural information processing systems, 27*, 2014.

[6] B. Wang, C. Yuan, B. Li, X. Ding, Z. Li, Y. Wu and W. Hu, "Multi-scale low discriminative feature reactivation for weakly supervised object localization," *IEEE Transactions on Image Processing,* vol. 30, pp. 6050-6065, 2021.

[7] X. Zhou, Y. Li, G. Cao and W. Cao, "Master-CAM: Multi-scale fusion guided by Master map for high-quality class activation maps," *Displays,* vol. 76, p. 102339, 2023.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, 2020.

[9] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick and P. Dollár, "Microsoft COCO: Common Objects in Context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 2014.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009.

[11] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig and V. Ferrari, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *International Journal of Computer Vision,* vol. 128(7), pp. 1956-1981, 2020.

[12] W. Bae, J. Noh and G. Kim, "Rethinking class activation mapping for weakly supervised object localization," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August*

23–28, 2020, Proceedings, Part XV 16, 2020.

[13] D. Zhang, J. Han, G. Cheng and M.-H. Yang, "Weakly Supervised Object Localization and Detection: A Survey," *IEEE transactions on pattern analysis and machine intelligence,* vol. 44(9), pp. 5866-5885, 2021.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in *Proceedings of the IEEE international conference on computer vision*, 2017.

[15] K. K. Singh and Y. J. Lee, "Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-supervised Object and Action Localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[16] Z. W. Haofan Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel and X. Hu, "Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020.

[17] X. Zhang, Y. Wei, G. Kang, Y. Yang and T. Huang, "Self-produced Guidance for Weakly-supervised Object Localization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[18] J. Choe and H. Shim, "Attention-based Dropout Layer for Weakly Supervised Object Localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[19] H. Xue, C. Liu, F. Wan, J. Jiao, X. Ji and Q. Ye, "DANet: Divergent Activation for Weakly Supervised Object Localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[20] S. Desai and H. G. Ramaswamy, "Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020.

[21] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba, "Object Detectors Emerge in Deep Scene CNNs," arXiv preprint arXiv:1412.6856, 2014.

[22] X. Ma, Z. Ji, S. Niu, T. Leng, D. L. Rubin and Q. Chen, "MS-CAM: Multi-Scale Class Activation Maps for Weakly-Supervised Segmentation of Geographic Atrophy Lesions in SD-OCT Images," *IEEE Journal of Biomedical and Health Informatics,* vol. 24, no. 12, pp. 3443-3455, 2020.

[23] T. Liu, H. Zheng, J. Bao, P. Zheng, J. Wang, C. Yang and J. Gu, "An Explainable Laser Welding Defect Recognition Method Based on Multi-Scale Class Activation Mapping," *IEEE Transactions on Instrumentation and Measurement,* vol. 71, pp. 1-12, 2022.

[24] C. Robinson, L. Hou, K. Malkin, R. Soobitsky, J. Czawlytko and B. Dilkina, "Large Scale High-Resolution Land Cover Mapping with Multi-Resolution Data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[25] K. KC, Z. Yin, D. Li and Z. Wu, "Impacts of background removal on convolutional neural networks for plant disease classification in-situ," *Agriculture,* vol. 11, no. 9, p. 827, 2021.

[26] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[27] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *In International conference on machine learning*, 2019.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[29] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, "Pyramid Scene Parsing Network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[30] G. Lin, A. Milan, C. Shen and I. Reid, "RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[31] K. H. R. G. J. S. Shaoqing Ren, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in neural information processing systems*, 2015.

[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," in *Proceedings of the IEEE international conference on computer vision*, 2017.

[33] D. A. D. E. Wei Liu, C. Szegedy, S. Reed, C.-Y. Fu and A. C. Berg, "SSD: Single Shot MultiBox Detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 2016.

[34] Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

Authors

Nguyen Bui Ngoc Han

She received a B.S. degree in Information Technology from Saigon University of Vietnam in 2019. Since 2021, She is currently pursuing a Master degree in ICT System and Convergence from Chonnam National University, South Korea. Her research interests are Machine Learning, Deep Learning, Computer vision.

Ju Hwan Lee

He received a B.S. degree in Department of Earth and Environmental Sciences from Chonnam National University of South Korea. Since 2019, he is currently pursuing a Integrated PhD program in ICT System and Convergence from Chonnam National University, South Korea. His research interests are Deep Learning, Computer vision, Knowledge Distillation, and ExplainableAI.

Jin Young Kim

He received his B.S., M.S. and Ph.D. degree in Computer Science from Seoul National University, South Korea in 1986, 1988 and 1994, respectively.
Since 1995, he has been a research professor in Department of Electrical Engineering, Chonnam National University, South Korea. His research interests are Digital signal processing, video processing, audio signal processing, machine learning, deep learning.