IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 June 2024, accepted 26 June 2024, date of publication 4 July 2024, date of current version 26 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3423697

APPLIED RESEARCH

A Weakly Supervised Learning Framework Utilizing Enhanced Class Activation Map for Object Detection in Construction Sites

JAEHUN YANG[1], EUNJU LEE[2], JUNEHYOUNG KWON[3], DONGMIN LEE[1], YOUNGBIN KIM[2,3], (Member, IEEE), CHANSIK PARK[1], AND DOYEOP LEE[1]

[1]ConTILab, Department of Architectural Engineering, Chung-Ang University, Seoul 06974, South Korea
[2]Department of Image Science and Arts, Chung-Ang University, Seoul 06974, South Korea
[3]Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Doyeop Lee (doyeop@cau.ac.kr)

This work was supported in part by the National Research and Development Project for Smart Construction Technology funded by Korea Agency for Infrastructure Technology Advancement through the Ministry of Land, Infrastructure, and Transport Managed by Korea Expressway Corporation under Grant RS-2020-KA156291; and in part by the National Research Foundation of Korea (NRF) Grant funded by Korea Government, Ministry of Science and ICT (MSIT), under Grant 2022R1G1A1012897.

ABSTRACT Computer vision has emerged as a promising tool for improving safety at construction sites through automatic scene recognition. However, traditional approaches require significant labor-intensive and time-consuming efforts for annotations. Although weakly supervised learning has the advantage of localizing objects without location information, the conventional class activation map (CAM) technique struggles with small and linear object localization and background noise at construction sites. For effective scene recognition related to construction safety, the spatial relationships between precisely localized objects are crucial. Due to the limitations of traditional CAM techniques in localizing small and linear objects, using CAM for construction safety monitoring remains inaccurate. Therefore, this study proposes a weakly supervised learning approach with an improved CAM for enhancing object detection in construction sites. The improved CAM localizes objects of various scales and is robust against background interference. Spatial relationships between localized objects are employed to determine the status of scenes for construction safety monitoring. Experiment using datasets associated with falls from ladders (FFL) demonstrates that the improved CAM surpasses the traditional CAM in mIoU (mean intersection over union) for the object localization performance as well as in the accuracy and F1-Score for recognizing unsafe scenes. This demonstrates the robust potential of employing CAM as a weakly supervised learning strategy, underlining its substantial feasibility for preventing hazards in construction sites. The proposed framework can minimize annotation efforts, demonstrating the potential of CAM as a viable computer vision technique for efficiently detecting hazards at construction sites.

INDEX TERMS Computer vision, weakly supervised learning, class activation map, scene recognition, safety monitoring.

I. INTRODUCTION

Computer vision has great potential for preventing accidents at construction sites through automatic scene recognition. Following significant advancements in deep learning, computer vision methods have been applied to effectively identify and recognize unsafe behaviors and conditions [1]; detection speed and accuracy have greatly improved in recent years. Object localization and classification are fundamental computer vision tasks that are required for implementing

The associate editor coordinating the review of this manuscript and approving it for publication was Mingbo Zhao.

subsequent tasks at construction sites, such as object tracking and action recognition, and for scene-based identification of unsafe conditions and behaviors (missing safety equipment, failure to wear personal protective equipment (PPE)) [2]. Deep learning-based computer vision applications require datasets to accomplish such tasks. Customized datasets featuring objects specific to construction sites are required to train deep-learning models for object detection, including localization and classification, in the construction industry [3]. However, accumulating datasets is labor-intensive and time-consuming. The requirement for a large amount of annotated data, including location annotations such as segmentation and bounding boxes, is a major challenge in the application of computer vision [4]. Given the distinct challenges inherent in industrial applications, annotation of specialized datasets tailored to specific domains is imperative [5]. In the construction industry, the characteristics of a given scene tend to vary significantly from one construction site to another. Thus, computer vision applications used at one site cannot be assumed to translate seamlessly to other sites. This is particularly evident for construction safety monitoring; a large custom dataset is necessary to address unique requirements to ensure safety in diverse construction environments. In addition, determining the standards for labeling objects (e.g., annotating range and method, and processing occlusion) is challenging because the diverse shapes and rules in construction sites make it difficult to establish clear definitions [6], potentially affecting the performance of the computer vision model [7].

Weakly supervised learning has garnered significant interest in computer vision research owing to the time- and labor-intensive processes of accumulating datasets. Weakly supervised learning is a type of deep learning in which the model is trained by leveraging weaker forms of labels to generate annotated data, and is less costly and time-consuming than using fully annotated data. The annotated data used in weakly supervised learning may be noisy or incomplete and may not contain all information required for accurate classification or regression [8]. The class activation map (CAM) technique has been extensively applied to images in weakly supervised learning. A CAM enables localization of image regions relevant to a specific class [9], [10]. It can be used for visualization where a neural network is ''looking'' at an image to make a classification decision. As the CAM can be generated only using image-level annotations, it is frequently utilized in weakly supervised learning method for capturing location of objects. The CAM can be used to generate a heat map of an image that highlights the regions most relevant to the predicted class, showcasing its potential for object detection problems.

Despite its promising potential, CAM technique remains a challenge for site safety monitoring in the construction industry. For recognizing the hazards at construction sites, the spatial relationship between multiple objects is a primary correlation required to detect the hazards in a scene [11], [12]. The location information of an object has a major influence in determining the status of the scene (safe or unsafe) using spatial relationships (correlations between objects); thus, precise object localization is required. However, applying CAM technique to small-object detection remains challenging and often results in visualization with over-highlighted backgrounds rather than the target objects [13]. Moreover, localization of small and linear objects at construction sites is often interrupted by unrelated backgrounds. Because CAM is based on weak label information (a dataset without location information), object detection in a complex (cluttered) background or a background with a shape and texture similar to that of a target object (light-gray scaffolding in front of light-gray concrete) becomes even more difficult. Developing a sophisticated, CAM architecture tailored to construction sites is imperative for exploiting the full benefits of a weakly supervised learning approach for construction safety monitoring. Such an architecture should be capable of localizing objects with precision, while effectively addressing the challenges of extracting small and linear objects at construction sites. Moreover, a strategy for identifying hazardous status through the utilization of spatial relationships among detected objects must be formulated to operationalize this architecture in real-world cases.

This study proposes a weakly supervised learning approach with an improved CAM (small and linear objects in construction sites, SOS-CAM) using only image-level labels for construction safety monitoring while addressing the aforementioned gaps. Since the SOS-CAM can provide more precise location information, spatial relationships to recognize the scene can be leveraged to monitor construction safety. To validate the feasibility of the proposed approach, this study investigated previous research that used spatial relationships to recognize scenes, prepared a specific ladder-related dataset composed of image-level labels, and evaluated it through quantitative and qualitative analyses. This study makes the following contributions: (1) To the best of our knowledge, this study represents the first attempt to exploit the location information of detected objects using a weakly supervised learning approach for scene recognition at construction sites. Owing to the improved performance of the SOS-CAM in detecting objects, (2) application of spatial relationships in scene recognition is more seamless, establishing the potential applicability of the CAM. (3) Attention can be revitalized with the use of the CAM for recognizing scenes on the application of computer vision to detect small and linear objects at construction sites, and (4) this approach enables the use of a large number of web images without requiring bounding boxes, and can improve model generalization.

The remainder of this paper is organized as follows. Section II reviews the current state of computer vision technology for construction safety and weakly supervised learning methods. The SOS-CAM architecture and process of applying spatial relationships are presented in Section III. The dataset preparation, model training procedures, and results

are described in Section IV. Section V presents the discussion points, significant contributions, and limitations. Conclusions and future research are presented in Section VI.

II. RELATED WORKS

A. CURRENT COMPUTER VISION APPROACHES FOR CONSTRUCTION SAFETY MONITORING

The construction industry is widely recognized as being the most hazardous due to its complex and dynamic environment [1], [2], [14], [15]. According to the Occupational Safety and Health Administration (OSHA), approximately 20% of worker fatalities occur in the construction industry [16]. Site monitoring to assess rule compliance is crucial and frequently utilized in construction to assess the potential risks associated with ongoing work and the current state of the site [2]. However, such observational methods can be costly and time-consuming because they require manual observations by supervisors or managers [16], [17]. Manual observation is hindered by untimely and inaccurate or missing information [18]. The severity of the construction industry's challenges has prompted extensive research on computer vision to prevent various accidents and incidents.

Object detection, including object localization and classification, provides one of the most basic pieces of information for various tasks of computer vision [19], [20]. Many studies have sought to detect specific objects, such as PPE and workers, in construction site monitoring using computer vision's object detection capabilities as a means to mitigate construction accidents [21], [22], [23], [24]. Beyond the object detection, there were considerable efforts to use spatial relationships, which means the correlation between detected objects, to recognize the scene status [11], [12]. Mneymneh et al. [21] coordinated a bounding box to identify workers wearing PPE. Chern et al. [25] used the overlap ratio of detected PPE and workers as the primary logic to verify whether a worker was equipped with proper PPE, including a hardhat, safety harness, safety strap, and safety hook. Fang et al. [26] combined computer vision algorithms with ontology to construct relationships between objects and automatically identify hazards using the coordinates of the detected objects. Khan et al. [12] proposed a handcrafted rule-compliance algorithm based on the coordinates of bounding boxes using a Mask R-CNN for mobile scaffolding. Anjum et al. [27] used an SSD for object detection and the coordinates of bounding boxes to check for rule compliance during ladder operations. The spatial relationships in previous studies that used bounding box coordinates can be summarized as follows: (a) WITHIN; (b) OVERLAP; (c) AWAY [26] [12]; (d) calculating size [27], as shown in Fig. 1. Most studies have used the coordinates of detected objects to leverage spatial relationships, reporting successful scene recognition.

Meanwhile, in order to apply spatial relationships between objects, the foremost prerequisite is the successful implementation of object detection. To achieve this, specific datasets tailored to construction domain to training the model on the objects related to construction safety is imperative [3]. Domain-specific dataset plays a crucial role in enabling the object detection process and subsequently, facilitating the application of spatial relationships among the detected objects. Several studies have accumulated datasets for the construction industry. Xuehui et al. [7] released a dataset that included moving objects at construction sites with polygon and bounding-box annotations. In their dataset, 41,668 images comprising workers and heavy equipment including tower cranes, excavators, and loaders were collected; four experts performed precise annotations. To overcome the challenges of detecting hardhats, Wu et al. [28] proposed a benchmark dataset for hardhats with bounding box annotations composed of numerous small-scale instances less than 32×32 pixels.

However, the methods of accumulating datasets used in previous studies were labor-intensive and time-consuming [29], [30]. Large-scale datasets require precise annotation for object localization [31], [32]; annotation accuracy affects training and model performance [32]. Although precise annotation is important, establishing a standard of location annotations for objects at construction sites is difficult due to ambiguity resulting from missing information and complexity.

B. LEVERAGING WEAKLY SUPERVISED LEARNING FOR ADDRESSING CHALLENGES IN ANNOTATING DATASETS

To address difficulties in annotating datasets, researchers have exploited scene- or image-level tags as weak supervision tools for localizing objects in images, enabling detection of objects without location annotations [33]. Weakly supervised learning uses training data with incomplete annotations (inexact information, image-level labels) to learn detection models [34]. Since Zhou et al. [35] first proposed a CAM technique for localization of objects with weakly labeled images, it has become a common method for weakly supervised object localization [34] that can be used for approximate object detection.

Li et al. [36] used Grad-CAM [37] and Grad-CAM++ [38] to create a heatmap that could localize infrastructure damage in an image and quantify its severity. They verified the feasibility of weakly supervised learning for assessing disasters resulting from infrastructure damage. Park et al. [39] proposed use of CAM technique to classify images conveying material and human factors into nine classes: an outrigger on temporary equipment, PPE, and working conditions. These studies focused on creating a CAM to classify the import of an image and succeeded in demonstrating the potential of weakly supervised learning for object localization and classification in the construction industry.

However, extending localization information to recognize the image was overlooked. Information on localized objects applied to the spatial relationships of multiple objects is essential for detecting construction hazards in an image [26],

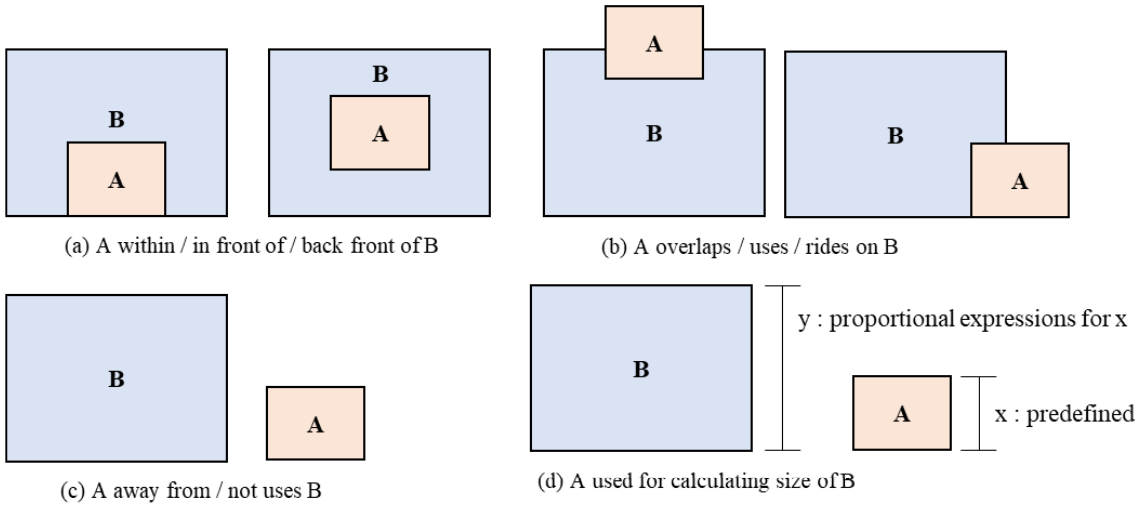

FIGURE 1. Spatial relationships identified in previous studies.

[40]. Moreover, objects related to construction hazards are of many sizes and shapes, including linear and small objects, limiting what the CAM can detect. As a CAM focuses on the most discriminating features of an object, less discriminating features may be ignored or not identified [41], [42]. Various attempts [13], [43] have been made to improve CAM and accurately detect small objects by overcoming noise, but adapting it to construction safety monitoring is challenging due to the characteristics of construction sites where similar and complex objects are present in the background. A few studies have been conducted to mitigate the problem of focusing on incorrect objects in construction sites. These include incorporating additional object size information from building information modeling (BIM) to aid weakly supervised segmentation in indoor environments [6], and integrating CAM techniques to enhance fully supervised learning-based object detection models in indoor environments [44].

C. POINT OF DEPARTURE

Despite the considerable achievement of previous studies on CAM, it still requires additional efforts, such as making BIM models or using other object detection models, to detect/recognize the objects/scene. Relying solely on image-level labels for CAM is challenging due to potential confusion caused by objects with similar shapes and textures, which are difficult to accurately locate and distinguish. As a result, it is not entirely accurate to claim that annotation efforts have been mitigated.

To overcome the challenges associated with annotation efforts and fully harness the potential of CAM, there is a need for the development of a new CAM architecture, that should possess to accurately localize small and linear target objects with complex backgrounds at construction sites, relying solely on image-level labels. Also, the utilization of correlation of localized object to recognize the scene would be instrumental in enhancing construction site safety monitoring.

III. METHODOLOGY

This study proposes the SOS-CAM as a weakly supervised learning-based object localization method using only image-level labels, overcoming the challenges of complex backgrounds in extracting precise location information for small and linear target objects.

This section describes the approach for adapting CAM to recognize images and determining the status of a scene (i.e., safe or unsafe) on construction sites by defining spatial relationships between target objects. Additionally, the architecture of SOS-CAM, and the process of applying spatial relationships for construction safety monitoring is presented. SOS-CAM employs the utilization of multiscale feature maps to effectively capture and discern small and linear objects within the images. This methodology enhances the feature extraction of small and linear objects. Furthermore, a refined module is integrated into the architecture to ameliorate noise-related challenges. The extracted location information from SOS-CAM is applied to the spatial relationships between target objects to determine the status of a scene.

A. APPROACH OF APPLYING WEAKLY SUPERVISED LEARNING

As shown in Fig. 2, the process of applying spatial relationships to determine status consists of three steps: (a) preprocessing: preparing a dataset with image-level labels; (b) object detection: localizing target objects and generating bounding boxes with classification using SOS-CAM, and (c) post-processing: applying established logic to the spatial relationship between target objects, as shown in Fig. 1.

The SOS-CAM is constructed by training on an image-level labeled dataset and generating a CAM for each class. Images for training the model must be captured in

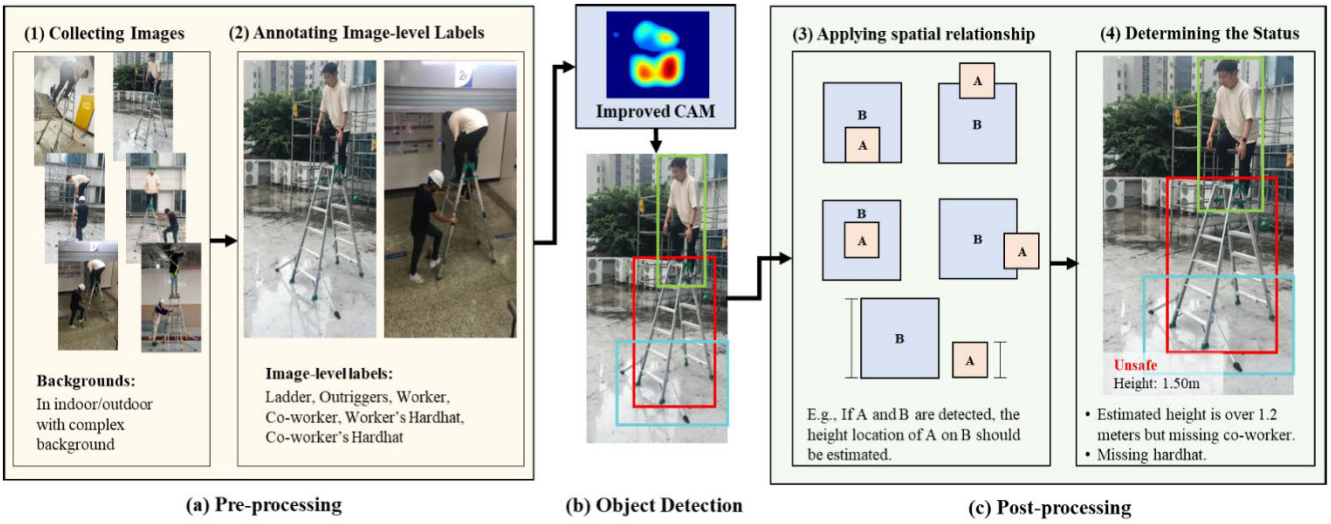

FIGURE 2. Leveraging weakly supervised learning approach for construction safety monitoring.

construction site environments where hazards often occur. Thereafter, only image-level labels that represent the class names of the target objects are annotated on the images to train the network.

The trained network can focus on target objects and generate bounding boxes to extract localization information. Localization information is applied to logic (spatial relationships) to determine scene status. To generate accurate bounding boxes for extracting the localization information of target objects, SOS-CAM concatenates low-level features and applies a refining process, as described in Section III-B.

From the object detection results (after generating the bounding boxes), the logic from the spatial relationships [24], [25], [26], [27] is executed to recognize the image status as safe or unsafe. The spatial relationships can be summarized as follows:

(1) The WITHIN case may cover the following examples:

- Worker (A) is located in front of equipment (B), such as the scissor lift.
- Worker (A) is located at the back of transparent equipment (B), such as scaffolds.
- Worker (A) wears PPE (B), such as a safe vest.
- Worker (A) is located on the deck plate (B).

(2) The OVERLAP case may cover the following examples:

- Worker (A) rides equipment (B), such as a step ladder.
- Worker (A) holds equipment (B), such as a hand saw.
- Worker (A) is close to an opening (B), such as a hole in a slab.
- Worker (A) is equipped with PPE (B), such as a safety harness.

(3) The AWAY case may cover the following examples:

- Worker (A) is located near equipment (B), such as the guardrails on the edge.
- Worker (A) does not use equipment (B), such as scaffolds.
- Worker (A) fastens safety hook (B) to the safety line at a distance.

(4) ESTIMATING how close workers are to hazardous equipment and calculating the height or distance are required.

- The bounding box height for a hole (A) is defined as 0.8 m; the distance to the worker (B) can be estimated via a proportional expression.

As the coordinates of the bounding boxes including the centroid, bottom, and top of the bounding box are used in the spatial relationship, accurate localization of bounding boxes is essential for implementing weakly supervised learning-based scene recognition. Consequently, the performance of determining status is related to the performance of object localization; thus, the performance evaluation should encompass both object localization and status determination.

B. SOS-CAM FOR LOCALIZING AND CLASSIFYING OBJECTS

Weakly supervised learning methods that use image-level labels use a CAM to detect object locations. The trained network produces a CAM as a localization map by aggregating deep feature maps using a class-specific fully connected layer [35]. The major obstacle in using the CAM technique for object detection is capturing the entire object region rather than its most discriminating part [9]. This makes it difficult to use a CAM, as it cannot provide accurate location information required to recognize risks through spatial relationships. Moreover, because only the last convolutional layer is used to obtain the spatial feature with the smallest dimensions for generating the CAM, it may result in coarse visualization, with an over-highlighted background and missing small objects [13].

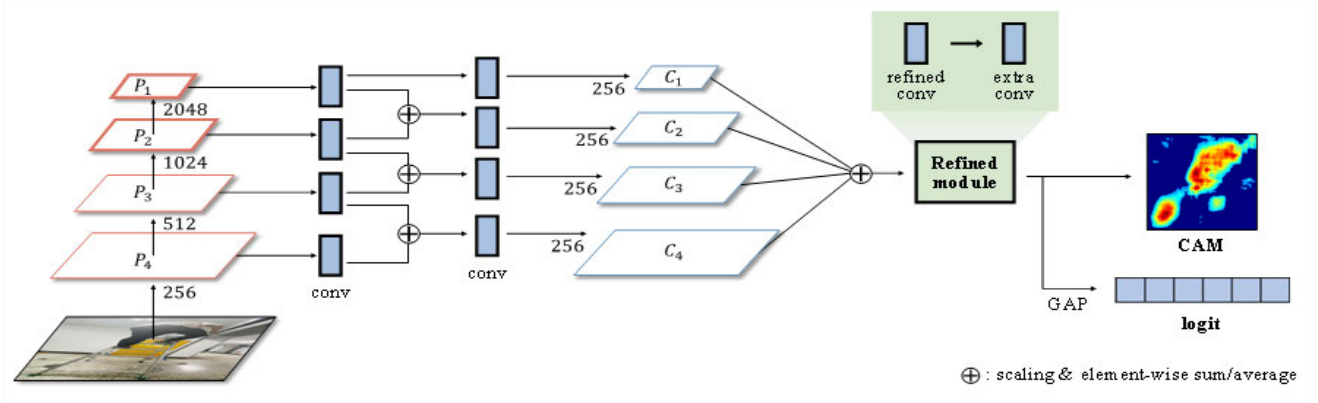

FIGURE 3. SOS-CAM architecture.

1) EXTRACTING MULTISCALE FEATURE MAPS

SOS-CAM is inspired by the Feature Pyramid Network (FPN) [45], which detects multiple objects and provides accurate location information; it is used to apply spatial relationships for recognizing scene status in construction sites while mitigating the limitations of conventional CAM technique. High-level features in the deep layer have more semantic meaning and low resolution; low-level features in the shallow layer are more content-descriptive and high-resolution [46]. Thus, high-level features are more beneficial for detecting textures with the most discriminating features; low-level features contain spatial information (edges and curves), making them beneficial for detecting small and linear objects.

SOS-CAM uses ResNet-50 with an FPN to learn scale-invariant representations using feature maps of different spatial information. The SOS-CAM architecture is illustrated in Fig. 3.

(1) From the convolutional network, a feature map with twice the resolution size difference was extracted from each level. In ResNet-50, four feature maps with twice the resolution size difference were extracted using the output of the last layer of each stage. Accordingly, four feature maps $P_i(i$ in {1,2,3,4}) of different sizes were obtained (i is the index of the feature map; the number decreases as the layer becomes deeper).

(2) The spatial size of the feature map was upsampled at each level to increase the resolution and match it with that of the feature map at the lower level. For example, through upsampling, P_4 matches the resolution of P_3, P_3 matches the resolution of P_2, and P_2 matches that of P_2. For P_1, a feature map of the original size without upsampling was used. The nearest-neighbor upsampling method, which is generally used as a parameter-free operation [47], was used. As P_1 was obtained from the lowest level, it had the same resolution as the original image; thus, it was not upsampled.

(3) All feature maps were passed through a 1×1 convolution layer to reduce the channel size to 256 pixels. To aggregate the information on the feature maps at different levels, an element-wise summation was performed on feature maps of the same resolution. Four aggregated feature maps were input into a 3×3 convolutional layer. Through these processes, feature map C_i of each level includes information on objects of different sizes.

Feature maps of different resolutions were extracted from the network layers and integrated. These feature maps contain information on both low- and high-level features, and are beneficial for capturing small objects in an image.

2) REFINING MULTISCALE FEATURE MAPS

However, using low-level features has an adverse effect in that it increases noise and can interrupt object localization. The integrated feature map contains information from multiple network layers. This means that the integrated low-level features include low-level attributes (e.g., texture and edge) in the image, which are unrelated to the class labels. To effectively remove information that is irrelevant to these classes and accurately capture objects, a refined module is applied to the integrated feature maps.

The refined module consists of a refined convolutional layer for feature refinement, and an extra convolutional layer that is used to reduce the number of channels to align with the desired number of classes for prediction. The refined and extra-convolutional layers employ a 1×1 convolutional filter. The refined module results in an integrated feature map containing rich semantics and aggregates the class-relevant attributes. Thus, the activation in the feature map that is irrelevant to the target labels is suppressed. The CAM can be obtained using this additional refinement process.

Upsampling and downsampling were performed based on the intermediate-resolution feature map to utilize the information on the four feature maps in a balanced manner. For example, if the reference feature map is C_3, the resolution is upsampled four and two times for C_1 and C_2, respectively, whereas C_4 is downsampled by half the number of times through maximum pooling.

After all feature maps had the same resolution, an element-wise summation was performed for all feature maps and

averaged according to the number of levels. However, an aggregated feature includes a large volume of spatial information because it contains a feature map obtained from a lower layer, which may cause a relatively large amount of noise. It is impossible to capture an object's location accurately by generating a noisy activation map.

To alleviate noise, SOS-CAM adds a refined module after element-wise summation, and a more discriminatory feature with semantic information can be extracted. The refined module consists of a convolutional layer that performs refinement and a 1×1 convolutional layer that reduces the number of channels to match the number of classes to be predicted.

Global average pooling (GAP) was performed on the refined feature map to obtain a class logit $\hat{y}$. A localization map was generated based on the obtained activation map. If the activation value of the CAM was equal to or less than the threshold, 0 was allocated. For classification, the multi-label soft margin loss (L) in (1) is obtained from class label y and the predicted logit. p_c is the prediction for the cth class, $\sigma(\cdot)$ denotes the sigmoid function, and C represents the total number of class labels.

$$L = -\frac{1}{c}\sum_{c=1}^{C} y_c log\sigma\,(p_c) + (1 - y_c)\log\left[1 - \sigma\,(p_c)\right] \tag{1}$$

As the conventional CAM [35] utilized only the output of the final layer, it failed to identify small and linear objects in an image by learning a scale-variant feature. The SOS-CAM can effectively capture objects of different scales by extracting and integrating the features of different layers. Thus, small and linear objects such as hardhats and outriggers can be detected in a scene, and an accurate activation map is generated, overcoming the challenges resulting from complex backgrounds.

3) GENERATING BOUNDING BOX

A CAM is an active region generated using a trained classification network and is a coarse prediction of the location of an object based on the active region of the CAM. After resizing the image to several scales {480, 640, 768} to obtain a relatively scale-invariant prediction, it was fed into a trained network. The final localization map results from aggregating the activation maps obtained from the input images of different scales.

To generate bounding boxes from the final localization map and apply spatial relationships, the localization map should be segmented. In this study, the threshold value was set to be above 50% of the maximum value in the localization map. The bounding boxes covering the connected components of the segmented localization map were generated.

IV. EXPERIMENTS

This section describes validation of the SOS-CAM. The dataset preparation and logic for checking rule compliance were described according to the defined scenario. The results of the experiments are presented in three steps: (1) generating the CAMs, (2) creating bounding boxes and determining the scene status, and (3) comparing qualitative and quantitative evaluations from the previous CAM and SOS-CAM.

A. TARGET SCENARIO

Falls from height (FFH) are the most frequent and impactful type of accidents at construction sites [48], [49]. The leading cause of trauma fatalities during construction is FFH, which accounts for 35.8% of all incidents [50]. Despite continuous efforts over the past few decades, FFH remain the main cause of accidents in the construction industry. Halabi et al. [51] reported that approximately 23,000 FFH accidents have occurred at construction sites over the past 20 years; this number is projected to increase significantly.

Of the types of FFH, falls from ladders (FFL) are responsible for a large proportion. Ladders are among the most widely used tools for accessing vertically distant surfaces. According to a statistical survey of construction companies [52], ladders are a leading cause of FFH, second only to roofing. Between 2011 and 2015, over 23% of fatal falls occurred on ladders [50]. In Korea, detailed guidelines for using step ladders [53] have been established and transmitted to workers and managers at construction sites to prevent ladder-oriented FFHs. The essential guidelines are presented as follows:

- Outriggers: A ladder should be installed with outriggers to prevent collapsing.
- Buddy system: If a worker is located at more than 1.2 m (height), coworkers should be able to grab the ladder.
- PPE: Workers should wear a hardhat.

To determine whether the scene shown in an image related to FFL is compliant, detecting small (hardhat) and linear objects (ladder and outriggers), and adopting several spatial relationships from Section III-A (e.g., overlap and within relationships for determining whether a worker is located on a ladder and estimating height for dependent rule compliance) are required. As mentioned in Section II-B, the previous CAM encountered challenges in detecting small objects and handling complex backgrounds at construction sites. Thus, scenarios for rule compliance when using a ladder can facilitate object localization and classification to validate SOS-CAM and are adequate for adapting spatial relationships.

B. DATASET PREPARATION

In this study, 3,615 images with image-level labels were prepared as datasets. The images were extracted from video clips collected at the Chung-Ang University and a construction site in South Korea. As shown in Fig. 4, videos were collected from indoor locations that had features similar to those of finishing works at construction sites, and outdoors covering complex backgrounds, including scaffolding and exterior walls. Furthermore, to prevent overfitting and obtain reasonable results, the authors ensured diversity in distances and viewpoints while collecting videos and checked for similarities in the extracted images. The annotations for each

FIGURE 4. Examples of FFL dataset with image-level labels.

TABLE 1. Dataset distribution.

SCENE STATUS	OVERALL	TRAINING	VALIDATION	TEST
Safe	1,432	996	294	142
Unsafe	2,183	1,530	426	227
Overall	3,615	2,526	720	369

object were only given as class names, as shown in Fig. 4; thus, the SOS-CAM was able to train with image-level labels that had multiple classes on an image.

After annotating the image-level labels, the dataset was randomly divided into training, validation, and test sets in a ratio of 7:2:1, corresponding to 2,526 images for training, 720 images for validation, and 369 images for testing as presented in Table 1. The dataset distribution ratio was followed to previous computer vision studies [54], [55], [56] that, on small datasets, have demonstrated successful results using same ratio.

To evaluate the performance of object localization, bounding box annotations for each object were provided only in the test set images. Additionally, the ground truth of the images for the scenario was divided into two classes (safe and unsafe) to evaluate the prediction of scene status. Maintaining balance in the dataset distribution is important for preventing biased results. However, the types of unsafe scenarios are diverse, such as missing worker's and co-worker's hardhats, a missing co-worker when the worker Is at a height of more than 1.2 m (buddy system), and missing outriggers. In addition, these unsafe situations can occur simultaneously; thus, the proportion of unsafe scenes in the dataset is higher than that of safe scenes. Although the dataset may appear to have an imbalanced distribution, significantly more images related to unsafe situations were expected when considering the diverse types of unsafe scenarios.

The number of classes for objects annotated in the dataset was six excluding background, as listed in Table 2. According to the fundamental rules in Section IV-A, the outrigger, ladder, worker, coworker, worker's hardhat, and coworker's hardhat were defined as object classes that served as the main items for checking rule compliance. In addition, background images that did not contain target objects related to FFL were set as the background class to improve the focus on the essential features of the target object rather than irrelevant background features [57].

As the FFL is the target scenario, a ladder and worker appear in most images. As the other objects (coworker, outrigger, and hardhat) are dependent objects that can appear when a worker and ladder are present, the number of instances of co-workers and ladders has an imbalanced ratio.

C. DEPLOYING SPATIAL RELATIONSHIPS FOR FFL-RELATED RULE COMPLIANCE

After extracting the bounding boxes and their coordinates, logic for adapting the spatial relationships between the target objects is required to extract the meaning of the scene

in the image (safe or unsafe). According to the KOSHA guidelines, spatial relationships (WITHIN, OVERLAP, AWAY FROM, and ESTIMATION) can be used to check hardhat use, ensure that workers, co-workers, and outriggers are present, and to estimate height. However, this study annotated these rules as objects and aimed to detect the spatial features of objects, including small ones, without manual effort. The height at which a worker is located is crucial information because a co-worker is required if the height exceeds 1.2 m, according to the guidelines. Moreover,

the co-worker hardhat is another dependent variable that becomes relevant only when a co-worker appears in the image. Regarding the importance of a specific factor (height location), height estimation was used as a spatial relationship to determine the image status.

To estimate the height at which a worker was located, the bounding box coordinates for the worker, ladder, and outrigger were applied to determine the spatial relationships between OVERLAP and ESTIMATION. The OVERLAP relationship was checked before estimating the height location. The coordinates of the bounding box were used to verify that the worker and ladder objects overlapped before performing height location ESTIMATION.

The angular points of the bounding boxes of the worker $(X_1^w, Y_1^w, X_2^w, Y_2^w)$ and ladder $(X_1^l, Y_1^l, X_2^l, Y_2^l)$ were used to calculate the width overlap ($W^{overlap}$ in (2)) and height overlap ($H^{overlap}$ in (3)) with respect to their minimum and maximum values. The overlap ratio in (6) was calculated by considering the area of the ladder in (5) and the overlap in (4).

$$W^{overlap} = \min\left(X_2^w, X_2^l\right) - \max\left(X_1^w, X_1^l\right) \quad (2)$$

$$H^{overlap} = \min\left(Y_2^w, Y_2^l\right) - \max\left(Y_1^w, Y_1^l\right) \quad (3)$$

$$overlap\_area = W^{overlap} * H^{overlap} \quad (4)$$

$$ladder\_area = W^l * H^l \quad (5)$$

$$ratio = overlap\_area/ladder\_area * 100 \quad (6)$$

If the conditions of ratio > 0 and $Y_2^w < Y_2^l$ are satisfied, the height is estimated by comparing the number of pixels between the lowest points of the worker bounding box Y_2^w and the ladder bounding box Y_2^l and considering the number of pixels on the y-axis of the ladder bounding box.

$$ladder'sheight : H^L = working\ height : d\left(Y_2^l, Y_2^w\right)$$

Then,

$$working\ height = \left(ladder'sheight * d\left(Y_2^l, Y_2^w\right)\right)/H^L \quad (7)$$

Algorithm 1 presents the pseudocode of Eqs. 2–7, which are proposed to adopt the spatial relationships, OVERLAP and ESTIMATION. To avoid unnecessary height calculations when the worker is not on the ladder, OVERLAP is checked by considering the bounding-box area when related objects (the worker and ladder) are detected. For height estimation, the ladder height as a hyperparameter was set to the height of 1.8 meters used in the dataset. Additionally, if the ladder is not detected by the CAM, the outrigger is utilized as an indicator for height estimation in the same manner.

D. TRAINING AND RESULTS

1) TRAINING SETUP

For training and validation, a pretrained ResNet-50 model with ImageNet was used as the classification network. The input image was resized to a resolution of 768 × 768, and random horizontal flipping, color jittering, and random cropping to 640 × 640 augmentations were implemented. The batch size, initial learning rate, weight decay, and momentum were set as 5, 0.001, 0.005, and 0.9, respectively. Stochastic gradient descent was used as the optimizer, and the network was trained for 20 epochs. Because batch normalization was applied, the dropout rate was set to zero due to disharmony [58]. Training was executed on a server equipped with an i9 × 10940X CPU and a single NVIDIA RTX 3090.

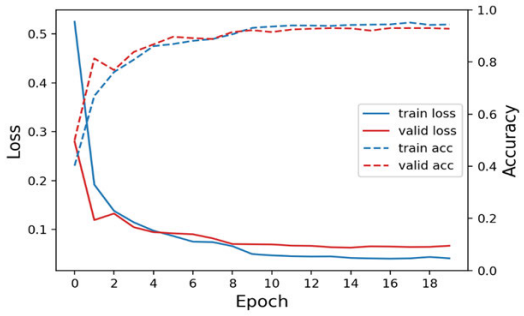

FIGURE 5. Loss and accuracy of training and validation.

2) VALIDATION AND RESULTS

The performance of the proposed SOS-CAM was validated using object localization and multi-label classification (object classification) for status determination. Spatial relationships were adopted to determine the status of the scene and to validate the feasibility of the SOS-CAM for construction safety. Object localization, classification, and status prediction were performed based on quantitative and qualitative evaluations.

The quantitative evaluation used the mean intersection over union (mIoU), also known as the mean Jaccard index, which is commonly used to evaluate the localization accuracy and F1-Score for classifying object and scene status. The scene status was divided into two categories: safe and unsafe. All quantitative evaluation results were the averages of five random runs. The training ended at epoch 20 due to the minimum validation loss, as shown in Fig. 5; the dotted line represents accuracy and the solid line depicts loss.

Table 3 presents the mIoU measurement results for object localization associated with the six classes related to FFL. SOS-CAM exhibited an improvement of 0.0478 mIoU compared to the conventional CAM [35], indicating an overall enhancement in the localization performance of the target objects. In particular, SOS-CAM, which incorporates an FPN, demonstrated superior localization capabilities for linear objects (e.g., ladders and outriggers) and small objects (e.g., hardhat), overcoming the challenges posed by complex backgrounds.

As presented in Table 4, although SOS-CAM showed a slight decrease in accuracy for object classification, the F1-score increased by 0.0027, and is considered a more crucial factor [59], especially with imbalanced datasets. Furthermore, for FFL, because the status of the scene is dependent on the worker location (working height) and the proposed technique applies bounding box information from SOS-CAM

TABLE 2. Instances in image.

CATEGORY	BACKGROUND	WORKER	WORKER'S HARDHAT	CO-WORKER	CO-WORKER'S HARDHAT	LADDER	OUTRIGGER
INSTANCES	384	2,604	1,743	874	693	3,066	1,751

Algorithm 1 Pseudocode for overlap and estimation

Input: Bounding box of worker, ladder, and outrigger, extracted top-left bounding box coordinates

Output: Working height of worker relative to ladder and outrigger, overlapping ratio of worker and ladder to ladder

worker_bbox = $(X_1^w, Y_1^w, X_2^w, Y_2^w)$ \\ bounding box of worker

ladder_bbox = $(X_1^l, Y_1^l, X_2^l, Y_2^l)$ \\ bounding box of ladder

outrigger_bbox = $(X_1^o, Y_1^o, X_2^o, Y_2^o)$ \\ bounding box of outrigger

$W^w = X_2^w - X_1^w; H^w = Y_2^w - Y_1^w$ \\ width and height of worker

$W^l = X_2^l - X_1^l; H^l = Y_2^l - Y_1^l$ \\ width and height of ladder

LADDER_HEIGHT =1.8

if ladder_bbox is exist:

ladder_area = $W^l * H^l$

overlap_area = calculate_overlap_area(worker_bbox, ladder_bbox) // find overlapping area (worker & ladder)

  overlap_ratio = overlap_area / ladder_area * 100

  if overlap_ratio > 0 and $Y_2^w < Y_2^l$: \\ lower end of the worker is higher than the lower end of the ladder

  // ladder_height: H^l = working_height: working_height_pixel

  working_height_pixel = ladder_lower - Y_2^w

  working_height = (working_height_pixel * LADDER_HEIGHT) / H^l

  elif outrigger_bbox is exist:

  if $Y_2^w < Y_2^o$: \\ lower end of worker is higher than the lower end of outrigger

  working_height_length = outrigger_lower - Y_2^w

  working_height_outrigger = (working_height_length * WORKER_HEIHGT) / H^w

  return working_height, working_height_outrigger, overlap_ratiod

to recognize an image, determining the status of the scene can be considered as an indicator of localization performance and feasibility of using SOS-CAM for construction safety monitoring. The experimental results demonstrated that SOS-CAM achieved a 0.0092 increase in accuracy and a 0.0079 increase in F1-Score, as presented in Table 4. The increase of 0.0448 in the precision metric, indicating that the model correctly determined the status, coincided with a small decrease in the recall metric, reflecting a trade-off relationship between the two measures.

In Fig. 6, the points of focus of the CAMs for comparison with the previous CAM [35] and SOS-CAM were extracted from the test images. The results of creating the bounding boxes are compared in Fig. 6. Although both CAM techniques succeeded in classifying the object in the images, the sizes of the bounding boxes, which have an important influence on the height estimation, differed. The sky blue, blue, light blue, red, light green, and green bounding boxes represent ladders, co-workers, co-worker hardhat, worker hardhat, workers, and outriggers, respectively. As shown in Fig. 6(a), the SOS-CAM created a precise bounding box for the worker and outrigger, whereas the previous CAM created a larger bounding box with the wrong focus. Moreover, the previous CAM tended to misfocus the ladder, as shown in Fig. 6(b), (c), and (d), whereas the SOS-CAM succeeded in maintaining focus with complex backgrounds. Given that the backgrounds of the images used in this study were scaffolds or finished with a grid pattern that resembled a ladder, a linear object, the SOS-CAM can presumably detect target objects better than the previous CAM. Furthermore, the SOS-CAM can detect small objects such as the outrigger and hardhat more accurately, as shown in Fig. 6(c) and (d).

The image recognition and status determination results are shown in Fig. 7. After extracting the coordinates from the generated bounding boxes, a spatial relationship was adopted to determine the status of the scene. As shown in Fig. 7(a), the SOS-CAM method exhibited enhanced object localization capabilities, resulting in more accurate estimation of the

TABLE 3. Performance for localizing objects.

METHOD	IoU_{WORKER}	$IoU_{WORKER'S\ HARDHAT}$	$IoU_{CO\text{-}WORKER}$	$IoU_{CO\text{-}WORKER'S\ HARDHAT}$	IoU_{LADDER}	$IoU_{OUTRIGGER}$	mIoU
CAM [35]	0.4047	0.0284	0.3088	0.0630	0.3606	0.3167	0.2470
SOS-CAM	0.4584	0.0562	0.3214	0.1012	0.3986	0.4332	0.2948

TABLE 4. Performance in classifying objects and determining status.

METHOD	CLASSIFICATION OF OBJECTS				DETERMINING STATUS			
	ACC	PRECISION	RECALL	F1 SCORE	ACC	PRECISION	RECALL	F1-SCORE
CAM [35]	0.9375	0.9744	0.9770	0.9719	0.9415	0.8984	0.9617	0.9270
SOS-CAM	0.9345	0.9773	0.9808	0.9746	0.9507	0.9432	0.9291	0.9349

worker's position and correct status determination. Although both CAM techniques successfully assessed the status in Fig. 7(b) and (d), SOS-CAM effectively addressed the challenges posed by complex backgrounds such as scaffolds and grid-patterned finishes, enabling more precise height estimation. Additionally, the SOS-CAM demonstrated superior performance in localizing small and linear objects (e.g., hardhat and outriggers), as shown in Fig. 7(c) and (e), compared with the conventional CAM technique. The SOS-CAM achieved a close estimation of the ground truth for height estimation owing to its improved ladder detection capability. However, in Fig. 7(c) and (e), both CAM techniques inaccurately localize the ladder. Nevertheless, the SOS-CAM yielded a height estimation that was closer to the ground truth than the conventional CAM technique.

Although the SOS-CAM exhibited improved object localization performance, resulting in a more accurate status determination, challenges in precise object detection persisted. The bounding box was not perfectly extracted to fit the ladder, including the worker, leading to inaccurate detection, even though the size and location of the bounding box for the ladder in the SOS-CAM were more accurate than those in the previous CAM. The results were attributed to the dataset used in this study, which was constructed to determine the status of the FFL. Thus, the presence of ladders was nearly ubiquitous across most images within the dataset, as shown in Table 2.

These results suggest that the proposed SOS-CAM exhibits better performance in detecting target objects related to FFL at construction sites. SOS-CAM has considerable potential for localizing objects at construction sites because it improves the localization of linear and small-sized objects in complex backgrounds.

V. DISCUSSION AND LIMITATIONS

The weakly supervised learning method using image-level labels effectively reduces annotation costs by approximating the locations of objects based on the CAM. This study proposed weakly supervised learning approach using SOS-CAM and spatial relationships for construction safety monitoring. SOS-CAM, in the proposed approach, utilizes both low- and high-level features to generate CAM and identifies small-size and linear objects without highlighting complex background regions. SOS-CAM uses an FPN to generate a CAM by concatenating both low- and high-level features. Concatenating low-level features can have an adverse effect, causing noise in the activation map and leading to incorrect localization and sizing of bounding boxes. To alleviate noise, a refined module was adopted after concatenating the low-level features before GAP. Thus, a bounding box was generated from a refined feature and leveraged for predefined spatial relationships. A case study using 3,615 image-level labeled images related to the FFL was conducted to validate the SOS-CAM for qualitative and quantitative analyses. The validation results revealed that creating the CAM and bounding boxes using the proposed methodology yielded more accurate localization information. The experimental results revealed an accuracy of 0.2948 mIoU for object localization, 0.9746 F1-Score, 0.9345 accuracy for object classification, and 0.9349 F1-Score and 0.9507 accuracy in determining whether the scene was safe or unsafe. The SOS-CAM achieved a classification accuracy comparable to that of the conventional method, but exhibited superior localization performance, confirmed by the status accuracy. Thus, SOS-CAM for object localization and applying spatial relationships to determine the scene status has sufficient potential for use in detection of construction hazards.

This study highlights achievements in the following areas.

1) This study made initial attempts to use a CAM with only image-level labels to assist in checking rule compliance using spatial relationships. Although meaningful research on the CAM has been undertaken in the construction industry, using only image-level labels for object localization and aiding in compliance determination is highly uncommon. This study succeeded in using only the CAM technique, without any other tools or techniques to check for rule compliance.

2) The performance of the SOS-CAM facilitated smoother integration of spatial relationships into scene recognition tasks, reinforcing the potential practicality of CAM. SOS-CAM uses an FPN to overcome complex backgrounds for

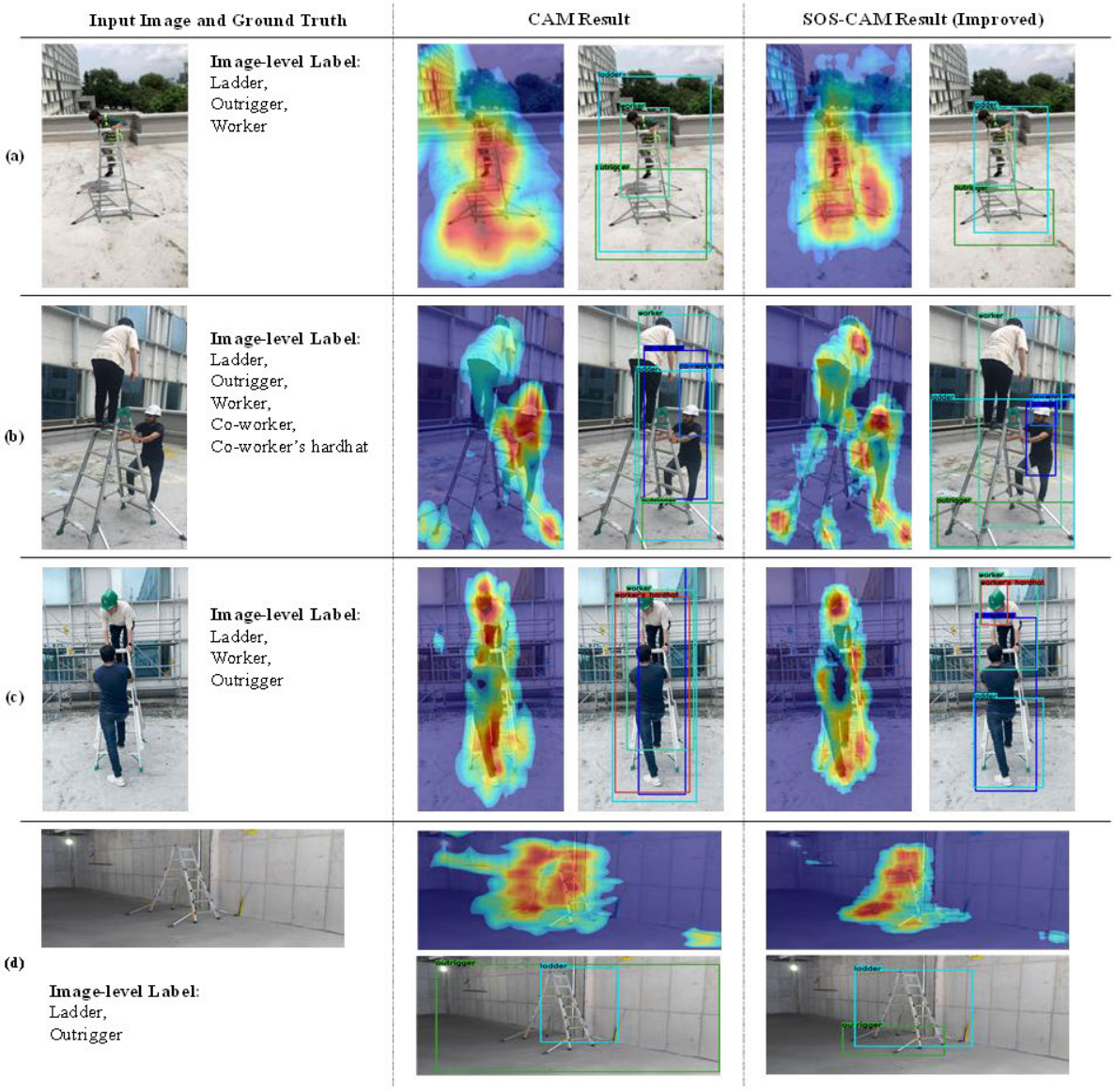

FIGURE 6. Object localization results of conventional CAM and SOS-CAM.

small and linear object localization in construction sites; a refined module is used to reduce noise resulting from use of low-level features, improving the accuracy of object location information and reinforcing the effectiveness of spatial relationship applications in scene recognition.

3) Application of the CAM to localize and classify small and linear object at construction sites, has great potential. This study indicates the feasibility of using image-level labels for weakly supervised learning in construction safety monitoring and lays the foundation for expanding computer vision research beyond fully supervised learning to include weakly supervised learning methodologies.

4) The dataset can be accumulated through image-level labels; images on the web can be used more effectively without creating bounding boxes or polygons. Use of a large number of images can generalize computer vision models for construction sites.

The objective of this study is to pioneer an innovative approach using CAM as a weakly supervised learning dedicated to construction safety monitoring. To enhance the robustness and practicality of the proposed methodology, conducting a meticulous comparison with contemporary advanced CAM techniques is necessary in future research. Although the SOS-CAM has improved object localization performance and allows for accurate determination of status, precise object detection still poses challenges. The SOS-CAM successfully extracted the bounding box and determined the status more accurately than the previous CAM; however, the bounding box was fitted to the ladder and included the worker. There were also incorrect detections of

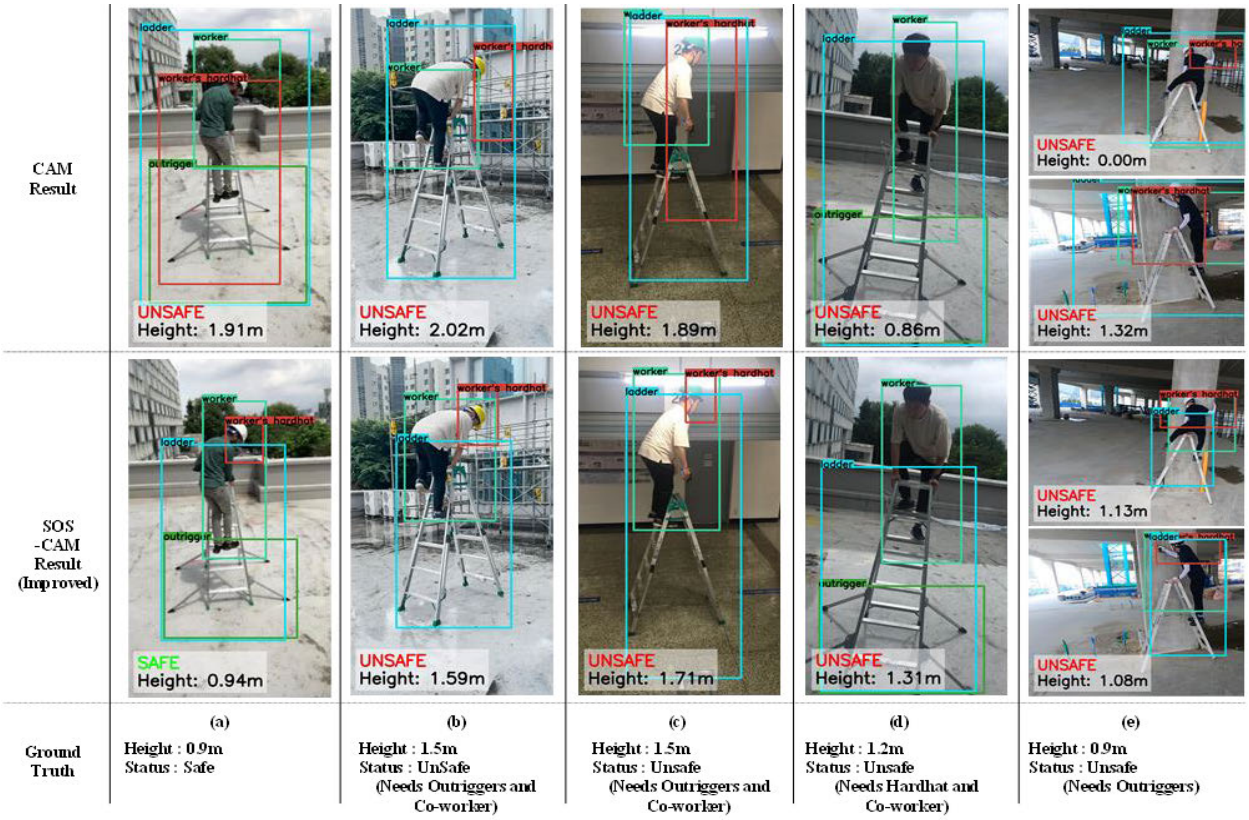

FIGURE 7. Prediction results of conventional CAM and SOS-CAM.

hardhats, possibly attributed to similarities in body parts or clothing colors. The dataset used in this study was limited; incorrect localization may have been resolved by training with larger amounts of data. It is anticipated that the use of large-scale benchmark datasets such as MOCS [7] and SODA [3] will improve the performance of the proposed CAM, specifically by enhancing its ability to localize objects. Furthermore, the dataset used in this study was limited to the already installed step ladder, which is opened and transparent; thus, the trained SOS-CAM as IP(Image Processing) is able to recognize the scene related to only the trained type of ladder. Given the constrained nature of the experiment, additional research including more scenarios such as verifying secure attachment of safety hooks to scaffolds where spatial relationships can be applied, is necessary to fully demonstrate the robustness of the SOS-CAM, ensuring that the proposed methodology can be effectively used across a broad spectrum of construction safety management contexts.

In addition, to further enhance the performance and applicability of the SOS-CAM, the feature aggregation method should be improved. The SOS-CAM integrates the low- and high-level features of a network in a one-way manner, which limits the flow of information from different layers. Moreover, feature maps of all levels are aggregated with the same weight; the importance of each level of the feature map is ignored. In future research, integration of feature maps at multiple levels in different directions by considering the weight of each resolution can overcome the limitations of this study.

VI. CONCLUSION

Computer vision has significant potential in automatic scene recognition to enhance safety measures and prevent accidents at construction sites. Computer-vision applications for construction safety still require accumulation of datasets, despite the emergence of weakly supervised learning methods. Due to the limitations of conventional CAM in overcoming complex backgrounds and effectively capturing small and linear objects, research on utilizing CAM to recognize unsafe scenarios such as FFL at construction sites by leveraging the spatial relationships between detected objects has been limited. In this study, SOS-CAM, as a weakly supervised learning method, was designed to localize small and linear objects while overcoming complex backgrounds at construction sites; extracting multiscale feature maps to use low- and high-level features for capturing small and linear objects in an image, and the refined module addresses the problem of noise. Spatial relationships were used to determine the scene status to validate the feasibility of the SOS-CAM for construction safety monitoring. To train, validate, and test the SOS-CAM, a dataset comprising 3,615 images associated with FFL was prepared using only image-level labels.

The SOS-CAM demonstrated precise localization of target objects, yielding improvements in performance metrics, including a 0.0478 increase in mIoU, a 0.0092 increase in scene status determination accuracy, and a 0.0079 increase in the F1-score. The proposed SOS-CAM enables more precise localization of objects, making it a potential computer vision monitoring technique in challenging environments such as construction sites, where dataset annotation remains a challenge.

To overcome the limitations of this study, the authors plan to improve SOS-CAM in terms of object detection performance at construction sites by comparing existing models. It can mitigate the limitations of CAM technique, which focuses on local features, and enhance the weakly supervised learning method for monitoring construction sites. Moreover, through the CAM of images used in this study, currently commercialized detectors can be improved, and monitoring systems can be facilitated as it allows the object detection reasoning of the detector to be understood. In addition, the proposed method can evolve to assist or replace fully supervised learning-based detectors. The code for SOS-CAM is available at https://github.com/EJLEE5826/SOS-CAM.

ACKNOWLEDGMENT

(Jaehun Yang and Eunju Lee are co-first authors.)

REFERENCES

[1] W. Fang, L. Ding, P. E. D. Love, H. Luo, H. Li, F. Peña-Mora, B. Zhong, and C. Zhou, "Computer vision applications in construction safety assurance," *Autom. Construct.*, vol. 110, Feb. 2020, Art. no. 103013, doi: 10.1016/j.autcon.2019.103013.

[2] J. Seo, S. Han, S. Lee, and H. Kim, "Computer vision techniques for construction safety and health monitoring," *Adv. Eng. Informat.*, vol. 29, no. 2, pp. 239–251, Apr. 2015, doi: 10.1016/j.aei.2015.02.001.

[3] R. Duan, H. Deng, M. Tian, Y. Deng, and J. Lin, "SODA: A large-scale open site object detection dataset for deep learning in construction," *Autom. Construct.*, vol. 142, Oct. 2022, Art. no. 104499, doi: 10.1016/j.autcon.2022.104499.

[4] S. Paneru and I. Jeelani, "Computer vision applications in construction: Current state, opportunities & challenges," *Autom. Construct.*, vol. 132, Dec. 2021, Art. no. 103940, doi: 10.1016/j.autcon.2021.103940.

[5] C. Ge, J. Wang, J. Wang, Q. Qi, H. Sun, and J. Liao, "Towards automatic visual inspection: A weakly supervised learning method for industrial applicable object detection," *Comput. Ind.*, vol. 121, Oct. 2020, Art. no. 103232, doi: 10.1016/j.compind.2020.103232.

[6] L. Yang and H. Cai, "Cost-efficient image semantic segmentation for indoor scene understanding using weakly supervised learning and BIM," *J. Comput. Civil Eng.*, vol. 37, no. 2, pp. 1–15, Mar. 2023, doi: 10.1061/jccee5.cpeng-5065.

[7] A. Xuehui, Z. Li, L. Zuguang, W. Chengzhi, L. Pengfei, and L. Zhiwei, "Dataset and benchmark for detecting moving objects in construction sites," *Autom. Construct.*, vol. 122, Feb. 2021, Art. no. 103482, doi: 10.1016/j.autcon.2020.103482.

[8] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, Jan. 2018, doi: 10.1093/nsr/nwx106.

[9] W. Bae, J. Noh, and G. Kim, "Rethinking class activation mapping for weakly supervised object localization," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Springer Science and Bus. Media), Deutschland, Germany, vol. 12360, 2020, pp. 618–634.

[10] H. Chen, Q. Hu, B. Zhai, H. Chen, and K. Liu, "A robust weakly supervised learning of deep conv-nets for surface defect inspection," *Neural Comput. Appl.*, vol. 32, no. 15, pp. 11229–11244, Aug. 2020, doi: 10.1007/s00521-020-04819-5.

[11] K. Kim, H. Kim, and H. Kim, "Image-based construction hazard avoidance system using augmented reality in wearable device," *Autom. Construct.*, vol. 83, pp. 390–403, Nov. 2017, doi: 10.1016/j.autcon.2017.06.014.

[12] N. Khan, M. R. Saleem, D. Lee, M.-W. Park, and C. Park, "Utilizing safety rule correlation for mobile scaffolds monitoring leveraging deep convolution neural networks," *Comput. Ind.*, vol. 129, Aug. 2021, Art. no. 103448, doi: 10.1016/j.compind.2021.103448.

[13] X. Shi, S. Khademi, Y. Li, and J. van Gemert, "Zoom-CAM: Generating fine-grained pixel annotations from image labels," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 10289–10296.

[14] C. Melchior and R. R. Zanini, "Mortality per work accident: A literature mapping," *Saf. Sci.*, vol. 114, pp. 72–78, Apr. 2019, doi: 10.1016/j.ssci.2019.01.001.

[15] H. Y. Chong and T. S. Low, "Accidents in Malaysian construction industry: Statistical data and court cases," *Int. J. Occupational Saf. Ergonom.*, vol. 20, no. 3, pp. 503–513, Jan. 2014, doi: 10.1080/10803548.2014.11077064.

[16] *Commonly Used Statistics | Occupational Safety and Health Administration*. Accessed: Sep. 5, 2022. [Online]. Available: https://www.osha.gov/data/commonstats

[17] D. Gyo, Y. Jun, S. S. Wook, and D. Young, "Analyzing the relationship between the critical safety management tasks and their effects for preventing construction accidents using IPA method," *Korean J. Constr. Eng. Manag.*, vol. 23, no. 5, pp. 77–86, 2022, doi: 10.6106/KJCEM.2022.23.5.077.

[18] K. Shrestha, P. P. Shrestha, D. Bajracharya, and E. A. Yfantis, "Hard-hat detection for construction safety visualization," *J. Construction Eng.*, vol. 2015, pp. 1–8, Feb. 2015, doi: 10.1155/2015/721380.

[19] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proc. IEEE*, vol. 111, no. 3, pp. 257–276, Mar. 2023.

[20] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11218. Cham, Switzerland: Springer, 2018, pp. 816–832, doi: 10.1007/978-3-030-01264-9_48.

[21] B. E. Mneymneh, M. Abbas, and H. Khoury, "Vision-based framework for intelligent monitoring of hardhat wearing on construction sites," *J. Comput. Civil Eng.*, vol. 33, no. 2, pp. 1–20, Mar. 2019, doi: 10.1061/(asce)cp.1943-5487.0000813.

[22] W. Fang, L. Ding, H. Luo, and P. E. D. Love, "Falls from heights: A computer vision-based approach for safety harness detection," *Autom. Construct.*, vol. 91, pp. 53–61, Jul. 2018, doi: 10.1016/j.autcon.2018.02.018.

[23] Z. Xie, H. Liu, Z. Li, and Y. He, "A convolutional neural network based approach towards real-time hard hat detection," in *Proc. IEEE Int. Conf. Prog. Informat. Comput. (PIC)*, Dec. 2018, pp. 430–434, doi: 10.1109/PIC.2018.8706269.

[24] Q. Fang, H. Li, X. Luo, L. Ding, H. Luo, and C. Li, "Computer vision aided inspection on falling prevention measures for steeplejacks in an aerial environment," *Autom. Construct.*, vol. 93, pp. 148–164, Sep. 2018, doi: 10.1016/j.autcon.2018.05.022.

[25] W.-C. Chern, J. Hyeon, T. V. Nguyen, V. K. Asari, and H. Kim, "Context-aware safety assessment system for far-field monitoring," *Autom. Construct.*, vol. 149, May 2023, Art. no. 104779, doi: 10.1016/j.autcon.2023.104779.

[26] W. Fang, L. Ma, P. E. D. Love, H. Luo, L. Ding, and A. Zhou, "Knowledge graph for identifying hazards on construction sites: Integrating computer vision with ontology," *Autom. Construct.*, vol. 119, Nov. 2020, Art. no. 103310, doi: 10.1016/j.autcon.2020.103310.

[27] S. Anjum, N. Khan, R. Khalid, M. Khan, D. Lee, and C. Park, "Fall prevention from ladders utilizing a deep learning-based height assessment method," *IEEE Access*, vol. 10, pp. 36725–36742, 2022, doi: 10.1109/ACCESS.2022.3164676.

[28] J. Wu, N. Cai, W. Chen, H. Wang, and G. Wang, "Automatic detection of hardhats Worn by construction personnel: A deep learning approach and benchmark dataset," *Autom. Construct.*, vol. 106, Oct. 2019, Art. no. 102894, doi: 10.1016/j.autcon.2019.102894.

[29] Y.-F. Li, L.-Z. Guo, and Z.-H. Zhou, "Towards safe weakly supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 334–346, Jan. 2021, doi: 10.1109/TPAMI.2019.2922396.

[30] J. Fan, Z. Zhang, C. Song, and T. Tan, "Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4282–4291. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Fan_Learning_Integral_Objects_With_Intra-Class_Discriminator_for_Weakly-Supervised_Semantic_Segmentation_CVPR_2020_paper.html

[31] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang, "Bounding box regression with uncertainty for accurate object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2883–2892.

[32] S. Guo, W. Huang, H. Zhang, C. Zhuang, D. Dong, M. R. Scott, and D. Huang, "CurriculumNet: Weakly supervised learning from large-scale web images," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11214, 2018, pp. 139–154. [Online]. Available: https://openaccess.thecvf.com/content_ECCV_2018/html/Sheng_Guo_CurriculumNet_Learning_from_ECCV_2018_paper.html

[33] Y. Li, W. Chen, Y. Zhang, C. Tao, R. Xiao, and Y. Tan, "Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning," *Remote Sens. Environ.*, vol. 250, Dec. 2020, Art. no. 112045, doi: 10.1016/j.rse.2020.112045.

[34] W. Gao, F. Wan, X. Pan, Z. Peng, Q. Tian, Z. Han, B. Zhou, and Q. Ye, "TS-CAM: Token semantic coupled attention map for weakly supervised object localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2866–2875. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Gao_TS-CAM_Token_Semantic_Coupled_Attention_Map_for_Weakly_Supervised_Object_ICCV_2021_paper.html

[35] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/Zhou_Learning_Deep_Features_CVPR_2016_paper.html

[36] X. Li, D. Caragea, H. Zhang, and M. Imran, "Localizing and quantifying infrastructure damage using class activation mapping approaches," *Social Netw. Anal. Mining*, vol. 9, no. 1, pp. 1–15, Dec. 2019, doi: 10.1007/s13278-019-0588-4.

[37] 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 *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html1

[38] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847, doi: 10.1109/WACV.2018.00097.

[39] J. Park, H. Lee, and H. Y. Kim, "Risk factor recognition for automatic safety management in construction sites using fast deep convolutional neural networks," *Appl. Sci.*, vol. 12, no. 2, p. 694, Jan. 2022, doi: 10.3390/app12020694.

[40] H. Wu, B. Zhong, H. Li, P. Love, X. Pan, and N. Zhao, "Combining computer vision with semantic reasoning for on-site safety management in construction," *J. Building Eng.*, vol. 42, Oct. 2021, Art. no. 103036, doi: 10.1016/j.jobe.2021.103036.

[41] S. Jo and I. J. Yu, "Puzzle-cam: Improved localization via matching partial and full features," in *Proc. Int. Conf. Image Process. ICIP*, Sep. 2021, pp. 639–643, doi: 10.1109/ICIP42928.2021.9506058.

[42] S. Jang, J. Kwon, K. Jin, and Y. Kim, "Weakly supervised semantic segmentation via graph RecalibratiOn with scaling weight uNit," *Eng. Appl. Artif. Intell.*, vol. 119, Mar. 2023, Art. no. 105706, doi: 10.1016/j.engappai.2022.105706.

[43] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "LayerCAM: Exploring hierarchical class activation maps for localization," *IEEE Trans. Image Process.*, vol. 30, pp. 5875–5888, 2021, doi: 10.1109/TIP.2021.3089943.

[44] J. Hwang, K. Lee, M. M. Ei Zan, M. Jang, and D. H. Shin, "Improved discriminative object localization algorithm for safety management of indoor construction," *Sensors*, vol. 23, no. 8, p. 3870, Apr. 2023, doi: 10.3390/s23083870.

[45] Y. Zhao, R. Han, and Y. Rao, "A new feature pyramid network for object detection," in *Proc. Int. Conf. Virtual Reality Intell. Syst. (ICVRIS)*, Sep. 2019, pp. 428–431, doi: 10.1109/ICVRIS.2019.00110.

[46] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 821–830, doi: 10.1109/CVPR.2019.00091.

[47] D. Mazzini, "Guided upsampling network for real-time semantic segmentation," 2018, *arXiv:1807.07466*.

[48] Health and Safety Executive. (2020). *Workplace Fatal Injuries in Great Britain 2020*. [Online]. Available: https://www.hse.gov.uk/statistics/pdf/fatalinjuries.pdf

[49] KOSHA. (2020). *Analysis of Industrial Accidents in 2020*. Accessed: Mar. 31, 2023. [Online]. Available: https://www.kosha.or.kr/kosha/data/industrialAccidentStatus.do?mode=view&articleNo=436868&article.offset=0&articleLimit=10

[50] *The Construction Chart Book*, CPWR, Silver Spring, MD, USA, 2016.

[51] Y. Halabi, H. Xu, D. Long, Y. Chen, Z. Yu, F. Alhaek, and W. Alhaddad, "Causal factors and risk assessment of fall accidents in the U.S. construction industry: A comprehensive data analysis (2000–2020)," *Saf. Sci.*, vol. 146, Feb. 2022, Art. no. 105537, doi: 10.1016/j.ssci.2021.105537.

[52] R. M. Choudhry. (May 2015). *Investigation of Fall Protection Practices in the Construction Industry of Pakistan Hydrology for Environment, Life and Policy (HELP) View Project Construction Engineering and Management At NUST Pakistan View Project*. [Online]. Available: https://www.researchgate.net/publication/275037120

[53] Korea Occupational Safety & Health Agency. (2021). *Guidelines for Using Ladder*. Accessed: Feb. 12, 2022. [Online]. Available: https://www.kosha.or.kr/aicuration/tr/FileDownload.jsp?med_seq=43740&file_name=20211207091635436.pdf&file_path=2021/12&down_name=[2021-BusinessGeneralHeadquarters-779]_MobileDinnerBridgeSafetyWorkGuidelines_OPS.pdf

[54] Z. R. Himami, A. Bustamam, and P. Anki, "Deep learning in image classification using dense networks and residual networks for pathologic myopia detection," in *Proc. Int. Conf. Artif. Intell. Big Data Analytics*, Oct. 2021, pp. 1–6, doi: 10.1109/ICAIBDA53487.2021.9689744.

[55] D. Luthfy, C. Setianingsih, M. W. Paryasto, and N. Amelia, "Utilizing YOLO for efficient Indonesian sign language recognition," in *Proc. Int. Conf. Comput., Control, Informat. Appl. (IC3INA)*, Oct. 2023, pp. 430–435, doi: 10.1109/ic3ina60834.2023.10285793.

[56] M. F. Humayun, F. A. Bhatti, and K. Khurshid, "IVision MRSSD: A comprehensive multi-resolution SAR ship detection dataset for state of the art satellite based maritime surveillance applications," *Data Brief*, vol. 50, Oct. 2023, Art. no. 109505, doi: 10.1016/j.dib.2023.109505.

[57] K. Li, Z. Wu, K. C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 9215–9223, doi: 10.1109/CVPR.2018.00960.

[58] X. Li, S. Chen, X. Hu, and J. Yang, "Understanding the disharmony between dropout and batch normalization by variance shift," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2677–2685, doi: 10.1109/CVPR.2019.00279.

[59] Q. Zou, S. Xie, Z. Lin, M. Wu, and Y. Ju, "Finding the best classification threshold in imbalanced classification," *Big Data Res.*, vol. 5, pp. 2–8, Sep. 2016, doi: 10.1016/j.bdr.2015.12.001.

JAEHUN YANG received the B.S. and M.S. degrees in architectural engineering from the School of Architecture and Building Science, Chung-Ang University, Seoul, South Korea, where he is currently pursuing the Ph.D. degree. His research interests include computer vision and blockchain-based safety management systems and applications.

EUNJU LEE received the B.S. and M.S. degrees in imaging engineering from Chung-Ang University, Seoul, South Korea, in 2020 and 2022, respectively, where she is currently pursuing the Ph.D. degree in imaging engineering with the Graduate School of Advanced Imaging Science, Multimedia and Film. Her current research interests include deep learning and computer vision.

JUNEHYOUNG KWON received the B.S. degree in philosophy from Kyung Hee University, Seoul, South Korea, in 2020, and the M.S. degree in digital imaging from Chung-Ang University, in 2022, where he is currently pursuing the Ph.D. degree with the Graduate School of Artificial Intelligence. His research interests include deep learning and computer vision.

DONGMIN LEE received the B.E. and Ph.D. degrees in civil and architectural engineering from Korea University, Seoul, South Korea. Since 2021, he has been an Assistant Professor with the School of Architecture and Building Science, Chung-Ang University. His research interests include the integration of construction equipment, methods, planning, scheduling, and control to support a better human–robot collaborative working environment. In this context, his current research focuses on improving project performance, such as cost, schedule, quality, safety, and sustainability, in the built environment by developing and testing of digital twin of physical assets, such as robots, workers, and materials, which can be used to simulate ''what-if'' scenarios using AI-based techniques, such as deep reinforcement learning.

YOUNGBIN KIM (Member, IEEE) received the B.S. and M.S. degrees in computer science and the Ph.D. degree in visual information processing from Korea University, in 2010, 2012, and 2017, respectively. From August 2017 to February 2018, he was a Principal Research Engineer with Linewalks. He is currently an Assistant Professor with the Graduate School of Advanced Imaging Science, Multimedia and Film, Chung-Ang University. His current research interests include data mining and deep learning.

CHANSIK PARK received the B.E. and M.E. degrees in architecture from Chung-Ang University, Seoul, South Korea, the M.S. degree from the University of Colorado at Boulder, Boulder, CO, USA, and the Ph.D. degree in construction management from the University of Florida. Since 1995, he has been a Professor with the School of Architecture and Building Science and the Former Dean of the Graduate School of Construction Engineering, Chung-Ang University. He is one of the founders and the Former President of KICEM, the Founder of ICCEPM, and the Vice President of Building Smart Korea.

DOYEOP LEE received the master's degree in architectural engineering and the Ph.D. degree in construction engineering and management from Chung-Ang University, Seoul, South Korea. His current research interests include construction automation, computer vision, job hazard analysis, construction safety, construction quality, blockchain, and BIM.