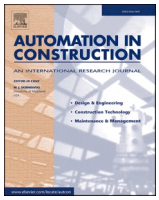
Automation in Construction 149 (2023) 104779

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon



Context-aware safety assessment system for far-field monitoring

Wei-Chih Chern [a], Jeongho Hyeon [b], Tam V. Nguyen [c], Vijayan K. Asari [a], Hongjo Kim [b,*]

[a] Department of Electrical and Computer Engineering, University of Dayton, Dayton, OH 45469, USA
[b] Department of Civil & Environmental Engineering, Yonsei University, Seoul, South Korea
[c] Department of Computer Science, University of Dayton, Dayton, OH 45469, USA

ARTICLE INFO

Keywords:
Context-aware safety monitoring
Object detection
Semantic segmentation
Depth estimation
Personal protective equipment
Construction safety

ABSTRACT

It is important to identify the working context of workers for a precise safety assessment for a personal protective equipment compliance check. However, very little attention has been paid to context-aware safety monitoring based on computer vision. To fill the knowledge gap, this study presents vision-based monitoring methods for context-aware PPE compliance checks, building upon a modularized analysis pipeline composed of object detection, semantic segmentation, and depth estimation models. Under the proposed system, two different approaches and their effectiveness are tested with real-world construction site data called YUD-COSAv2. In experiments, the proposed system was able to differentiate workers at height and on grounds to apply different PPE compliance rules to assess the safety status of workers, reporting Average Precision of 78.50% and 86.22% with the depth estimation model and the semantic segmentation model, respectively. This work will generate fresh insight into designing context-aware construction safety monitoring in far-field settings.

1. Introduction

Statistics showed [1] that fall accidents accounted for 33.5% of all construction deaths in the United States in 2018. Workers should properly use fall protection equipment when working at heights as an essential way to prevent fall accidents. That is, workers should attach their safety hooks to a stationary object to prevent falls. However, this safety rule is frequently violated due to the inconvenience of attaching and detaching safety hooks when moving from one place to another at heights.

Monitoring safety compliance using CCTV (closed-circuit television) could be a possible approach to alert the absence or improper use of fall protection equipment. Since CCTV is usually installed at an elevated position and captures a wide area in far-field monitoring settings, workers at heights and on the grounds can be recorded in the same video footage. However, monitoring systems could produce false alarms for workers on the grounds due to the absence of fall protection equipment if the systems cannot differentiate the working context—working at heights or on the grounds. As video footage is a 2D projection of the 3D real world, raw images do not provide depth information. Therefore, it is hard to calculate an elevation for each worker, which can be further used to recognize the working context.

Although previous studies have presented promising ways to analyze safety information, there has been no detailed investigation of a safety analysis method in far-field monitoring by distinguishing workers at heights and on grounds directly from videos. It could be a significant limitation as many false alarms can be produced if monitoring systems evaluate the safety status of workers based on the assumption that they are in the same working context. Multiple workers with different working contexts are often observed in far-field monitoring settings. Therefore, the limitation could be more disadvantageous in making robust safety monitoring systems. In addition, identifying personal safety equipment for each worker at close distances has not been investigated in detail in the previous literature. If a safety hook is detected between two workers, a monitoring system should be able to determine who possesses the hook for accurate safety monitoring.

To fill the knowledge gap, this study proposes a context-aware safety analysis method to identify the absence of proper personal protective equipment (PPE), such as a hardhat, a safety harness, a safety strap, and a safety hook, considering the working context. Specifically, the proposed method distinguishes workers at heights and on grounds to determine required PPE types and analyze their safety status. To this end, two different monitoring system variants and their effectiveness for context-aware safety monitoring are investigated: One with three

* Corresponding author.
E-mail addresses: chernw1@udayton.edu (W.-C. Chern), hyeon9404@yonsei.ac.kr (J. Hyeon), tnguyen1@udayton.edu (T.V. Nguyen), vasari1@udayton.edu (V.K. Asari), hongjo@yonsei.ac.kr (H. Kim).

https://doi.org/10.1016/j.autcon.2023.104779
Received 27 July 2022; Received in revised form 6 January 2023; Accepted 31 January 2023
Available online 14 February 2023

different convolutional neural networks (CNNs) for object detection, semantic segmentation, and depth estimation is proposed. The other one with CNNs for object detection and semantic segmentation that also perform height estimation with explicit height labels is presented. The object detection and semantic segmentation models are employed to recognize individual workers and their PPE for PPE assignment and PPE connectivity analysis to generate a relation vector for the safety classification of each worker. The depth estimation model is employed to estimate the distance between a camera and a worker, and the estimated distance is used to identify the working context. Based on the working context and the instance segmentation results, the safety status of a worker is evaluated.

Experiments were conducted on images collected from real construction sites dataset YUD-COSAv2, an extension of the YUD-COSA dataset introduced in [2]. In YUD-COSAv2, two construction scenes were added: fence installation for a training set and the other with installed fences for a testing set, as shown in Fig. 1. The experimental results showed that the proposed method could distinguish workers at heights or on grounds and analyze their safety status. The importance and originality of this study are that it explores the potential of depth estimation on 2D images without complex multi-view geometry algorithms (approach #1) and the use of explicit height labels for semantic segmentation (approach #2). Additionally, this study demonstrates the effectiveness of a novel context-aware safety analysis method in assessing the safety status of workers in far-field monitoring settings. The findings should make an important contribution to the construction safety management and automation field.

2. Related work

2.1. Computer vision-based safety monitoring

In recent years, there has been an increasing amount of literature on deep learning applications in the construction domain. Construction safety monitoring is one of the major areas in this research trend because the construction industry continues to be the most dangerous industry considering the number of fatal occupational injuries in 2016–2020 [3]. Before the introduction of AlexNet [4], most previous studies utilized traditional machine learning models to analyze construction site images [5]. Along with the initial stage of visionbased image analysis, promising ways to analyze construction safety in images have been proposed. To name a few applications, spatial contexts of construction entities have been analyzed by using visual information such as the location and class of objects to measure the risk of struck-by accidents [6]. The potential of a motion-capture approach has also been presented to analyze construction work-related musculoskeletal disorder [7]. The findings of the previous studies laid the foundation of recent vision-based methods in construction site monitoring.

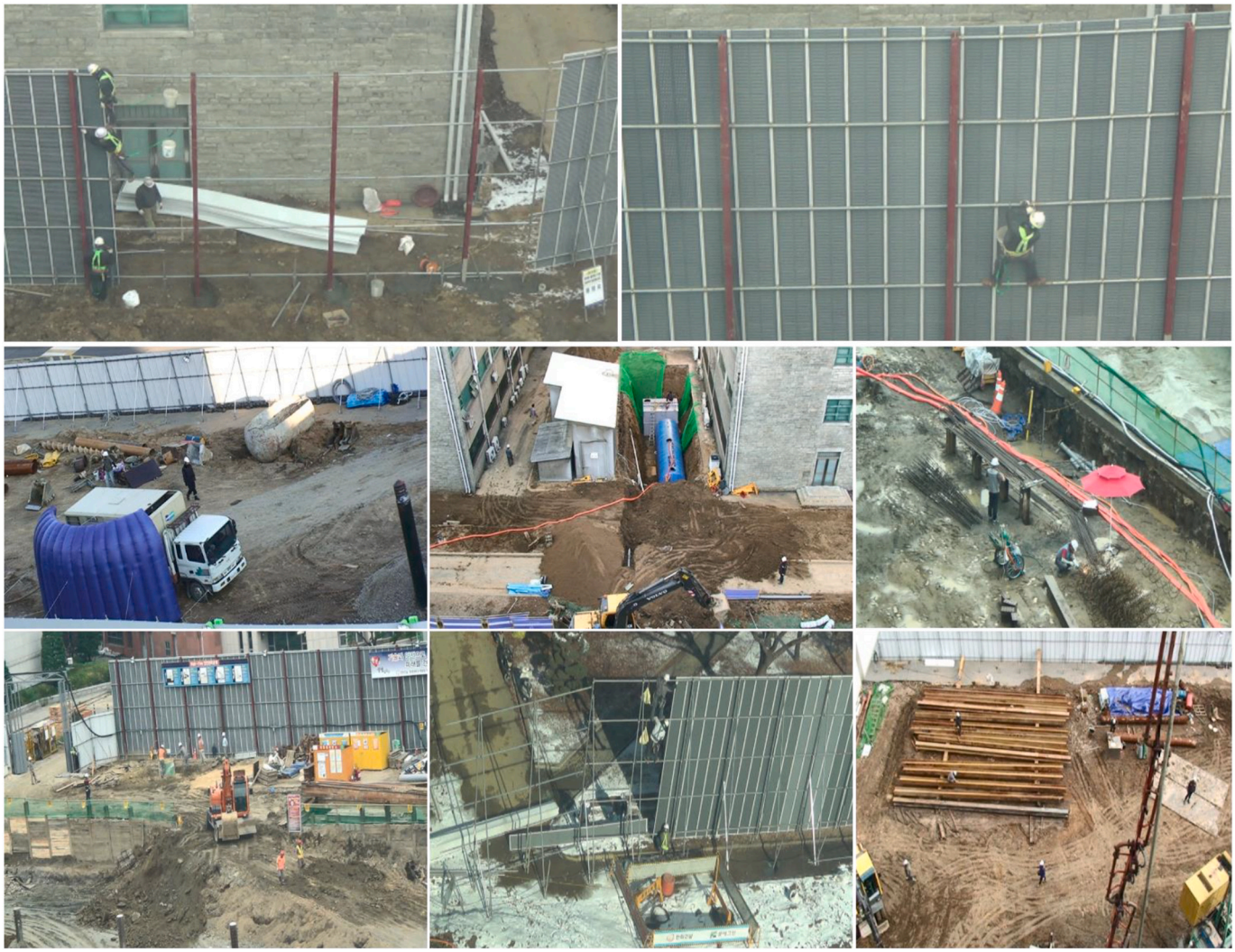


Fig. 1. Two extra scenes were added in YUD-COSAv2 for experiments: the top-left scene for training, and the top-right scene for testing. Various scenes from 2_{nd} and 3_{rd} rows were displayed from the training set.

Deep learning models have been actively investigated for analyzing construction site and infrastructure images in recent years. Early studies examined the applicability of CNN-based object detection models for construction site monitoring [8] and infrastructure assessment [9], demonstrating the superior performance of CNNs. Owing to the high accuracy of deep learning models for visual tasks, researchers suggested novel safety monitoring methods for various applications. Some studies showed that spatial information of construction sites could be reconstructed from 2D images using CNN models to prevent struck-by accidents [10,11]. The findings of these studies suggest that 3D spatial information can be reconstructed from monocular images. Researchers employed object detection models based on CNN architecture to recognize PPE such as hardhats and harnesses that are hard to be recognized by using the traditional detection methods due to lack of visual features [12].

Instance segmentation provides pixel-level object localization beside object bounding boxes. Segmentation masks are helpful for context analysis based on precise localization, and their use can be extended to object tracking. Fang et al. [13] proposed a framework to facilitate early safety warning using the prior knowledge model with object bounding boxes, semantic masks, and key points detection. Khan et al. [14] proposed a computer vision-based risk recognition system to detect workers' safety behaviors using a Mask-RCNN and object correlation detection for mobile scaffolds monitoring. Kang et al. [15] proposed a one-stage instance segmentation model trained to detect workers under various weather conditions. Xiao et al. [16] proposed a worker tracking framework based on instance segmentation, association, and assignment, which handle occlusion and scale variations for worker tracking in off-site construction. Overall, these studies highlight the use of instance segmentation in construction site monitoring. However, there remains a paucity of knowledge on using instance segmentation in far-field monitoring. In such monitoring conditions, recognition of small objects is challenging in general. Moreover, there has been little investigation into identifying the working context (working at heights or on the ground) based on instance segmentation results in far-field monitoring, despite its importance in safety monitoring applications.

2.2. Detection of fall protection equipment

To prevent falls from heights, region-based CNN models were used to identify the wearing of fall protection equipment [17] and recognize unsafe behaviors [18]. Xiong and Tang [19] proposed a pose-guided anchoring method to select specific body regions for PPE usage classification. Shen et al. [20] proposed detecting safety helmets using a face detector with bounding-box regression to locate the helmet area. Chen and Demachi [21] proposed using YOLOv3 and OpenPose framework for scene information identification that can generate worker's relation with PPE and tool for hazards identification. Ma et al. [22] proposed a lightweight YOLOv4 detector that directly recognizes the difference between workers, workers with hardhats, and workers with harnesses to monitor improper usage of PPE. Li et al. [23] proposed a framework of YOLOv5 and OpenPose to generate features for a 1DCNN classifier to inspect hardhats and harnesses on workers. Cheng et al. [24] proposed to monitor workers among different cameras by combining worker re-identification with similarity loss and PPE classification with a weighted-class strategy. The previous studies showed promising ways of recognizing PPE and its proper use. Nevertheless, a systematic understanding of how and which pieces of PPE are connected to a specific worker in far-field monitoring is still lacking.

2.3. Far-field monitoring

Nowadays, CCTV (Closed Circuit Television) is prevalent in construction sites due to its advantage of monitoring wide areas with a simple camera system. However, there are challenges in analyzing CCTV video frames collected by far-field monitoring settings, such as varying scales of target objects [2]. While addressing such a challenge, previous studies showed great potential for far-field monitoring. Luo et al. [25] proposed a two-step method of deep action recognition and a Bayesian non-parametric hidden semi-Markov model to infer worker activities in far-field surveillance videos for tasks such as worker fatigue monitoring. Yan et al. [26] proposed a 3D bounding box reconstruction for detected construction workers and vehicles, which can be used to understand the 3D spatial relation for analyzing the risk of struck-by accidents. Zeng et al. [27] proposed an improved YOLOv3 architecture and grey wolf optimized extreme learning machine to detect various scales of construction vehicles in far-field monitoring. Assadzadeh et al. [28] proposed a far-field camera calibration without manual intervention using object detection and geometrical scene analysis for construction safety and productivity analysis. Fang et al. [29] proposed a Mask R-CNN based method to estimate the working context of workers by retrieving bounding boxes and segmentation masks to analyze the relations between workers and elevated structural supports to mitigate fall accidents. The previous studies suggested securing reliable recognition performance in far-field monitoring settings for safety or productivity analysis. Nonetheless, there has been no detailed investigation of analyzing the working context of workers based on depth information that can identify who is working at height or on the ground.

Overall, the literature review reveals a knowledge gap in far-field construction site monitoring for assessing the safety status of workers considering their working context—working at height or on the ground—based on a single 2D image. Although the working context estimation is beneficial to accurately evaluate the safety status of workers considering different safety rules, there has been little investigation on how to consider working contexts in safety assessment. Especially in recognizing working contexts based on height information, very little attention has been paid to the role of semantic segmentation and depth estimation models. Moreover, previous safety monitoring systems mainly relied on detection results to check the presence of target objects related to PPE, which could produce false alarms as they could not be aware which PPE belongs to which workers due to occlusions. To fill the knowledge gap, this study investigates far-field monitoring methods that can infer the working context and analyze the safety status of each worker in an image with visual recognition results.

3. Proposed method

The proposed method aims to classify the safety status of workers considering the presence of essential PPE (hardhat, harness, safety strap, safety hook) based on their working context—working at height or ground—using object detection, semantic segmentation, and depth estimation. To this end, this study proposes two approaches: One with a depth estimation model and the other using a segmentation model with explicit height labels, as shown in Fig. 2. The first approach directly estimates a working height (an elevation where a worker stands up) of a worker using a depth estimation model. This approach is designed for monitoring environments with CCTV installed at an angle looking down from an elevated location, assuming that workers at height are located closer to the camera than workers on the ground. If a camera is installed at an angle looking up from the ground, the method should reversely assume that workers at height are located further away than workers on the ground. The other approach does not estimate the working height. However, it predicts the working context explicitly using a semantic segmentation model with working context labels such as 'working at height' or 'working on the ground.' In Section 5.2, the pros and cons of each method are discussed.

In the second approach, one observation inspired the idea of using a semantic segmentation model for the working context estimation. Depth estimation models in recent studies adopt encoder-decoder designs as semantic segmentation models. That is, depth information of a specific pixel is estimated by information of surrounding pixels. Similarly, the working context labels— height (working at height) or ground (working

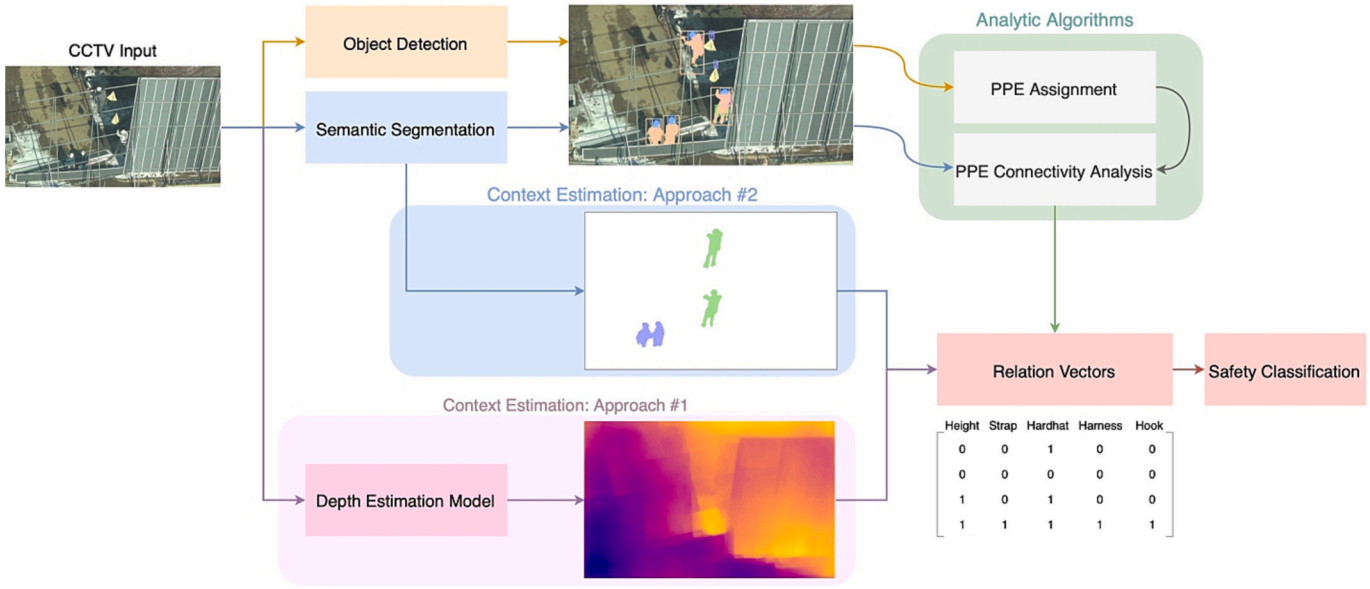


Fig. 2. The overview of the proposed safety monitoring system.

on the ground)—can be predicted in the same manner if the segmentation masks of workers have those additional labels. A similar finding was presented by Nath et al. [30]. The authors used a detection model to detect workers as safe or unsafe—training the detection model to recognize necessary information, such as PPE, with safe and unsafe bounding box labels only. In light of the observation and the previous study, the second approach considers the working context estimation via explicit labels as a subset of a data-driven problem in supervised learning.

Three main challenges are associated with workers' safety assessment in construction site monitoring. First, target objects have varying scales and may have similar colors compared to backgrounds in far-field monitoring. The first challenge often results in a lack of visual features of workers and PPE, which leads to suboptimal accuracy for object detection and semantic segmentation. Second, identifying the working context of workers is not a trivial task since the input data, video frames, does not have 3D information. Thus, estimated working heights of workers based on 2D images could have a large error range depending on estimation algorithms, camera settings, or weather effects. Third, it is difficult to identify which PPE belongs to which workers based on detection and segmentation results. There might be occlusion or disconnection between target objects, which hinders checking PPE compliance in different working contexts. To address these challenges, the two proposed methods take different strategies in estimating working contexts but share common recognition modules such as object detection, semantic segmentation, and analytic algorithms for safety assessment. Both systems have a modular system design, which does not depend on specific modules in detection, segmentation, or depth estimation parts. Each module can be replaced by other models tailored for recognition performance or computational efficiency.

3.1. Recognition models

This study uses object detection to distinguish object instances in the same object class. Semantic segmentation is used to find accurate boundaries of target objects. A modified instance segmentation method based on object detection and semantic segmentation is designed, which has a different mechanism than a common instance segmentation approach, which only segments pixels in bounding boxes. The modified instance segmentation method first detects target objects. Then, semantic segmentation is conducted on all pixels of the whole image. The reason is that some small objects, such as safety straps and hooks, might not be detected by a detection model. However, a segmentation model could recognize them due to a different processing mechanism between the models. The independent use of detection and segmentation could recognize more challenging target objects even if the object detector fails to identify them. Therefore, the two models are complementary in generating visual information from image data and analyzing the working context.

A YOLOv5 (You Only Look Once Version 5) [31] model is used for object detection as it is considered as a decent and fast object detection model for various applications. However, a different detection model can be used according to the needs for specific monitoring scenes and applications. For example, a lighter and faster detector such as YOLOv7 (You Only Look Once Version 7) [32] can be employed to improve the performance of the safety assessment. YOLOv5 comprises the BottleNeck Cross Stage Partial Network (BottleNeckCSP) and Spatial Pyramid Pooling (SPP) Layers. That is, the model includes the architecture of reducing and resuming the number of feature maps (BottleNeck), the architecture of reusing and passing feature maps into layers in the back and preventing duplicate gradient back-propagation during training (Cross Stage Partial Network: CSP), and the architecture of fusing pooling results of varying scales of convolutional kernels for object detection. An FPN (Feature Pyramid Network) [33] segmentation model is used for semantic segmentation. The FPN model possesses an encoder-decoder architecture with skip connections that provide extracted features from an encoder part of the network to a decoder part for better performance. Additionally, the FPN model utilizes each stage's feature maps (feature pyramid) from the decoder in the output layer for prediction. For training the FPN model, compound loss functions composed of Jaccard and focal losses were used to improve the performance of more challenging objects [2]. Target labels of FPN will vary depending on the two different safety assessment approaches. The first one uses five target labels such as 'worker', 'hardhat', 'harness', 'strap', and 'hook'. The second approach uses two additional labels, 'working at height' and 'working on the ground.'

3.2. Checking the presence of PPE for workers

It is necessary to assign detected PPE to individual workers, to verify whether a worker is equipped with proper PPE based on his or her working context. This study defines this process as 'PPE Assignment.' An

object detection model, YOLOv5, is employed to identify instances of each class since semantic segmentation cannot differentiate individual instances of the same class. After applying object detection, bounding box coordinates are used for PPE assignment by calculating an intersection over the PPE area between detected workers and PPE, as shown in Eq. 1. Each PPE will be used to calculate overlap scores between all workers and assigned to the worker that shares the highest degree of overlap, and the score should be larger than 0.1. The equation of overlap calculation given bounding box coordinates is formulated as follows:

$$Overlap(p_i, w_j) = \frac{I(p_i, w_j)}{A(p_i)}, \tag{1}$$

$$I(p_i, w_j) = max(0, A(min(p_{ibr}, w_{jbr}) - max(p_{itl}, w_{jtl}))), \tag{2}$$

$$A(pi) = pibr - pitl \tag{3}$$

where I is a function to calculate the intersection, A is a function to calculate the total area given top-left and bottom-right coordinates, p_i represents detected PPE, w_j represents detected workers, p_i and w_j contain top-left and a bottom-right 2D coordinates denoted with $_{tl}$ and $_{br}$ subscripts, respectively. Functions I and A are defined as below:

However, there is an exception for a safety hook of a worker at height. It can be shown away from workers as they are attached to scaffolds. This case results in a lower overlap score, although a worker is equipped with it. This study uses a two-step PPE assignment for safety hooks to address this issue. PPE connectivity analysis is conducted to check whether pixel regions are connected between a worker and assigned PPE based on segmentation masks. For that, a new bounding box that includes a worker and assigned PPE will be produced and used as input for PPE connectivity analysis. This analysis will generate a sub-region segmentation mask of the bounding box to ensure the worker wears the PPE based on the connectivity between target classes. Then, a depth estimation model or an extended semantic segmentation model can estimate the height estimation to determine whether a worker is on scaffolds or on the ground to apply safety standards accordingly. By gathering all the information, a relation vector is generated for each worker for safety classification.

Algorithm 1 PPE assignment

```
Input:
    BBoxes: Bounding boxes results from the object detector
Output:
    WorkerPPE: BBoxes of workers and their assigned PPE

function PPE Assignment(BBoxes)
    for BBox_ppe in BBoxes do
        if BBox_ppe type is hook then
            continue
        for BBox_worker in BBoxes do
            Scores ← overlap(BBox_ppe, BBox_worker)
        if max(Scores) ≥ 0.1 then
            WorkerPPE[worker_idx] ← BBox_ppe
        else
            continue

    return WorkerPPE
end function
```

3.3. PPE connectivity analysis

A semantic segmentation model is used to acquire pixel-level localization of workers and PPE for connectivity analysis that verifies if the assigned PPE from object detection is connected to the assigned worker. In this study, an FPN model is used to extract segmentation masks of workers and PPE in the entire image region. This way, target objects that the object detector might miss could be identified, and the identified objects are considered in the PPE connectivity analysis. Then, the connectivity between a worker and PPE in a bounding box can be analyzed. The segmentation masks are used to determine what PPEs are connected to which worker, as shown in Algo. 2. The bounding boxes of a worker and assigned PPEs are used to extract the segmentation masks for connectivity analysis.

An image processing algorithm, Flood Fill, is used to proceed with the connectivity analysis between workers and assigned PPE. The Flood Fill algorithm recursively searches for neighbor pixels with values within a threshold to validate the connectivity between a worker and assigned PPE. In bounding boxes that include workers and their assigned PPE, sub-region segmentation masks for target classes are extracted by FPN. Then, the masks are used as the input of Flood Fill for the connectivity analysis, as shown in Fig. 3-A. Before applying the Flood Fill algorithm, pixel intensities of segmentation masks are amplified so that each class can be better visualized and assessed because the original pixel intensities of foregrounds range from 1 to 5. The amplification results are shown in Fig. 3-B. Then, a seed point should be determined as a starting point of Flood Fill to search for connected pixels between a worker and PPE. With the sub-region segmentation masks, it is possible to find the center point of the worker's pixels and select it as the seed point, as shown in Fig. 3C. The Flood Fill algorithm outputs a mask of the worker and PPE pixels with intensity values of 255 as an indication of connected objects. The bit-wise AND operation is then performed between the results in Fig. 3-A and Fig. 3-C to generate the final connectivity analysis result (Fig. 3-D). Eventually, Fig. 3-D is treated as an array to search for unique index values, indicating the PPE connected to the worker, as shown in Line 7 in Algo. 2.

3.4. Working context analysis

This study proposes two approaches for working context analysis, using a pre-trained depth estimation model [34] or explicit working context labels, respectively. In the first approach, the depth estimation model consists of an encoder-decoder architecture for depth estimation. With an architecture similar to semantic segmentation models, the depth estimation model estimates distances from a camera to each pixel. The depth estimation model used in this study was trained by the NYU Depth V2 dataset [35]—an indoor furniture dataset collected by a Microsoft Kinect depth camera. The use of the depth estimation model is based on an assumption about the camera location and its view. For instance, workers at height are closer than workers on the ground if a camera is recording a video at an angle looking down at the scene. Conversely, when a camera is on the ground recording a video at an angle looking up, workers at height are farther away than workers on the ground. In this study, the testing images were collected from a camcorder with a top-down angle. Therefore the workers with smaller depth values will be considered as at height. Since the depth model was not trained for construction scenes, the accuracy of the estimated depth map is not ideal, as shown in the middle column of Fig. 4. The accuracy could be better if training data is prepared based on construction site scenes, but it is not trivial due to a hardware limitation of the depth camera. Considering the safety monitoring application, it would be sufficient if workers' working contexts could be correctly classified based on the estimated depth map. For that, this study binarizes the estimated depth map into 'ground' and 'scaffold' regions using a threshold value. Different thresholding methods have been experimented such as Otsu's method, mean value, and simple thresholding. It was found that the mean value acquired the best safety assessment performance among other thresholding methods for the dataset in this study. The binarization of the depth map is conducted using the mean value of the estimated distance values of all pixels. As a result, the working contexts—working at height or working on the ground—can be determined, as shown in the rightmost column of Fig. 4. However, it should be noted the optimal value for thresholding would vary for different construction scenes.

Algorithm 2 PPE connectivity analysis.

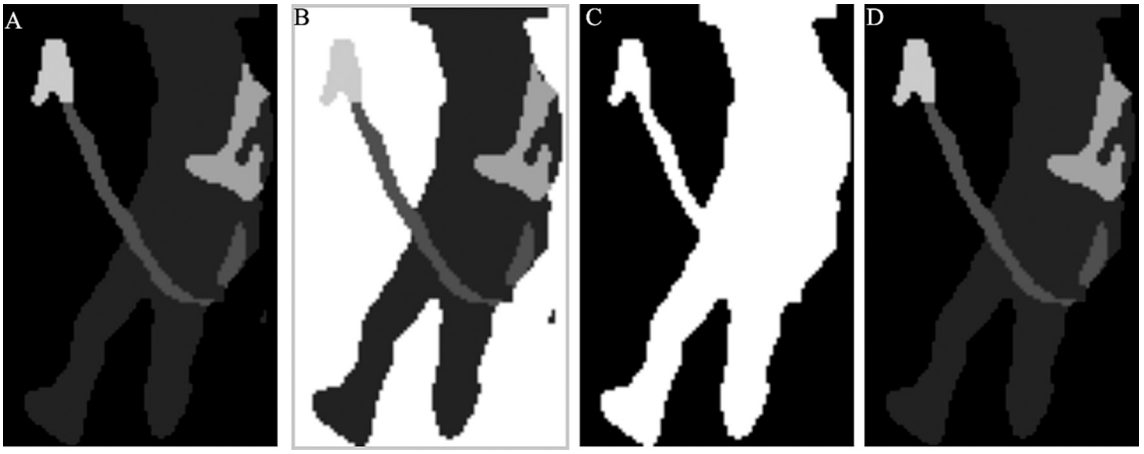


Fig. 3. The process of PPE connectivity analysis. Worker and PPE pixel intensity values are amplified for better visualization except for step C. (A) Sub-region segmentation mask of a worker and assigned PPE using overlap comparison from object detection. (B) Pre-processing step to turn background pixels to the intensity value of 255 for Flood Fill. (C) An output mask of the Flood Fill algorithm. Connected foreground pixels are assigned with the intensity value of 255. (D) A bitwise AND operation is performed between results A and C, then the unique indices for connected PPE to the worker are determined.

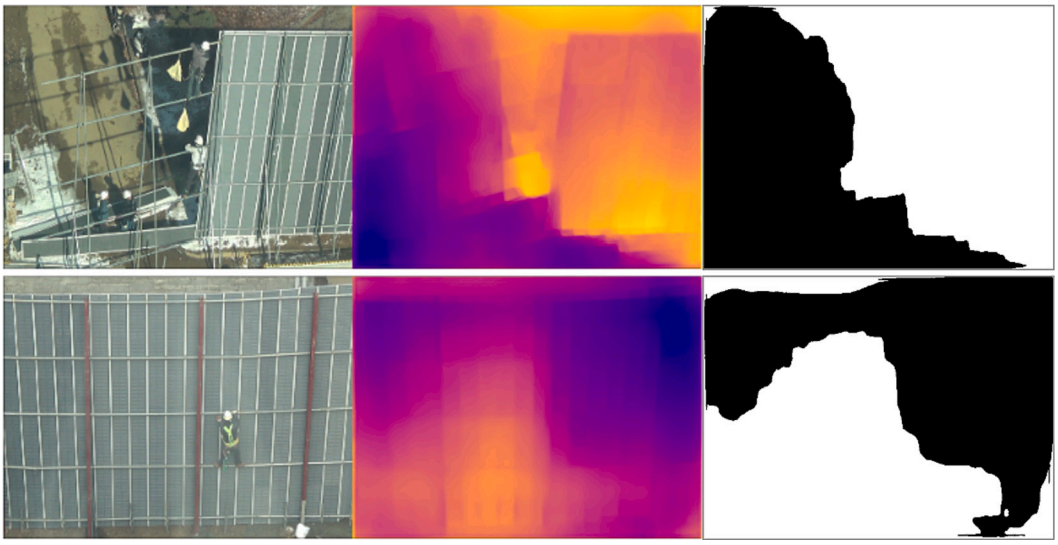


Fig. 4. Results of depth estimation and mean value thresholding. (From left to right columns) Original images, Depth maps, and Thresholded maps.

```
Input:
    BBox: BBoxes that includes workers with assigned PPE
    Mask: Segmentation mask
Output:
    WorkerPPE: Workers' connected PPE after analysis

function PPE Analysis(BBox, Mask)
    for BBox_i in BBox do
        Sub_Mask_i ← Mask[BBox_i]
        Sub_Mask_i[background] = 255
        ConnectMask ← FloodFill(Sub_Mask_i, Seed)
        ConnectMask ← AND(ConnectMask, Mask_i)
        PPE_indice ← Unique(ConnectMask)
        WorkerPPE_i ← PPE_indice
    return WorkerPPE
end function
```

In the second approach to estimating the working context, target class labels of semantic segmentation are extended to estimate a working context besides workers and PPE directly, as shown in Fig. 5. The output layer of the original segmentation model has the shape of W × H × C, where C equals 6 (background and five foreground classes). Additional two classes are included to have a total of 8 target classes–background, worker, hardhat, strap, harness, hook, worker on the ground, and worker at height–for the output layer. This version of semantic segmentation is used in the second safety assessment approach. In this approach, a pixel can be labeled with two classes:' worker' and 'working at height.' Except for the class 'worker', the other classes still have a single target class label without a working context label. The softmax function is also replaced by the sigmoid function in the output layer since a pixel with the 'worker' label has an additional label of working context. Now, the ground truth label of each pixel can be single or double. In annotation, this study reuses the original worker polygon annotations and includes the extra labels of worker on the ground and worker at height. As a result, the segmentation model is trained to recognize workers and workers' working context together during the training process, as shown in Fig. 5.

3.5. Safety analysis

Based on the recognition results from the previous processing steps

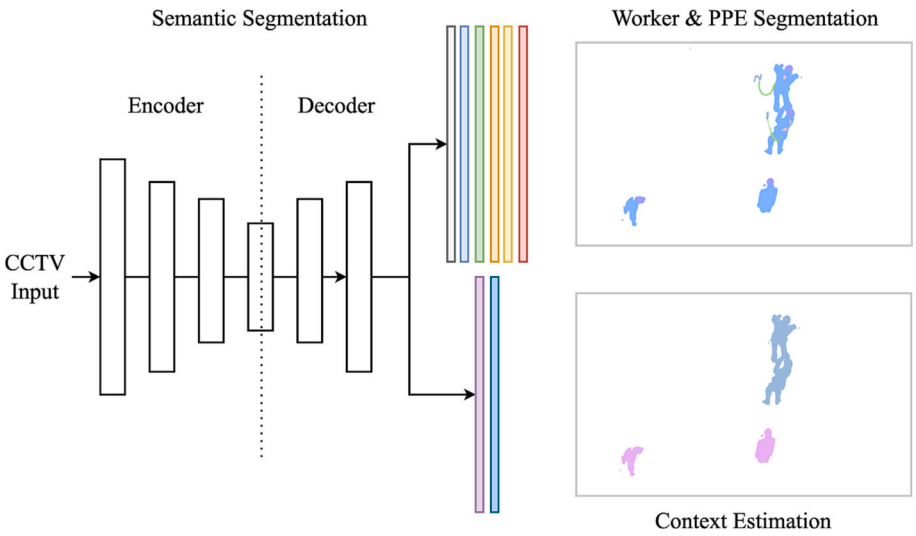


Fig. 5. The extended semantic segmentation model for worker, PPE, and working context. For working context estimation, the polygon annotations of workers are reused and additional labels are included such that workers on ground (purple) and at height (blue) can be learned by the model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(PPE assignment using object detection, PPE connectivity analysis using semantic segmentation, and working context analysis), the safety status of a worker can be determined considering the different working contexts. The criteria are: For a worker on the ground, he or she should wear a safety hardhat. For working at height, he or she should wear a hardhat, a harness with a safety strap, and a safety hook. These rules can be checked by the presence of necessary PPE for workers considering their working contexts, as shown in Algorithm 3. This algorithm assumes that if a worker and necessary PPE appear together in close proximity, he or she uses PPE correctly.

A relation vector is generated for each detected worker to determine whether a worker is wearing and carrying the necessary PPE. The relation vector represents the working status of a detected worker and his or her PPE. For example, if a relation vector indicates that a worker on the ground is wearing a hardhat, such as [00100], the safety status of the worker is determined as safe. Likewise, a relation vector indicates that a worker at height is wearing a hardhat and harness with a safety strap and a safety hook, such as [11111], the status will be determined as safe. If a worker misses one of the essential PPE, then the status will be considered unsafe. The bounding box coordinates between a worker and an assigned hardhat are used to confirm the proper use of a hardhat. A hardhat assigned to a worker should be located at the top position of the bounding box of the worker. This study considers the top position by dividing the bounding box of a worker into four equally spaced boxes in a vertical direction, and the detected hardhat should be within the top box.

4. Experiment

4.1. Experimental settings

A workstation equipped with two NVIDIA RTX 3090 GPUs and an AMD Ryzen Threadripper 3970 × 32-Core CPU running on Ubuntu 20.04 was used for experiments. A variant of YOLOv5, YOLOv5m, was used as an object detector at the resolution of 960 × 544, which has a medium network size to balance the inference speed and the performance. A dataset used in this study was collected from real construction sites. It is named YUD-COSAv2 (Yonsei University - University of Dayton Construction Safety version 2), which includes additional image label information on top of the original YUD-COSA [2]. The total number of images in the dataset is 1089. The training set contains 840 images, and the testing set contains 249 images. The image resolution was adjusted to 960 × 544 in the experiments. The YOLOv5m pre-trained using Microsoft Common Objects in Context (COCO) [36] was employed for transfer learning. The YUD-COSA dataset updated the entire layers of the model. The network complexity and input resolution of YOLOv5 were selected based on empirical evidence from multiple experiments to balance detection performance and inference speed. The mean average precision (mAP) was used to evaluate the detection performance, as shown in Eq. 6. The data augmentation techniques used for training the YOLOv5 were Mosaic [37], vertical and horizontal flips, HSV color jittering, random perspective transform, and mixup [38] augmentations. The performance of the model is shown in Table 1.

Algorithm 3 Worker safety classification.

```
Input:
    F: New frame from CCTV
Output:
    Each worker's safety status

function Safety Classification(F)

    if UseDepthModel == True then                    ▷Acquire recognition results
        BBoxes ← ObjectDetection(F)
        Worker_PPE_Mask ← SemanticSegment(F)
        Depth ← DepthEstimation(F)
    else
        BBoxes ← ObjectDetection(F)
        Worker_PPE_Mask, Depth ← ExtendSegment(F)

    Worker_PPE ← PPE_Assignment(BBoxes)              ▷Start analysis
    Worker_PPE ←
Connectivity_Analysis(Worker_PPE, Worker_PPE_Mask)

     if PPE == hook then
         if Dist(BBox_hook, BBox_worker) < thres then
             Worker_PPE ← BBox_hook

    Worker_PPE ← Hardhat_Check(Worker_PPE)           ▷Hardhat position check
    Relation_Vector ← Worker_PPE                     ▷Output analysis results

    Cls_Results ← Relation_Vector

    return Cls_Results
end function
```

The input resolution for the FPN segmentation model was the same as 960 × 544. FPN was pre-trained on ImageNet [39] for transfer learning, and the YUD-COSA dataset updated the entire model. The data augmentation used for training FPN included horizontal flip, shifting,

Table 1
Performance of YOLOv5m and FPN with the resolution of 960 × 544. In the second row, the scores indicates Average Precision for each class. In the third row, the scores indicates IoU for each class.

Network	mAP*/mIoU	Worker	Strap	Hardhat	Harness	Hook	Height	Ground
YOLOv5m	85.3%	99.4%	58.8%	99.4%	91.9%	77.0%	–	–
FPN	75.74%	82.98%	50.45%	89.51%	71.48%	59.79%	89.03%	86.94%

* mAP score was calculated with overlapping (IoU) threshold of 0.5: mAP@0.5.

scaling, rotating, HSV color jittering, and Copy-Paste [40]. The Copy-Paste technique was used to augment more challenging target classes, such as safety straps, harnesses, and hooks, to learn the representation more frequently during training. The mean intersection over union (mIoU) was used as the performance metric to evaluate the segmentation performance, as shown in Eq. 7. The performance of the FPN model is shown in Table 1.

4.2. Evaluation metrics

For object detection, mAP was used as an evaluation metric for the detection model, and it is the mean of average precision (AP) of each class. AP is the average of multiple precision values with respect to different recall values depending on a range of prediction confidence thresholds (1000 intervals from 0 to 1) with a fixed IoU of 50% between ground truth and predicted bounding box [41]. The formulas are shown as follows:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$mAP = \frac{\sum_n (R_n - R_{n-1}) P_n}{n} \tag{6}$$

where *TP* represents true positive, *FP* represents false positive, *FN* represents false negative, R_n represents Recall at n_{th} confidence threshold, and P_n represents Precision at n_{th} confidence threshold. Precision represents a detection model's accuracy among the detected objects. Therefore, a detection model with fewer false positives will have a higher Precision score. Recall presents a detection model's accuracy to ground truth objects. Therefore, a detection model with less missing detection will have a higher Recall score. Since Precision and Recall are both important characteristics for evaluations, AP uses the area under the Precision-Recall curve as a metric that considers both Precision and Recall for performance evaluation.

For semantic segmentation, mIoU is used as an evaluation metric to evaluate segmentation performance. The formula is shown as follows:

$$mIoU = \sum_{n=1}^{C} \frac{IoU_n}{C} \tag{7}$$

where it sums up the IoU score of each class and averages it by the number of object classes, C. The mIoU score is a commonly used performance metric for semantic segmentation in reputable journals and conference proceedings. Datasets such as the Pascal Visual Object Classes (VOC) [42] and COCO also use the mIoU score as a performance metric for semantic segmentation tasks.

To evaluate the classification of the safety status of a worker, Precision (Eq. 4), Recall (Eq. 5), and F1 [43] were used. A confusion matrix, as shown in Table 3, is provided to demonstrate the safety system's statistics: TP, TN, FP, and FN. Similar to the mAP metrics for evaluations to object detection, the F1 score is another metric that takes both Precision and Recall into account for performance evaluation. The formula for F1 score is shown as follows:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{8}$$

4.3. Experimental results

Table 1 shows the performance of object detection and segmentation. The mAP score of YOLOv5m for five target classes—worker, strap, hardhat, harness, and hook—is 85.3%. The AP scores for worker, hardhat, and harness are above 90%, while the AP scores for more challenging objects are 58.8%, and 77% for strap and hook, respectively. Additionally, the precision-recall curve as shown in Fig. 6 also indicated that YOLOv5m was less optimized for the classes such as straps and hooks. The mIoU score of the FPN model for seven target classes—worker, strap, hardhat, harness, hook, height (working at height), and ground (working on ground)—is 75.74%. Similar to the YOLOv5, the segmentation performance of strap, harness, and hook are lower, recording 50.45%, 71.48%, and 59.79%, respectively. Additionally, the segmentation performance for working context estimation was decent, recording 89.03% for height and 86.94% for ground.

The safety assessment system based on the depth estimation model for the working context analysis scored the F1 of 91.18% with respect to the class 'Safe'. For the 'Unsafe' class, the system recorded the F1 of 70.53%.

On the other hand, the safety assessment system based on the second approach in which the extended segmentation model predicted the working context recorded the F1 score of 95.07% with respect to the class 'Safe'. For the 'Unsafe' class, the system recorded the F1 score of 83.52%. A confusion matrix, Precision, Recall, and F1 scores are shown in Table 3.

As shown in Table 2, the processing time for each method was separately calculated over 249 images from the entire testing set of the YUD-COSAv2. The YOLOv5m consumed an average time of 8.8 ms/image. The FPN consumed an average time of 65.8 ms/image. The depth estimation model consumed an average time of 61.9 ms/image. The safety analysis algorithm—PPE assignment, PPE connectivity analysis, and relation vector classification— consumed 6.4 ms/image.

5. Discussion

5.1. Performance of detection and segmentation

As shown in Section 4.3, YOLOv5m and FPN showed reasonable recognition performance on large object classes such as worker and hardhat, while recorded lower performance on small object classes such as harness, strap, and hook. This phenomenon was expected before the experiments, as previous studies, [44–46] identified the challenges of recognizing small objects in far-field monitoring settings. The challenges include a lack of visual features due to a small number of pixels, the color similarity between foreground and background, and irregular shapes of target objects. It was especially observed that safety harnesses suffer from color similarity more than other classes, resulting in poor recognition performance in both detection and segmentation. Moreover, the small objects such as harness, strap, and hook were more vulnerable to occlusion by other objects or workers' pose variance. Obviously, the most challenging object classes to recognize were safety straps and hooks due to their size, lack of visual features, and color similarity. Data

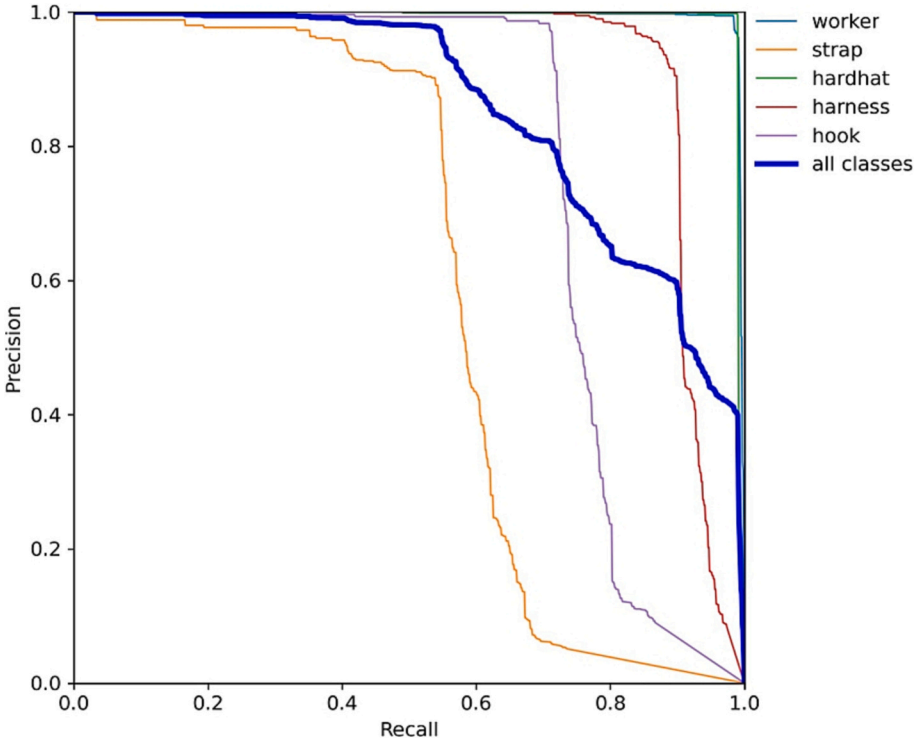


Fig. 6. Precision-recall curve of the YOLOv5m.

Table 2
Inference speed of each module from the proposed safety monitoring system.

Method	Speed
YOLOv5m	8.8 ms/img
FPN	65.8 ms/img
Depth Estimation	61.9 ms/img
Safety Analysis	6.4 ms/img

imbalance is also critical in recognizing the target objects in far-field monitoring. That is, large object classes such as worker and hardhat could significantly influence loss calculation, resulting in imbalanced recognition performance over target classes. In the YUD-COSAv2 dataset, data imbalance is evident, as shown in Table 4. Despite the challenges in recognition, the segmentation performance on small objects was able to be secured for reliable safety assessment in the following process by employing compound loss settings presented in Chern et al. [2] work. It was found that the combination of the YOLOv5 version 6 backbone and strong data augmentation techniques was able to secure the detection performance of the target classes. Recently, a more recent version of the YOLO series, YOLOv7 [32], was made public. For better performance, it can be used to increase the recognition performance on challenging objects, thereby improving the overall safety classification performance and processing speed.

The two proposed methods' most critical and frequent errors were observed from the fully protected workers at height. Those workers were classified as unsafe due to missing a safety hook in the relation vector. As a result, the precision for the Unsafe category suffers from lower precision for both methods (referred to Table 3). As discussed in Section. 3.2, a safety hook is assigned to a worker based on the distance between the worker and the assigned safety strap. A safety hook is assigned to a worker if the distance between the worker's strap and the hook is smaller than a threshold of the pixel distance 70. By increasing the distance threshold for the hook assignment, the F1 score for the Unsafe category could be improved as the precision could also be improved.

Table 3
Confusion matrices, and F1 scores of the proposed safety assessment systems for far-field safety monitoring. Mean F1 scores represents the average of F1 scores from the safe and unsafe categories

Method	Seg. Height		Depth Estimation	
	Unsafe	Safe	Unsafe	Safe
Predict Unsafe	180	66	152	94
Predict Safe	5	685	33	657
F1 Score	83.52%	95.07%	70.53%	91.18%
Mean F1	89.38%		80.85%	
Avg. Precision	86.22%		78.50%	

Increasing the distance threshold may improve the chance of detecting safety hooks but also increase the chance of false positives (false assignment to other adjacent workers).

5.2. Working context analysis

This study investigates context-aware safety assessment methods for construction workers with two different approaches. The first approach estimated the working context based on the depth estimation model. In the experiments, the first approach has demonstrated its effectiveness in estimating the working context, recording the F1 score of 91.18% for the

Table 4
The statistics of the YUD-COSAv2 dataset with an original image resolution.

Class	Train Pixel #	Test Pixel #
BK	1,455,882,317	602,669,399
Worker	17,069,619	14,256,146
Strap	748,357	440,181
Hardhat	1,493,783	1,443,986
Harness	1,949,468	986,841
Hook	223,496	209,847
Height	10,937,547	10,255,385
Ground	9,863,893	6,705,690

class 'Safe' and 70.53% for the class 'Unsafe' in the safety assessment. The results were interesting since the depth estimation model was not fine-tuned with data in the construction domain. It is expected that the performance of the safety assessment could be further improved if the depth estimation model is trained with data in the construction domain. The ground truth to train the depth estimation model could be obtained by LIDAR (Light Detection and Ranging) sensors.

The second approach estimated the working context based on the segmentation model with extended class categories—Height (working at height) and Ground (working on the ground). As a result, the model successfully estimated the working context, recording IoU of 89.03% for Height and 86.94% for Ground. The resulting performance in the safety assessment is better than the first approach. The mean F1 score of the safety assessment was 89.38%, which is 8.38% higher than the first approach.

The possible explanations for the difference in the safety assessment performance are conjectured as follows: The depth estimation performance of the first approach was not ideal since the depth estimation model was not re-trained with construction data sets. In the second approach, the segmentation model was able to learn essential visual features indicating the working context of the workers, such as scaffolds overlapping on worker's body, the presence of fall protection equipment (safety harnesses, straps, and hooks), and the surrounding pixels of workers in images. This can be explained by the concept of a receptive field in CNN models. A receptive field in segmentation models can be interpreted as the pixel region in the input image considered to make a final prediction of a pixel's category. As a pixel's category is determined by the surrounding pixels and their visual features, the working context could be well classified. However, since the YUD-COSAv2 was collected from CCTV-based monitoring settings, the construction site scenes are static in most cases. Thus, pixel locations could also be an important feature in describing the working context. Therefore, caution must be applied, as the findings might not be generalized to different monitoring environments where a camera can have various poses.

5.3. Modular Design of Object Recognition Pipeline

The target object recognition system is designed to be modular and complementary. Although the modular design, composed of object detection and semantic segmentation models, requires more computations, it can double-check the presence of target objects from two recognition results in the form of segmentation masks and bounding boxes. As demonstrated in Fig. 7, the PPE circled in red is the PPE that was not detected by the object detector and captured by the semantic segmentation. Therefore, the system successfully prevents false negatives of target classes due to recognition failures in object detection or segmentation. For example, as shown in Fig. 7, the object detector failed to detect straps and harnesses, but the segmentation model captured them. As a result, PPE connectivity analysis (2) could identify the missing PPE connectivity from object detection based on segmentation

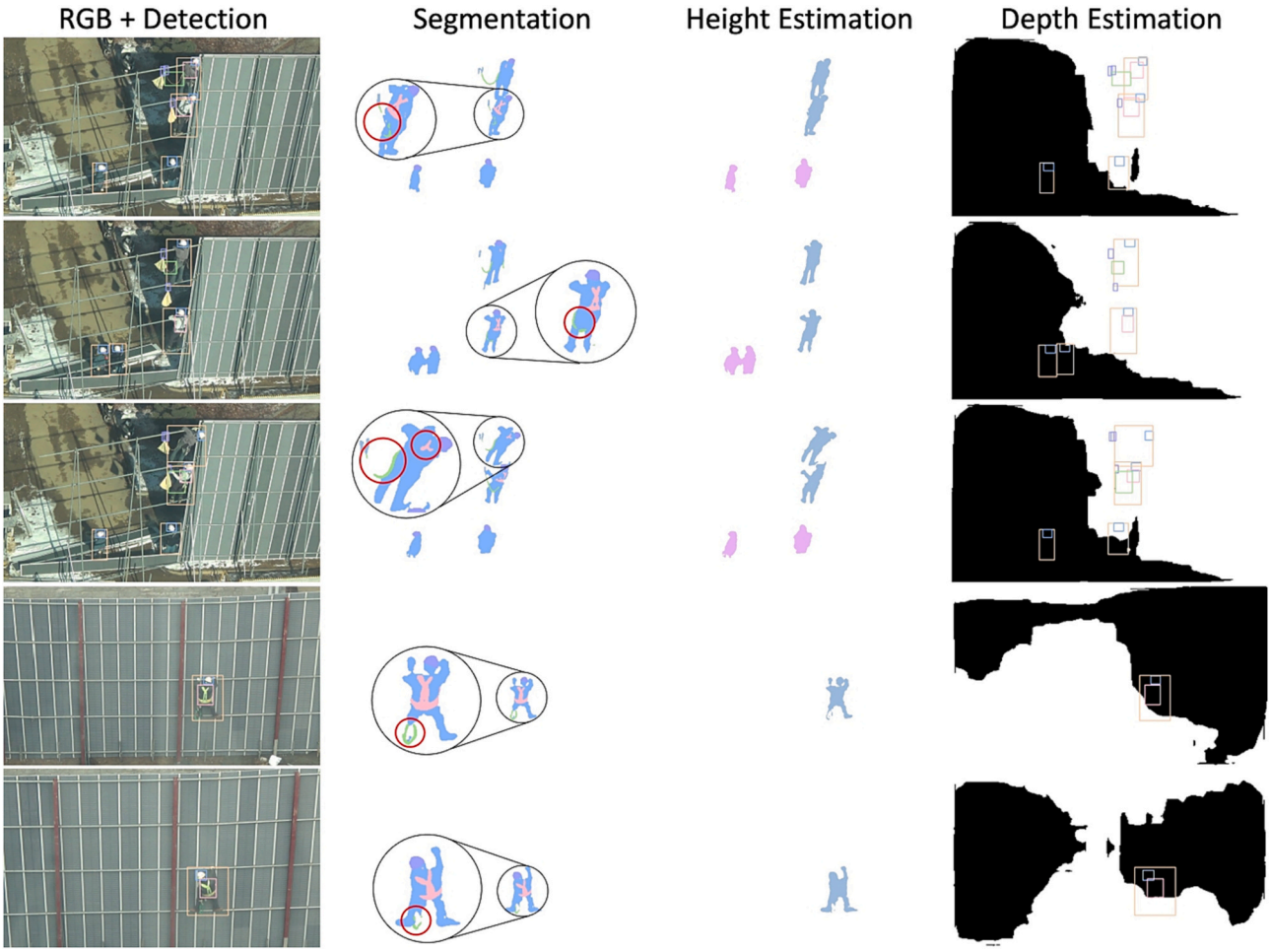


Fig. 7. Visualization results of each recognition models. From left to right, object detection, semantic segmentation, extended segmentation, and depth estimation. The recognition results showed the segmentation masks detect missing PPE (red circles) from object detection. For height estimation using extended semantic segmentation, the workers in green represent at height, and the workers in purple represent on ground. For depth estimation, the bright pixels are closer to the camera, and the dark pixels are further away to the camera. Bounding boxes are overlaid on depth maps for reference. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

results.

Thanks to the modular design of the safety monitoring system—object detection, semantic segmentation, and depth estimation—it allows users to replace each module for better performance. During the experiments, the performance of the safety assessment was improved as the segmentation model was better optimized. As a result, one can adapt the object detection architecture tailored for recognizing small objects as proposed by Zhan et al. [47] for YOLOv5. Alternatively, one can replace the semantic segmentation model with a more efficient architecture as proposed by Zhang et al. [48], which can run on ARMbased mobile devices with competitive performance. Likewise, a better depth estimation model can be trained on construction site-specific training data and be integrated into the proposed system.

6. Conclusion

This study proposed a context-aware safety assessment system to determine the safety status of workers based on the presence of essential PPE for different working contexts, such as working at height or on the ground. To this end, two different approaches—one with the depth estimation model and the other with extended semantic segmentation labels—were investigated to identify the working context of a worker. The experiments demonstrated the potential of both approaches, as they can consider the working context in assessing the safety classification. The second approach based on extended segmentation labels outperformed the first approach based on the depth estimation model, recording Mean F1 of 89.38% and Average Prevision of 86.22% in the determination of the safety status of workers. However, this result would not indicate the better suitability of the second approach for the contextaware safety assessment, as the depth estimation model was employed without fine-tuning to construction site datasets. Rather, the comparable performance, Mean F1 of 80.85% and Average Precision of 78.50%, by the first approach may indicate the potential of developing a robust context-aware safety assessment system. However, preparing training data for depth estimation models could be challenging due to the dynamic nature and the scale of construction sites in far-field monitoring. The findings suggest that using the extended segmentation labels is preferable if depth estimation models are hard to be fine-tuned by construction site datasets. The main contributions of this study can be summarized as follows:

- Proposing a modular system for context-aware worker safety monitoring related to fallaccidents.
- Demonstrating the effectiveness of explicit labels of working context in semantic seg-mentation to classify worker's work status.
- Presenting the potential of a depth estimation model for working context analysis
- Introducing an updated YUD-COSA dataset with new images and full annotations forobject detection, semantic segmentation for far-field monitoring research (the YUDCOSAv2 data set is available for non-commercial and research purposes upon request to the corresponding author.)

6.1. Limitation and future study

There are some limitations to the proposed method. First, the performance of the depth estimation model is insufficient to yield precise safety assessment results, as it was not fine-tuned to construction datasets. Second, the detection and segmentation processes are independent, resulting in an additional processing burden in computation. Third, the recognition performance on small object classes is not high compared to large object classes. Considering these limitations, the following topics would be fruitful areas for further work:

(1) Training and optimizing a depth estimation model with construction site-specific data. (2) Improving detection and segmentation recognition performance with advanced methods such as self-supervised learning-based training. (3) Integrating detection and segmentation pipelines into a single pipeline with multiple prediction heads.

However, some methods may improve the robustness of the proposed context-aware safety monitoring system. The first limitation can be addressed by collecting ground truth depth using LiDAR devices to train depth estimation models. Additionally, Miangoleh et al. [49] proposed an architecture that effectively merges depth results from different resolutions. It presented detailed depth information in crowded scenes such as audiences in sports games. The ability to produce fine details would be helpful in construction scenes to analyze precise working contexts in congested areas. For the second limitation, the methods of self-supervised learning [50,51] and weakly-supervised learning [52] can potentially be used to increase the training data by providing image-level classification labels. Such techniques can be used to minimize the efforts in data preparation, especially for polygon annotations, which are time-consuming and labor-intensive. Lastly, a detection head could be added to a semantic segmentation model to make the inference faster. In addition, a reparameterized module [53] can be applied to the single pipeline segmentation and detection architecture to improve the inference speed.

More issues must be considered from the perspective of safety far-field monitoring in practice. Occlusion by structural components or pose variances of workers could cause false alarms that would frequently distract the supervisors' attention on construction sites. For instance, the position transition of a scaffold worker from one end to the other would create many occlusion cases and body poses. It prevents object detection and semantic segmentation from recognizing PPE consistently. A reasonable approach to tackle this issue could be to assess the worker's safety based on multiple video frames. This way, an alarm will be triggered only when a worker has been classified as unsafe for a specific duration to reduce false alarms. In addition, it is a clear limitation that the proposed method cannot handle the situation where the worker at height is far from the camera but the worker on the ground is near the camera. The assumption made in this study is that workers at height are closer to the camera than workers on the ground, as the camera is installed at an elevated position. If a camera is installed on the ground, the assumption should be changed to recognize the working context.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

The authors appreciate Sai Nikhil Reddy Mandhada who first attempted an initial idea of the proposed work. This research was conducted with the support of the "2022 Yonsei University Future-Leading Research Initiative (No. 2022-22-0102)" and the "National R&D Project for Smart Construction Technology (No. 22SMIP-A158708-03)" funded by the Korea Agency for Infrastructure Technology Advancement under the Ministry of Land, Infrastructure and Transport, and managed by the Korea Expressway Corporation.

References

[1] National Census of Fatal Occupational Injuries in 2018, Technical Report USDL-19-2194, U.S. Bureau of Labor Statistics, 2019. URL: https://www.bls.gov/news.release/archives/cfoi_12172019.pdf. Accessed: 2023/01/06.

[2] W.-C. Chern, T.V. Nguyen, V.K. Asari, H. Kim, Impact of loss functions on semantic segmentation in far-field monitoring, Comp.-Aided Civ. Infrastruct. Eng. (2022), https://doi.org/10.1111/mice.12832.

[3] Census of Fatal Occupational Injuries Summary, Technical Report USDL-21-2145, U.S. Bureau of Labor Statistics, 2021.

[4] A. Krizhevsky, One weird trick for parallelizing convolutional neural networks, CoRR abs/1404.5997 (2014), https://doi.org/10.48550/arXiv.1404.5997. arXiv:1404.5997.

[5] K. Kim, H. Kim, H. Kim, Image-based construction hazard avoidance system using augmented reality in wearable device, Autom. Constr. 83 (2017) 390–403, https://doi.org/10.1016/j. autcon.2017.06.014.

[6] H. Kim, K. Kim, H. Kim, Vision-based object-centric safety assessment using fuzzy inference: monitoring struck-by accidents with moving objects, J. Comput. Civ. Eng. 30 (2016) 04015075, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000562.

[7] J. Seo, R. Starbuck, S. Han, S. Lee, T.J. Armstrong, Motion data-driven biomechanical analysis during construction tasks on sites, J. Comput. Civ. Eng. 29 (2015) B4014005, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000400.

[8] H. Kim, H. Kim, Y.W. Hong, H. Byun, Detecting construction equipment using a region-based fully convolutional network and transfer learning, J. Comput. Civ. Eng. 32 (2018) 04017082, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000731.

[9] Y.-J. Cha, W. Choi, O. Buyukozturk, Deep learning-based crack damage detection using convolutional neural network, Comp.-Aided Civ. Infrastruct. Eng. 32 (2017) 361–378, https://doi.org/10.1111/mice.12263.

[10] X. Yan, H. Zhang, H. Li, Computer vision-based recognition of 3d relationship between construction entities for monitoring struck-by accidents, Comp.-Aided Civ. Infrastruct. Eng. 35 (2020) 1023–1038, https://doi.org/10.1111/mice.12536.

[11] D. Kim, S. Lee, V.R. Kamat, Proximity prediction of mobile objects to prevent contact-driven accidents in co-robotic construction, J. Comput. Civ. Eng. 34 (2020) 04020022, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000899.

[12] M.W. Park, N. Elsafty, Z. Zhu, Hardhat-wearing detection for enhancing on-site safety of construction workers, J. Constr. Eng. Manag. 141 (2015) 04015024, https://doi.org/10.1061/(ASCE)CO.1943-7862.0000974.

[13] Q. Fang, H. Li, X. Luo, C. Li, W. An, A sematic and prior-knowledge-aided monocular localization method for construction-related entities, Comp.-Aided Civ. Infrastruct. Eng. 35 (2020) 979–996, https://doi.org/10.1111/mice.12541.

[14] N. Khan, M.R. Saleem, D. Lee, M.-W. Park, C. Park, Utilizing safety rule correlation for mobile scaffolds monitoring leveraging deep convolution neural networks, Comput. Ind. 129 (2021), 103448, https://doi.org/10.1016/j.compind.2021.103448.

[15] K.-S. Kang, Y.-W. Cho, K.-H. Jin, Y.-B. Kim, H.-G. Ryu, Application of one-stage instance segmentation with weather conditions in surveillance cameras at construction sites, Autom. Constr. 133 (2022), 104034, https://doi.org/10.1016/j.autcon.2021.104034.

[16] B. Xiao, H. Xiao, J. Wang, Y. Chen, Vision-based method for tracking workers by integrating deep learning instance segmentation in off-site construction, Autom. Constr. 136 (2022), 104148, https://doi.org/10.1016/j.autcon.2022.104148.

[17] W. Fang, L. Ding, H. Luo, P.E.D. Love, Falls from heights: a computer vision-based approach for safety harness detection, Autom. Constr. 91 (2018) 53–61, https://doi.org/10.1016/j.autcon.2018.02.018.

[18] W. Fang, B. Zhong, N. Zhao, P.E.D. Love, H. Luo, J. Xue, S. Xu, A deep learning-based approach for mitigating falls from height with computer vision: convolutional neural network, Adv. Eng. Inform. 39 (2019) 170–177, https://doi.org/10.1016/j.aei.2018.12.005.

[19] R. Xiong, P. Tang, Pose guided anchoring for detecting proper use of personal protective equipment, Autom. Constr. 130 (2021), 103828, https://doi.org/10.1016/j.autcon.2021.103828.

[20] J. Shen, X. Xiong, Y. Li, W. He, P. Li, X. Zheng, Detecting safety helmet wearing on construction sites with bounding-box regression and deep transfer learning, Comp.-Aided Civ. Infrastruct. Eng. 36 (2021) 180–196, https://doi.org/10.1111/mice.12579.

[21] S. Chen, K. Demachi, Towards on-site hazards identification of improper use of personal protective equipment using deep learning-based geometric relationships and hierarchical scene graph, Autom. Constr. 125 (2021), 103619.

[22] L. Ma, X. Li, X. Dai, Z. Guan, Y. Lu, A combined detection algorithm for personal protective equipment based on lightweight YOLOv4 model, Wirel. Commun. Mob. Comput. 2022 (2022) 1–11, https://doi.org/10.1155/2022/3574588.

[23] J. Li, X. Zhao, G. Zhou, M. Zhang, Standardized use inspection of workers' personal protective equipment based on deep learning, Saf. Sci. 150 (2022), 105689, https://doi.org/10.1016/j.ssci.2022.105689.

[24] J. Cheng, P.K.-Y. Wong, H. Luo, M. Wang, P.H. Leung, Vision-based monitoring of site safety compliance based on worker re-identification and personal protective equipment classification, Autom. Constr. 139 (2022), 104312, https://doi.org/10.1016/j.autcon.2022.104312.

[25] X. Luo, H. Li, X. Yang, Y. Yu, D. Cao, Capturing and understanding Workers' activities in far-field surveillance videos with deep action recognition and Bayesian nonparametric learning: capturing and understanding workers' activities, Comp.-Aided Civ. Infrastruct. Eng. 34 (2019) 333–351, https://doi.org/10.1111/mice.12419.

[26] X. Yan, H. Zhang, H. Li, Computer vision-based recognition of 3D relationship between construction entities for monitoring struck-by accidents, Comp.-Aided Civ. Infrastruct. Eng. 35 (2020) 1023–1038, https://doi.org/10.1111/mice.12536.

[27] T. Zeng, J. Wang, B. Cui, X. Wang, D. Wang, Y. Zhang, The equipment detection and localization of large-scale construction jobsite by far-field construction surveillance video based on improving YOLOv3 and grey wolf optimizer improving extreme learning machine, Constr. Build. Mater. 291 (2021), 123268, https://doi.org/10.1016/j.conbuildmat.2021.123268.

[28] A. Assadzadeh, M. Arashpour, A. Bab-Hadiashar, T. Ngo, H. Li, Automatic far-field camera calibration for construction scene analysis, Comp.-Aided Civ. Infrastruct. Eng. 36 (2021) 1073–1090, https://doi.org/10.1111/mice.12660.

[29] W. Fang, B. Zhong, N. Zhao, P.E. Love, H. Luo, J. Xue, S. Xu, A deep learning-based approach for mitigating falls from height with computer vision: convolutional neural network, Adv. Eng. Inform. 39 (2019) 170–177, https://doi.org/10.1016/j.aei.2018.12.005.

[30] N.D. Nath, A.H. Behzadan, S.G. Paal, Deep learning for site safety: real-time detection of personal protective equipment, Autom. Constr. 112 (2020), 103085, https://doi.org/10.1016/j.autcon.2020.103085.

[31] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, Tao Xie, J. Fang, imyhxy, K. Michael, A.V. Lorna, D. Montes, J. Nadar, Laughing, tkianai, yxNONG, P. Skalski, Z. Wang, A. Hogan, C. Fati, L. Mammana, AlexWang1900, D, D. Yiwei Patel, F. You, J. Hajek, L. Diaconu, M.T. Minh, ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference, 2022, https://doi.org/10.5281/zenodo.7002879.

[32] C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-ofthe-art for real-time object detectors, arXiv preprint (2022), https://doi.org/10.48550/arXiv.2207.02696 arXiv:2207.02696.

[33] T. Lin, P. Dollár, R.B. Girshick, K. He, B. Hariharan, S.J. Belongie, Feature pyramid networks for object detection, CoRR abs/1612.03144 (2016), https://doi.org/10.1109/CVPR.2017.106. arXiv:1612.03144.

[34] Ibraheem Alhashim, Peter Wonka, High Quality Monocular Depth Estimation via Transfer Learning, 2018 arXiv, 1812.11941, http://arxiv.org/abs/1812.11941.

[35] P.K. Nathan Silberman, Derek Hoiem, R. Fergus, Indoor Segmentation and Support Inference from Rgbd Images, ECCV, in, 2012, https://doi.org/10.1007/978-3-642-33715-4_54.

[36] T. Lin, M. Maire, S.J. Belongie, L.D. Bourdev, R.B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, CoRR abs/1405.0312 (2014), https://doi.org/10.48550/arXiv.1405.0312. arXiv:1405.0312.

[37] G. Jocher, Y. Kwon, J. Veitch-Michaelis, D. Suess Ttayu, F. Baltacı, G. Bianconi, Marc IlyaOvodov, C. Lee, D. Kendall, F. Reveriano Falak, Fu Lin, Google Wiki, J. Nataprawira, J. Hu, Lin Coce, A.I. Luke, Nir Zarrabi, O. Reda, P. Cohen, P. Skalski, Sergio Sanchez, Montes UAM, S. Song, T.M. Shead, Ultralytics/yolov3: v9.6.0 - YOLOv5 v6.0 Release Compatibility Update for YOLOv3, 2021, https://doi.org/10.5281/zenodo.5701405.

[38] H. Zhang, M. Cissé, Y.N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, CoRR abs/1710.09412 (2017), https://doi.org/10.48550/arXiv.1710.09412. arXiv:1710.09412.

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, Comp. Vision Pattern Recog. (2009), https://doi.org/10.1109/CVPR.2009.5206848.

[40] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T. Lin, E.D. Cubuk, Q.V. Le, B. Zoph, Simple copypaste is a strong data augmentation method for instance segmentation, CoRR abs/2012.07177 (2020), https://doi.org/10.48550/arXiv.2012.07177. arXiv:2012.07177.

[41] R. Padilla, W.L. Passos, T.L.B. Dias, S.L. Netto, E.A.B. da Silva, A comparative analysis of object detection metrics with a companion open-source toolkit, Electronics 10 (2021), https://doi.org/10.3390/electronics10030279.

[42] M. Everingham, S.M.A. Eslami, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: a retrospective, Int. J. Comput. Vis. 111 (2015) 98, https://doi.org/10.1007/s11263-014-0733-5.

[43] T.J. Sorensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons, Biol. Skar. 5 (1948) 1–34.. https://cir.nii.ac.jp/crid/1571417124651609984.

[44] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, K. Cho, Augmentation for small object detection, CoRR abs/1902.07296 (2019), https://doi.org/10.48550/arXiv.1902.07296. arXiv:1902.07296.

[45] J.-S. Lim, M. Astrid, H.-J. Yoon, S.-I. Lee, Small object detection using context and attention, in: 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2021, pp. 181–186, https://doi.org/10.1109/ICAIIC51459.2021.9415217.

[46] J. Gasienica-Józkowy, M. Knapik, B. Cyganek, An ensemble deep learning method with optimized weights for drone-based water rescue and surveillance, Integr. Comp.-Aided Eng. 28 (2021) 221–235, https://doi.org/10.3233/ICA-210649.

[47] W. Zhan, C. Sun, M. Wang, J. She, Y. Zhang, Z. Zhang, Y. Sun, An improved yolov5 real-time detection method for small objects captured by uav, Soft Comput.: A Fusion Found. Methodol. Appl. 26 (2022) 361–373, https://doi.org/10.1007/s00500-021-06407-8.

[48] W. Zhang, Z. Huang, G. Luo, T. Chen, X. Wang, W. Liu, G. Yu, C. Shen, Topformer: Token pyramid transformer for mobile semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12083–12093, https://doi.org/10.1109/CVPR52688.2022.01177.

[49] S.M.H. Miangoleh, S. Dille, L. Mai, S. Paris, Y. Aksoy, Boosting Monocular Depth Estimation Models to High-resolution Via Content-Adaptive Multi-resolution Merging, 2021, https://doi.org/10.48550/arXiv.2105.14021.

[50] W.-C. Chern, V. Asari, T. Nguyen, H. Kim, Weakly supervised pseudo label generation for construction vehicle segmentation, in: Proceedings of the 39th International Symposium on Automation and Robotics in Construction, International Association for Automation and Robotics in Construction (IAARC), 2022, pp. 41–46, https://doi.org/10.22260/ISARC2022/0008.

[51] Y. Wang, J. Zhang, M. Kan, S. Shan, X. Chen, Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, 2020, pp. 12272–12281, https://doi.org/10.1109/CVPR42600.2020.01229.

[52] Y.-T. Chang, Q. Wang, W.-C. Hung, R. Piramuthu, Y.-H. Tsai, M.-H. Yang, Weakly-supervised semantic segmentation via sub-category exploration, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8988–8997, https://doi.org/10.1109/CVPR42600.2020.00901.

[53] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, J. Sun, Repvgg: making vgg-style convnets great again, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13728–13737, https://doi.org/10.1109/CVPR46437.2021.01352.