Object Detectors for Construction Resources Using Unmanned Aerial Vehicles

Ivan Mutis, Ph.D., A.M.ASCE[1]; Virat Arun Joshi[2]; and Abhishek Singh[3]

Abstract: Project control operations in construction are mostly executed via direct observations and the manual monitoring of progress and performance of construction tasks on the job site. Project engineers move physically within job-site areas to ensure activities are executed as planned. Such physical displacements are error-prone and ineffective in cost and time, particularly in larger construction zones. It is critical to explore new methods and technologies to effectively assist performance control operations by rapidly capturing data from materials and equipment on the job site. Motivated by the ubiquitous use of unmanned aerial vehicles (UAVs) in construction projects and the maturity of computer-vision-based machine-learning (ML) techniques, this research investigates the challenges of object detection—the process of predicting classes of objects (specified construction materials and equipment)—in real time. The study addresses the challenges of data collection and predictions for remote monitoring in project control activities. It uses these two proven and robust technologies by exploring factors that impact the use of UAV aerial images to design and implement object detectors through an analytical conceptualization and a showcase demonstration. The approach sheds light on the applications of deep-learning techniques to access and rapidly identify and classify resources in real-time. It paves the way to shift from costly and time-consuming job-site walkthroughs that are coupled with manual data processing and input to more automated, streamlined operations. The research found that the critical factor to develop object detectors with acceptable levels of accuracy is collecting aerial images with for adequate scales with high frequencies from different positions of the same construction areas. **DOI: 10.1061/(ASCE)SC.1943-5576.0000598.** *© 2021 American Society of Civil Engineers.*

Author keywords: Unmanned aerial vehicles (UAVs); Project controls; Monitoring; Deep learning; Single-shot detector (SSD).

Introduction

The inefficient utilization of materials and equipment on a job site is a significant factor that contributes to lower productivity and higher costs in a construction project (NAS 2009). Observations and assessments for control operations are error-prone and ineffective in cost and time (Dozzi and AbouRizk 1993), particularly when concurrency of multiple construction tasks occurs in larger construction zones. It is critical to explore new methods and technologies to effectively assist performance control operations by rapidly capturing data from materials and equipment on the job site. The research capitalizes on the capabilities of unmanned aerial vehicle (UAV) technologies—sensing, data collection, and rapid displacements over the construction site's airspace—to investigate the implementation of deep-learning-based methods for effective object detection from images taken from the construction project airspace. Deep learning or deep-structured learning is a class of machine-learning algorithms that employs multiple layers to progressively extract higher-level features from raw input data (Deng 2014).

The research aim is to lay the foundations for research and technology implementations to execute control tasks. By facilitating assessments in real-time of the operational status, location, and movements of a sparsely located trade's construction equipment and materials in the job site, it is anticipated that future implementation of these concepts will streamline data collection by reducing the project engineers' physical displacements and by ensuring that activities are progressing as planned. The study addresses the challenges of using two proven and robust technologies for the real-time data collection and predictions for remote monitoring in project control activities. It is expected that successful application of the approach will support the decision making of project engineers, superintendents, and responsible crew members who are monitoring the progress of construction task and the use of resources on a job site.

There exist significant challenges for the detection of construction resources (object) using a UAV's sensing systems for image collection from the airspace of a construction zone. Detecting and collecting images from the airspace are considerably more problematic methods than those using standard camera devices on the ground in construction zones. Critical drawbacks of UAV images include scale variation (objects from the same category appearing at multiple scales), low resolution [limited visual information that makes it difficult to contrast to cluttered backgrounds (Zhou et al. 2019)], occlusion (objects of interest in a particular view being partially obstructed by other objects), and class imbalance (one object occurs in a view at a much higher rate of frequency when compared with the occurrence of others).

The factors that impact the implementation of object detectors for the monitoring of construction resources can be grouped into two categories:

- Those related to freedom of displacement (including characteristics of motion and position/elevation) when the UAV is used to collect images over the construction area generates arbitrary orientations and angle views in the image with reference to points

[1]Assistant Professor, Illinois Institute of Technology, 3201 South Dearborn St., Alumni Hall, Chicago, IL 60616 (corresponding author). ORCID: https://orcid.org/0000-0003-2707-2701. Email: imutissi@iit.edu

[2]Graduate Student, Dept. of Computer Science, Stuart Building, Illinois Institute of Technology, 3201 South Dearborn St., Alumni Hall, Chicago, IL 60616. ORCID: https://orcid.org/0000-0003-4942-0949. Email: vjoshi1@hawk.iit.edu

[3]Graduate Student, Dept. of Computer Science, Stuart Building, Illinois Institute of Technology, 3201 South Dearborn St., Alumni Hall, Chicago, IL 60616. Email: asingh137@hawk.iit.edu

Note. This manuscript was submitted on November 30, 2020; approved on March 26, 2021; published online on July 27, 2021. Discussion period open until December 27, 2021; separate discussions must be submitted for individual papers. This paper is part of the ***Practice Periodical on Structural Design and Construction***, © ASCE, ISSN 1084-0680.

on the ground. The elevations—the height above ground level (AGL)—create issues related to high densities over the distribution of objects of interest within the image area and major scale variations.
- Those related to the deep-learning approach (e.g., feature extraction and object prediction and localization methods), which impacts performance in terms of accuracy and inferential speed due to aerial images' features as input.

The research presented herein uses a new analytical conceptualization and a showcase demonstration to explore the impact that UAV aerial images have on the design and implement object detectors for construction resource monitoring. The proposed conceptualization is comprised of two parts: spatial resolution and temporal resolution. Spatial resolution refers to the instantaneous field of view (IFOV) within an image of objects and backgrounds on the ground with an adequate scale at a given instant in time. Temporal resolution refers to the frequency for obtaining imagery of the same particular area of a construction project.

The showcase demonstration is an implementation of a deep-learning object detector that uses aerial images from a construction project. The aim is to incorporate strategies that address spatial and temporal resolutions to demonstrate conditions that highlight performance difficulties related to implementation of deep-learning object detectors for construction resource monitoring (e.g., features of the existing image data set limit their use for training deep-learning models due to differences in aerial viewing angles, a condition that complicates training of the models because they need large amounts of training data to make more accurate predictions).

The research was developed following a five-phase approach: (1) definition and selection of the main factors that impact implementation of object detectors; (2) selection of a machine-learning object-detector algorithm; (3) data collection and postprocessing steps; (4) training the machine-learning convolutional neural network; and (5) evaluation of the results and analysis. In sum, the research goals were to explore and develop an analytical conceptualization and a showcase demonstration to detect construction resources within aerial images in real time using the latest machine-based algorithms.

It is expected that the exploration of the challenges identified herein will inform and guide strategies for object detectors' implementations, thereby benefitting construction stakeholders by enabling efficient data collection, processing, and interpretation in construction management tasks. The research will serve as a baseline for UAV technology use in construction planning. It will effectively build considerations for preventive and corrective actions in activity assessments—keeping accurate, timely information to support status updates and progress measurements. The rapid detection and tracking of resources require efficient implementations of project control methods to facilitate decision making, thereby assisting the workflows of monitoring tasks and improving the overall understanding of a project's status.

The presented approach offers four critical contributions:

1. Conceptually demonstrating methods for choice, training, datasource, and implementation of deep-learning-based object detectors for UAV low-altitude images from construction projects as input.
2. Studying the state-of-the-art object-detection approaches to better fit their implementations for construction projects' particular conditions by exploring and developing an analytical conceptualization and a showcase demonstration that detects construction resources within aerial images in real time.
3. Offering a guide for researchers on the challenges and requirements for aerial image data sets of construction resources (objects) for quality performance of deep-learning approaches.
4. Advancing knowledge on UAV capabilities to be used as an intervention to reduce human effort on construction-monitoring tasks and informing job-site activities by combining the power of UAVs with deep learning to subsequently detect construction resources based on their unique signatures.

Following a "Background" section explaining the uses of UAV in construction, this paper details the methodology, the showcase demonstration of implementation, and the results. The discussion and conclusion examine and summarize findings and lessons learned during the investigation.

Background

The first part of this section focuses on UAV advancement in construction. Implementations of object detectors using computer-vision techniques with standard image collection methods for construction project applications follow.

The UAV is ubiquitous technology that has proven to be safe and capable of providing information on construction engineering-related tasks faster and at lower costs (Albeaino et al. 2019). UAVs enable the simultaneous visualization of in situ construction resources, processes, and management of activities as they unfold over time, which in turn supports solutions to construction engineering tasks, such as following-up on progress in specific locations within project controls and construction safety tasks (Gheisari and Esmaeili 2019). UAVs as an intervention replace personnel operations in the job site that are difficult, costly, and inefficient to perform due to a safety hazard and high level of effort. It collects information to engage a unique bird's-eye view of reality experiences—a perspective that personnel in the job site would otherwise not be able to observe. The UAV information affords experiential observations for awareness, facilitating planning and goal-prioritizing activities essential for the success of project-site activities.

By air-capturing the physical construction site environment using a moving ultrahigh-definition camera and sensors (Engel et al. 2014), and by reducing the limitations of onsite fixed-location video devices deployed for information collection in locations such as on occlusions and congested zones, UAV use has significantly expanded in the last few years (Irizarry and Costa 2016; Zhou et al. 2018). Demonstrated applications include inspections, structural and health monitoring (Duque et al. 2018), transportation (Greenwood et al. 2019; Kwon et al. 2017), project control (Asadi et al. 2020), disaster management, preservation (Bakirman et al. 2020), energy efficiency in the built environment (Ficapal and Mutis 2019), and construction safety (Kim et al. 2019; Liu et al. 2019). However, most of these applications follow a postprocessing approach whereby users process and interpret information using UAV images and videos as input after its job-site collection, requiring a high level of human input for processing and interpretation. New research approaches should reduce the users' efforts to collect and process information from UAVs for decision making based on accurate detection of construction resources available from UAV images in real time.

Due to the recent availability of curated data sets, the researchers have deployed high-capacity supervised or discriminative deep-learning approaches for many construction management applications, ranging from object category recognition to object tracking. For example, large-scale data sets such as ImageNet (Deng et al. 2009) and GoogleNet (Szegedy et al. 2015) have facilitated deep-learning techniques in multiple applications in other domains with high levels of accuracy. Research in construction management using deep-learning methods is categorized into major clusters (Seo et al. 2015): object detection, object tracking, and action recognition.

Each of these clusters aims to obtain specific output from images, such as data of movement (e.g., tracked construction resources in the physical space) and activity markers [e.g., cycle times on productivity (Sherafat et al. 2020)].

For each cluster, there has been significant research in the last two decades. Employed methods range from histograms of oriented gradients (HOG) (Azar et al. 2013; Golparvar-Fard et al. 2013), latent support vector machines (SVM) (Zhu et al. 2017), optical flow (Kim et al. 2017) and stereoscopic vision [three-dimensional (3D)] for object detection (Brilakis et al. 2011), to deep-learning techniques using convolution neural network (CNN) (Fang et al. 2018) and long short-term memory (LSTM) (Slaton et al. 2020). Each method has a set of advantages and disadvantages, and the key has been to select the correct method for the given application. New contributions to the clusters are growing. The most significant existing implementations for object detection and activity recognition in construction specific to deep-learning techniques are presented in Table 1.

Using kinematic, image/video, and sound-based methods (Sherafat et al. 2020), there have been numerous approaches deployed for site monitoring systems that are less dependent on humans. Such systems aim to help project stakeholders control activities from activity recognition to activity tracking and performance monitoring using corrective actions and performance data. The collected information is analyzed for productivity (Roberts and Golparvar-Fard 2019), safety (Fang et al. 2020; Kim et al. 2019, 2017), and quality control and decision-making tasks (Luo et al. 2018) to prevent delays and enable safe and hazard-free environments.

Methodology

Fig. 1 describes the general steps of the method. The first two phases focus on understanding the problem and selecting the deep-learning-based approaches to achieve the research work. The last three phases contribute to the development, implementation, and evaluation experimentations of the chosen approach. The authors conceptualized continuous and iterative feedback in the model to achieve improved performance and results for detection and recognition, thereby improving the indicators for further decision making.

Phase 1: Aerial Image Analysis and Selection of Deep-Learning Object-Detector Methods

Object detection refers to capturing and detecting instances of a specific class in images from videos. Class means the type of construction equipment and materials of interest, such as bulldozer, crane, ladder, rebar, and formwork. There are two deep-learning approaches for object detection based on two- and one-phase models. For the two-phase model, the first step is object localization in the image by implementing algorithms to generate regions with a series of candidate frames (bounding boxes). The second step is a classification for detecting the class of objects in the proposed region by extracting features from each bounding box to determine if and which objects are present in the proposals using classifiers [e.g., region-based convolutional neural networks (R-CNNs)] (Girshick et al. 2014). CNNs are feed-forward networks in which multiple-input, output, and hidden layers form a convolutional layer. For the one-phase model, the algorithm uses regressions instead of a proposed region generation phase, thereby using one step to directly identify features maps with different resolutions to perform object localization and classification [e.g., YOLOv3 (Redmon and Farhadi 2018), RetinaNet (Lin et al. 2020), and single-shot detection (Liu et al. 2016)].

The use of UAV aerial images collected from the construction zone airspace involves factors for the selection and implementation of the object-detector approach. The factors can be viewed under a two-group category. The first group is associated with the UAV's motion and position (including elevation) conditions in the airspace relative to reference points on the ground during the image collection. The second group is related to the deep-learning algorithm performance factors when using UAV images that impact the selection and implementation of an object detector.

The researchers analyzed the motion and position conditions under the spatial and temporal resolutions to explore the first category. The free movement for displacement (motion and positions) over the construction area generates arbitrary orientations and angle views of the objects with reference to a point on the ground. Fig. 2 shows the shifting orientations and angle views during UAV data collection. The elevations—AGL—generate issues related to high densities over the distribution of objects of interest within the image-area and major scale variations in the image.

Low—or lack of optimal—spatial resolution indicates that the UAV image-sensing system (digital camera sensor and lens) procures images with high densities of objects. The IFOV is from positions in the airspace with a minimum of 40 m AGL for safe UAV operations, which in effect is high density. IFOV from positions in the airspace are not commonly found in the true examples of data sets used for CNN networks' training. Low resolutions also occur when objects in the images have sizes and aspect ratios—small or wide—with unique scales and views. IFOV of objects with very low scale values appears within a small proportion to the image's full size. The objects would likely appear bunched together with partial occlusions, adding hurdles to the substantial limitations to the existing amount of annotated data currently available in the true ground example. The issue is more significant when objects of interest are placed close to each other, usually the case at construction sites. Low scale values affect accuracy for predictions.

Spatial resolutions have a significant impact because the existing data set of images of objects (classes) used in construction projects is limited in sample object sizes and shapes from aerial viewing angles. Object-detection applications need large amounts of training images with representative sample cases of IFOV with optimal spatial resolutions—objects and backgrounds on the ground with an adequate scale at a given instant in time—to make predictions. Another effect of low spatial resolutions is class imbalances—the conditions of having a few main objects and extensive background in the images. Class imbalance impacts accuracy due to high values from the CNN loss functions of well-classified examples with high probabilities (e.g., the IFOV in the image should not present deformations, physical aggregation with multiple classes, and partial occlusions).

The UAV's physical moving features impact temporal resolutions—the frequency of obtaining imagery of the same construction project area. UAVs' image-sensing systems rapidly shift positions relative to the ground as opposed to surveillance sensing systems, where cameras are mostly fixed and maintain the same position covering the same area. The shifting positions imply having rapid changes of viewing angles of the covered areas during flight operations. Shifting positions generate continuous changes of locus of the UAV sensing system and IFOV. The generation of arbitrary IFOV from moving position impact the outcome of the deep-learning algorithm in terms of their accuracy and speed. Any of these issues' materialization impacts performance accuracy and inference time for deep-learning detectors' implementation. The algorithm requires short inference time to meet the real-time demands of rapidly changing images from video to process and accurately detect objects of interest.

Table 1. Summary of most salient machine-learning implementations for the detection of construction resources and construction management tasks

Objective	Architecture	Classes	Advantages	Limitations	Metrics	References
Detecting construction equipment	DetectNet (deep neural network)	Formwork	Applied DetectNet architecture to recognize and locate formwork in UAV videos.	Detects only a single class with a map of 44%—poor performance for inputs with multiple classes.	Mean average precision (mAP)	Jahr et al. (2018)
Generating metadata tags for construction images	VGG-16 (CNN)	Building, equipment, workers	Applied VGG-16 architecture to generate metadata tags in constructions images used for object detection and activity recognition. Uses data set to generate labels for single and multilabel classification.	The model can perform the only classification and cannot identify object boundaries within an image for detection. Performance metrics are comparatively lower for a classification task.	Accuracy, precision, recall	Nath et al. (2019)
Automated vision tracking of construction equipment	Coordinate triangulation using multiple cameras	Equipment, workers, and materials	Detects the spatial location of project entities like equipment and materials without sensors.	Cannot display recorded information to the interested user. The absence of sensor-based techniques affects the accuracy of coordinate tracking.	Absolute error, average error	Brilakis et al. (2011)
Annotation of construction footage using object detection	Histogram of oriented gradient (HOG) and Bayesian networks	Bulldozer, excavator, dump truck, grader, roller	Detects five equipment classes using HOG classifiers. Annotates classified videos using Bayesian networks to calculate the most probable action of the equipment class.	Diversity in data can adversely affect performance. Conditional probability deteriorates performance when classes are similar.	Precision–recall curves	Rezazadeh Azar (2017)
Real-time object detection on construction sites	Sliding-window detector, random forests	Steer loaders, backhoes	Uses sliding windows to detect objects in construction site videos.	Sliding windows are inefficient for high-resolution videos. The approach uses HOG detectors for speed but, in turn, has a low detection rate of 81.68%.	Miss rate, false-positive per window	Memarzadeh et al. (2012)
Surveillance of power grids using real-time object detection	Faster R-CNN	Excavator, bulldozer	Creates a fine-tuned data set of engineering vehicles intruding in power grids. Utilizers modified R-CNN to detect objects for surveillance purposes.	The use of Faster R-CNN prohibits use in real time because it can process fewer frames per second compared with algorithms such as single-shot detectors.	Mean average precision (MAPmAP)	Xiang et al. (2018)
Construction site monitoring	Single-shot detectors (SSD)	Construction equipment detected in surveillance footage	Uses single-shot detector and affinity propagation clustering to detect seven classes of equipment in construction site videos.	Approach results in a maximum map of 74%, which is very low for the given task. Greedy nonmax suppression restricts SSD performance.	Mean average precision (MAPmAP)	Thakar et al. (2018)
Detecting construction equipment	R-FCN	Excavator, dump truck, loader, road roller, mixer	Using a region-based convolution network to detect multiple objects on a construction site.	It cannot be used in real-time detection because FPS is low for the model.	Precision–recall curves	Jinwoo Kim (2015)

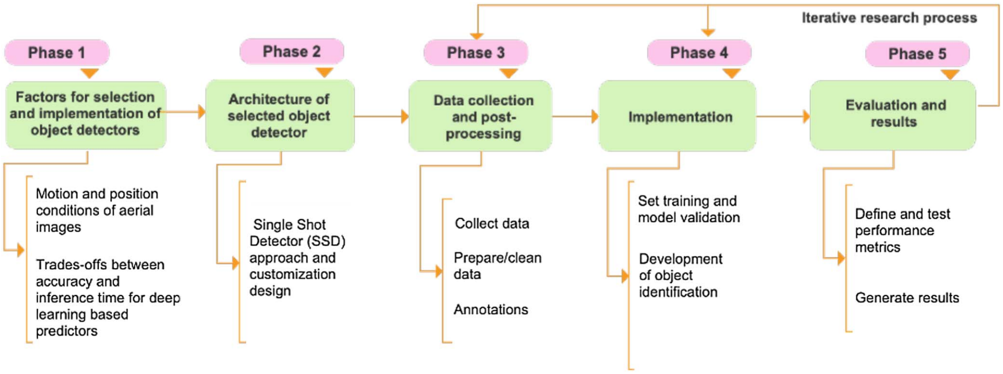

Iterative research process
Phase 1
Factors for selection and implementation of object detectors
Motion and position conditions of aerial images
Trades-offs between accuracy and inference time for deep learning based predictors
Phase 2
Architecture of selected object detector
Single Shot Detector (SSD) approach and customization design
Phase 3
Data collection and post-processing
Collect data
Prepare/clean data
Annotations
Phase 4
Implementation
Set training and model validation
Development of object identification
Phase 5
Evaluation and results
Define and test performance metrics
Generate results

Fig. 1. Research methodology.

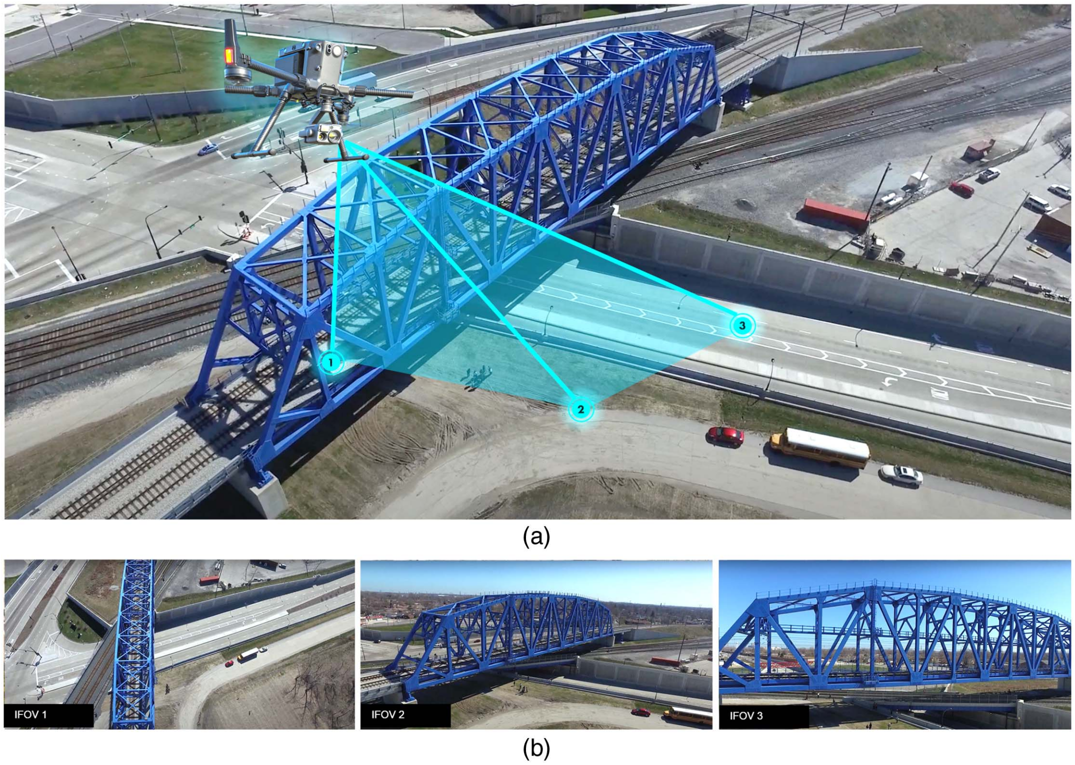


Fig. 2. Spatial and temporal resolutions: (a) shifting UAV motion and position in the construction zone airspace; and (b) IFOVs with different angle viewing, object sample sizes, and densities. (Image of UAV device and UAV images by authors.)

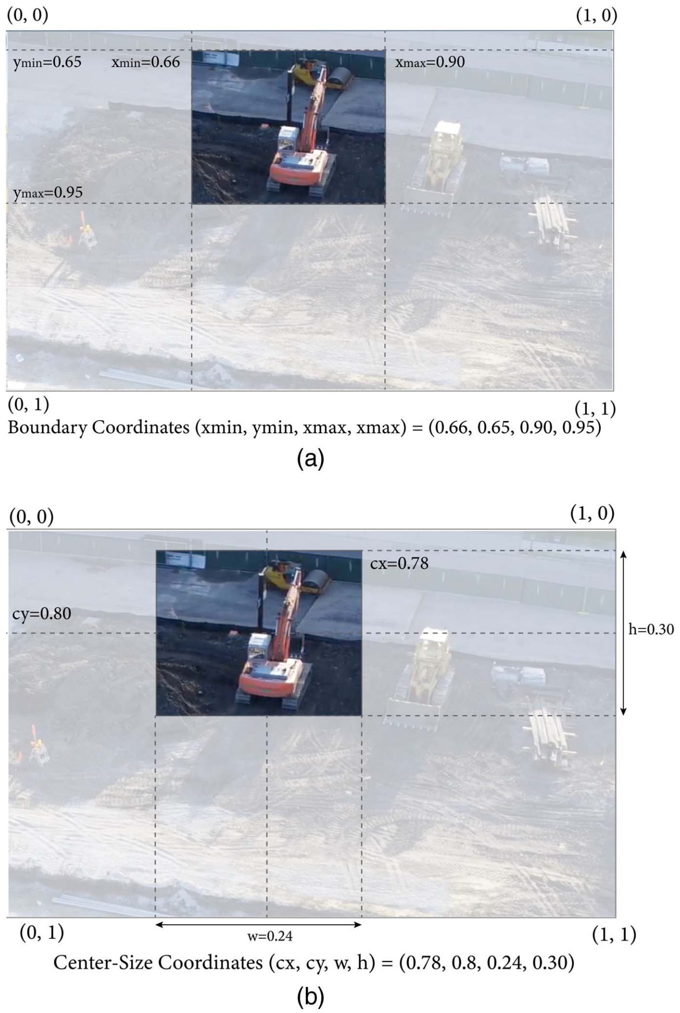
The selection and design custom of the CNN models must be made in the context of the problem. Some deep-learning algorithms have better accuracy but slower processing times and vice versa, i.e., lower accuracy but faster processing times. The speed–accuracy trade-off is a critical factor for the choice of deep network for object detection using UAV in applications. UAV images' spatial and temporal resolutions of the current for object predictors' implementations are significant limitations compared with standard nonaerial images. The CNN model's selection and its design customization hinge on the conditions of the spatial and temporal resolutions by adjusting balances between accuracy (how well it classifies and localizes objects of interest) and processing time (how long it takes to process predictions of classes).

Because real-time detection with classification and localization from UAVs' sensing system is the aim, the authors chose a one-phase CNN model to achieve an acceptable detection accuracy and shorter inference time to meet the real-time requirement. One-phase models have low inference time and can predict small object sizes within lower-resolution images in complex backgrounds. The two-phase model (R-CNN) was most suitable for non-real-time and postflight processing and analysis applications with higher prediction score results. The two-phase model required more time for inference (processing) the predictions and presented a poor performance for object localization on images when the object is relatively small—low spatial resolutions for the aerial image case—thereby adding limitations for UAVs image usage in real-time. For example, the maximum reported prediction speed for deep-learning models (benchmarking experiments), measured in frames per second (FPS) for Faster R-CNN (VGG 16) was 7 (Ren et al. 2015), SSD300 was 46 (Liu et al. 2016), Fast R-CNN was 0.5 (Girshick 2015), YOLOv3 was 22 (Redmon and Farhadi 2018), and SSD512 was 19 (Liu et al. 2016). The reported times offer an approximation of how they perform. The experiments were executed in disparate settings by different authors, which makes exact one-to-one not viable comparisons.

After small-scale implementations of other one-stage object detectors [e.g., Yolo3 (Redmon and Farhadi 2018) and SqueezeDet (Wu et al. 2017)], the authors concluded that the single shot detector (SSD) is an algorithm that best meets the conditions of spatial and temporal resolution for object detection from UAV in real time, in the context of construction projects. The authors concluded that SSD was a more suitable approach for detection from UAVs. The goal was to achieve high accuracy and precision values with relevant modifications to SSD architecture. An in-depth discussion of the SSD architecture is explained in Phase 2, as follows.

Phase 2: Architecture

SSD uses only a single shot to detect multiple objects within the image, unlike techniques that use region-proposal methods. SSD utilizes predetermined bounding boxes known as priors. Priors represent the coordinates of the detected object's location in the image, called boundary coordinates. They can be expressed in two ways (Fig. 3). The first specifies the normalized minimum and maximum of the *x*- and *y*-coordinates of the boxes in the Euclidian space. The second specifies the center coordinates along with the box's width and height in normalized form. The research presented herein features an implementation that utilizes both representation methods depending on the task being performed.



Boundary Coordinates (xmin, ymin, xmax, xmax) = (0.66, 0.65, 0.90, 0.95)

Center-Size Coordinates (cx, cy, w, h) = (0.78, 0.8, 0.24, 0.30)

Fig. 3. Two different methods of representing the bounding box for the objects: (a) boundary coordinates; and (b) center coordinates. (Images by authors.)

As a standard and popular algorithm, SSD has been used in research in construction domain applications primarily to monitor construction activities. An overview and analysis of the existing SSD approach to applications in construction are summarized in Table 2.

Single-Shot Detector Approach

The SSD CNN is split into three components. The first component features the base convolutions that generate feature maps based on an existing image classification architecture. The second is composed of the auxiliary convolutions that make high-level feature maps, followed by a convolution layer for predictions. This third component is a prediction layer that detects and locates the objects that match previously specified classes. Depending on the number of layers, the SSD can be referred to as either SSD300 or SSD512. The trailing number indicates the image size, with SSD300 identified as a suitable choice for the data in this research. Utilizing SSD300 in place of SSD512 is contrary to what could be a common expectation because SSD512 has a higher accuracy metric due to its use of higher-resolution images. The major drawback of SSD512 is that it requires more priors, which makes a model computationally expensive and challenging to deploy in scenarios where processing power is limited. Further, such a model would need a higher inference time (e.g., SSD300 and SSD512 can obtain 46 and 19 frames per second, respectively), rendering it, at times, impossible to feed the model with high-quality input, thus eventually affecting prediction accuracy.

The authors used existing image classification architectures for a class of construction resources because the current architectures have been adapted and improved through years of research to capture elementary features from a given image. Using existing architecture to build a new system is more efficient. For the convolution base (backbone), the system utilized a Visual Geometry Group (VGG)-16 architecture that was pretrained on the ImageNet Large-Scale Visual Recognition Competition (ILSVRC) classification task (Deng et al. 2009). The kernel size in the fifth pooling layer and change of the convolution stride facilitated calculations when training. The classifications used fully connected layers and were further removed for adaptability in the authors' implementation. Fig. 4 shows an example of the processed and fully connected layers fc6 and fc7 fitted into the convolutional layers as conv6 and conv7.

Phase 3: Data Collection and Postprocessing

There were two sources of data for the implementation, based on its source. The first was publicly available repositories of scraped single-class and multiclass images from the internet. The selection criteria were the spatial resolutions to maintain consistency and similarity in terms of objects' sizes and aspect ratios within the image as they were taken from UAV sensing systems. The second was a real project video from UAV sensing systems. This source helps maintain consistency based on the temporal resolution of the UAV images. Postprocessing was performed, selecting frames at intervals in image selection for footage from construction sites. Considerations for data collection and postprocessing activities follow.

A proper viewing angle and IFOV of the ground area with adequate scale were necessary for each image for adequate spatial and temporal resolutions. Viewing angle included changing viewing angles of covered areas in the ground because it implied a moving

Table 2. Summary of the most salient implementations of deep-learning approaches for object detection

Problem category	Description	Limitations	Performance	References
SSD for site monitoring	Replaces greedy nonmax suppression with affinity propagation clustering to improve performance for smaller objects. Analyzes performance of different SSD versions with existing neural networks.	Bounding boxes (representation of predictions) tend to be smaller than the actual object, thus making it difficult for use on classes such as cranes and ladders.	Improves mean average precision (mAP) of SSD by 3.77% on custom data sets	Jahr et al. (2018)
Object-detection models on embedded hardware for construction sites	Implements SSD and mobilenet in tandem on web-scraped and subset of ImageNet data set data to detect six classes of construction equipment. Uses edge computing to deploy the developed model on embedded hardware.	Merge bounding boxes of different classes in close vicinity. It can only detect one class per image.	Achieves an interpolated average precision (IAP) of 91.20% for all classes	Arabi et al. (2020)
Detect abnormal objects on railway tracks	Uses modified SSD and R-CNN to detect objects occluding railway tracks using a technique called region cutting.	The approach does not classify construction equipment. The size of data required for each class when training the model is considerably larger.	Achieves mAP of 88.9% for 20 classes.	Li et al. (2020)

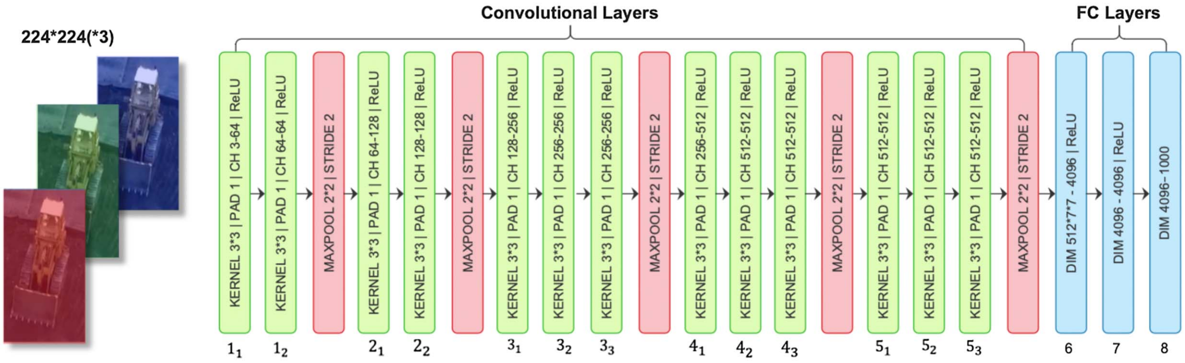

224*224(*3)
Convolutional Layers
FC Layers
KERNEL 3*3 | PAD 1 | CH 3-64 | ReLU
KERNEL 3*3 | PAD 1 | CH 64-64 | ReLU
MAXPOOL 2*2 | STRIDE 2
KERNEL 3*3 | PAD 1 | CH 64-128 | ReLU
KERNEL 3*3 | PAD 1 | CH 128-128 | ReLU
MAXPOOL 2*2 | STRIDE 2
KERNEL 3*3 | PAD 1 | CH 128-256 | ReLU
KERNEL 3*3 | PAD 1 | CH 256-256 | ReLU
KERNEL 3*3 | PAD 1 | CH 256-256 | ReLU
MAXPOOL 2*2 | STRIDE 2
KERNEL 3*3 | PAD 1 | CH 256-512 | ReLU
KERNEL 3*3 | PAD 1 | CH 512-512 | ReLU
KERNEL 3*3 | PAD 1 | CH 512-512 | ReLU
MAXPOOL 2*2 | STRIDE 2
KERNEL 3*3 | PAD 1 | CH 512-512 | ReLU
KERNEL 3*3 | PAD 1 | CH 512-512 | ReLU
KERNEL 3*3 | PAD 1 | CH 512-512 | ReLU
MAXPOOL 2*2 | STRIDE 2
DIM 512*7*7 - 4096 | ReLU
DIM 4096 - 4096 | ReLU
DIM 4096–1000
1_1
1_2
2_1
2_2
3_1
3_2
3_3
4_1
4_2
4_3
5_1
5_2
5_3
6
7
8

Fig. 4. VGG-16-modified architecture for the adaptability of the implemented SSD. (Bulldozer images by authors.)

condition of UAV sensing systems during flight operations. The limit number of images from data sets that are captured in a top-viewed angle. The images in the data set present objects at a lower viewing angle, reducing effectiveness in training the model because the features at lower angles are not transferable for aerial views. The goal was to procure images with densities of objects that best contribute to the model's learning.

Images in data sets have objects with large relative dimensions, adding limitations for their use in object detectors with UAV images as input. Objects within UAV images tend to appear at small scales due to UAV elevations—a spatial resolution problem. The researchers explored the option of using existing low-altitude aerial image data sets [e.g., UAV mosaicking and change detection (UMCD) data set (Avola et al. 2020) and Okutama (Barekatain et al. 2017)]. Their utility was minimal because they specialized for other functions (e.g., event recognition in surveillance, human action detection, humans' gestures, or plastic bottle localization). Some data sets had a general-purpose function, but the existing content did not intersect with objects used in construction (equipment and materials), limiting their use for UAV in construction projects.

The researchers used 1,200 images collected using web scraping and converting interval-separated image frames from existing UAV footage, then augmented. Not all web-scraped images had similar dimensions. Often, the target classes were occluded by construction workers or other objects—these issues needed to be addressed using suitable transformation methods. The augmentation was performed with a 50% probability, which means that only half the images were augmented at random. The resulting data consisted of 1,800 images balanced in their class distribution and loaded into the model using PyTorch's inbuilt DataLoader method. The images were manually annotated using open-source software called LblImg version 1.8.5 (Tzutalin 2015).

The collected data were augmented using a variety of data transformations (Fig. 5). The first was photometric distortion, considering variations in hue, brightness, saturation, and noise. The second was a geometric distortion. The researchers zoomed out the images to assist the model in learning smaller objects with ease and zoomed in for larger objects. They were horizontally flipped and normalized using their mean and standard deviation of SSD's pretrained weights. Finally, the researchers resized the images to a dimension of 300 × 300, which is the required input size of SSD300.

The transformed data were then ready to be used as input to the model. The authors annotated the data manually and uniformly categorized it into five primary classes of construction resources: bulldozer, excavator, ladders, rebar, and formwork. Each class comprises an object or piece of equipment commonly found at construction sites. The presence of these classes in a given image can be used to make inferences on the type of activity being executed on site. The developed approach is flexible in making it possible to incorporate additional classes and associated data as efficiently as possible for further training and adjustment of the model.

Phase 4: Implementation

The implemented approach used SSD's three different sets of convolutions (base, auxiliary, and prediction) to find and classify objects in an image, i.e., object detection. In the base convolution, the researchers designed a neural net similar to VGG-16 by modifying the last two fully connected layers, replacing them with convolutional layers. The researchers employed a pretrained weight from the ImageNet data set. The output from the base convolutions were two feature maps. These feature maps were used as an input to the auxiliary convolutional layers. At the end of this stage, the neural network completed its learning phase to proceed to the generation of the bounding boxes and predictions.

The previous step generated seven feature maps that fed the prediction convolutions, which subsequently returned the location of bounding boxes and class labels in each feature map. The layer utilized priors boxes—anchor boxes—to generate the bounding boxes around the desired class. The model's object-detection phase then used multibox detection to locate class objects in the images as input. Because there were many detected boxes, the most relevant boxes were filtered based on Jaccard similarity measures.

The researchers used transfer learning—the process of employing pretrained (CNN) as the basis for predicting a new given data set—to build the model. The researchers froze the pretrained weights and later unfroze them. The unfrozen part helped the model learn the data once the basic features were extracted using the pretrained weights. These weights—from the VGG-16 layer's base convolutions—are frozen because they have already been trained for object classification. The weights in the auxiliary and prediction convolutions were replaced by training the initializations, which were updated after training the model. One thousand images scraped from the web and uniformly distributed across all five classes were used as input to train the model. An important hyperparameter used for training was the stochastic gradient descent (SGD) optimizer.

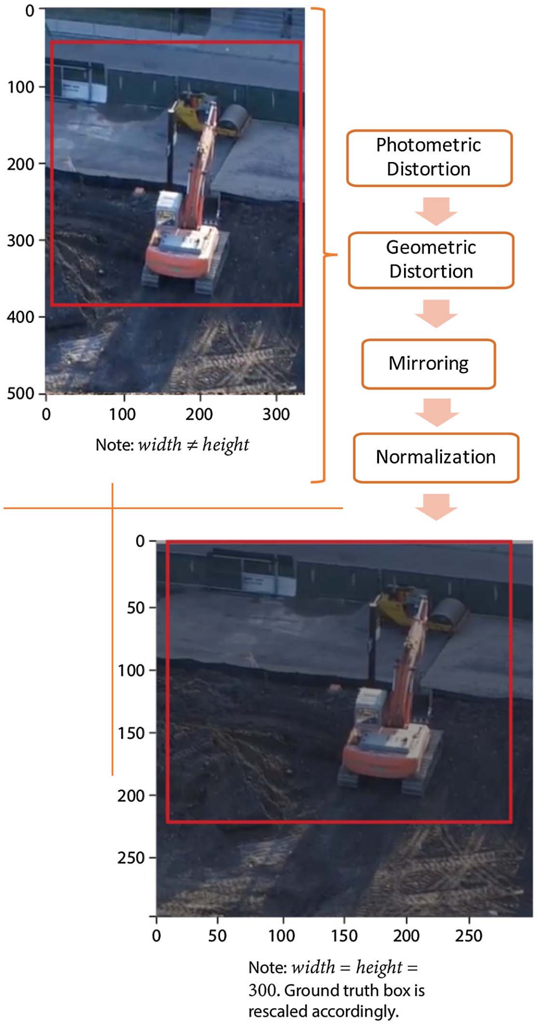


Fig. 5. Syntax mapping for activity registration. (Images by authors.)

To evaluate the model, the researchers used multibox loss. Multibox loss is a combination of the localization and confidence losses of the object detected in the images. A parameter alpha balanced the contribution of the localization loss, thereby helping the predictions approach the ground truth. The trained model using the previously mentioned evaluation measures was used to output the predictions.

The resulting data were trained using a model that implemented three approaches for training the deep neural network. The hyperparameters used in the model are as close to the ideal values as possible. Thus, instead of the initial tuning of the hyperparameters, this research approach focuses on how an ideal learning rate for the object-detection module was found. Once the ideal learning rate was achieved, the parameters were fine-tuned to get more accurate results. The three approaches used to find the ideal learning rate are (1) a constant-learning-rate method, (2) learning-rate annealing, and (3) cyclic learning rates. It is important to note that all three approaches had a different learning rate. The three different approaches enable comparison and verification of the methods used within the case example (use case) implementation.

The model was trained on a Nvidia GeForce TX 1070 GPU for 120 epochs using mini batches of size 32 for all three approaches, and the model was built and run using the PyTorch library. The first approach set a constant learning rate for each of the layers in the model (base, auxiliary, and convolution) of 0.0008 for the base layer, 0.001 for the extra layer, and 0.004 for the prediction layers. These values were decided on a trial-and-error basis.

The second approach set the learning-rate annealing values to 0.0008, 0.001, and 0.004 initially. After every 30 epochs, the value was dropped by a factor of 10. This approach was initially selected because it required a very high learning rate to identify the range in which the optimum minima would be found. The learning rate was gradually reduced to enable the rates to stay in the optimum minima vicinity.

The third approach used the cyclic learning rate, a strategy for balancing between a given range of minimum and maximum values. The range of values was set at 0.0002, 0.0008, and 0.006 as the minimum for each layer in the model, and 0.001, 0.008, and 0.001 as the maximum. To validate and monitor performance, the prediction, mean average precision (mAP), and loss values were continuously calculated during the training process using the validation data set. The comparisons helped ensure that the detector maintained the overall performance during training without overfitting or losing its generalization. The loss can be visualized for the learning-rate annealing approach in Fig. 6.

Phase 5: Evaluations and Results

The authors developed a case example to demonstrate the approach that included implementation and experimentation tasks. It uses UAV footage collected from several missions over a construction site. Each video belonged to a UAV mission, and each UAV mission was composed of a set of flights over a job site during a specific period in the construction phase. The demonstration analyses the output of the retrieved cluster of UAV videos critical to the job-site personnel. It focuses on detecting classes of construction resources (equipment and tools) that are typically found in construction task projects.

The accuracy of the object-detection modules was evaluated in terms of the mean average precision [mAP, as defined by

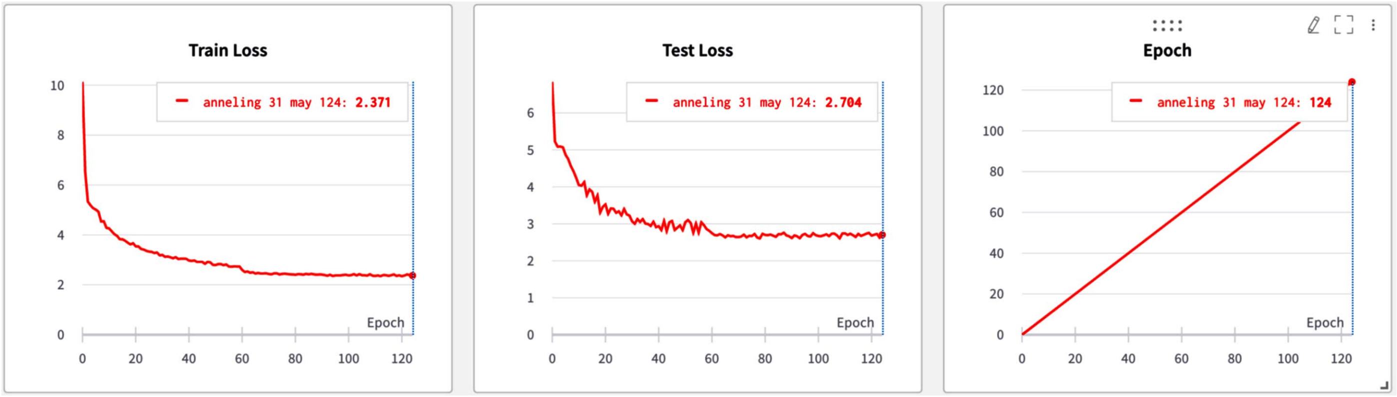


Fig. 6. Modeling: train and test loss: learning-rate annealing.

Everingham et al. (2009)]. The mAP, in simple terms, is the area under the precision-recall curve. The evaluation was performed using the raw predictions, then parsing and comparing them with the ground-truth values. To standardize evaluation, the nonmax suppression was set to a threshold of 0.45. Although hard-negative mining was used to select the images where the background was the most difficult to detect, the discarded images were considered in the evaluation data set to avoid having false positives. The results showed that the model performs well, given the small size of the data set.

The optimum learning rate was achieved by calculating the training and test loss and adjusting the rate along with the number of epochs required for training the data (one epoch corresponds to a full training cycle of the network without repetition on each data value). The implementation considered different approaches to arrive at the optimal learning rate, such as using a cyclic learning rate and learning-rate annealing. The hyperparameters such as weight decay, image resolution, and momentum were tweaked to account for the number of priors used. As a result, the authors achieved a comparable performance using less than half the data used in existing implementations (Jahr et al. 2018), thereby making the model in this research more effective. The results demonstrated that using the cyclic-learning-rate approach results in the highest mAP values for the model. However, the difference in the values is marginal, indicating that hyperparameters need to be adjusted and tuned in future work. The results also showed that the model performs very well for the excavator class (mAP 75.8%). Adding more data would provide better insight into the results for the given class.

The coordinates from the bounding boxes (indicators) can be drawn on the images to analyze the detected objects. Fig. 7 shows an output of the object identification module based on the SSD model, showing the bounding boxes and their associated annotations as indicators of the construction resource that was detected in the image. The bounding boxes are examples of detected classes for different UAV missions.

A record is generated within segments of one second by the object-detection algorithm. The approach summarizes the occurrence of an activity within 60 frames per second in a text-based activity list. To reduce the effect of false positives, the algorithm counts the indicators after registering a minimum of 35 frames per segment with positive indicators (i.e., if an object is present in more than 35 frames, the algorithm registers the object as existing). Using the activity indicators, it was inferred that the given activity was being performed in that video segment. If there are no objects detected, then the activity output, along with the segments, is null.

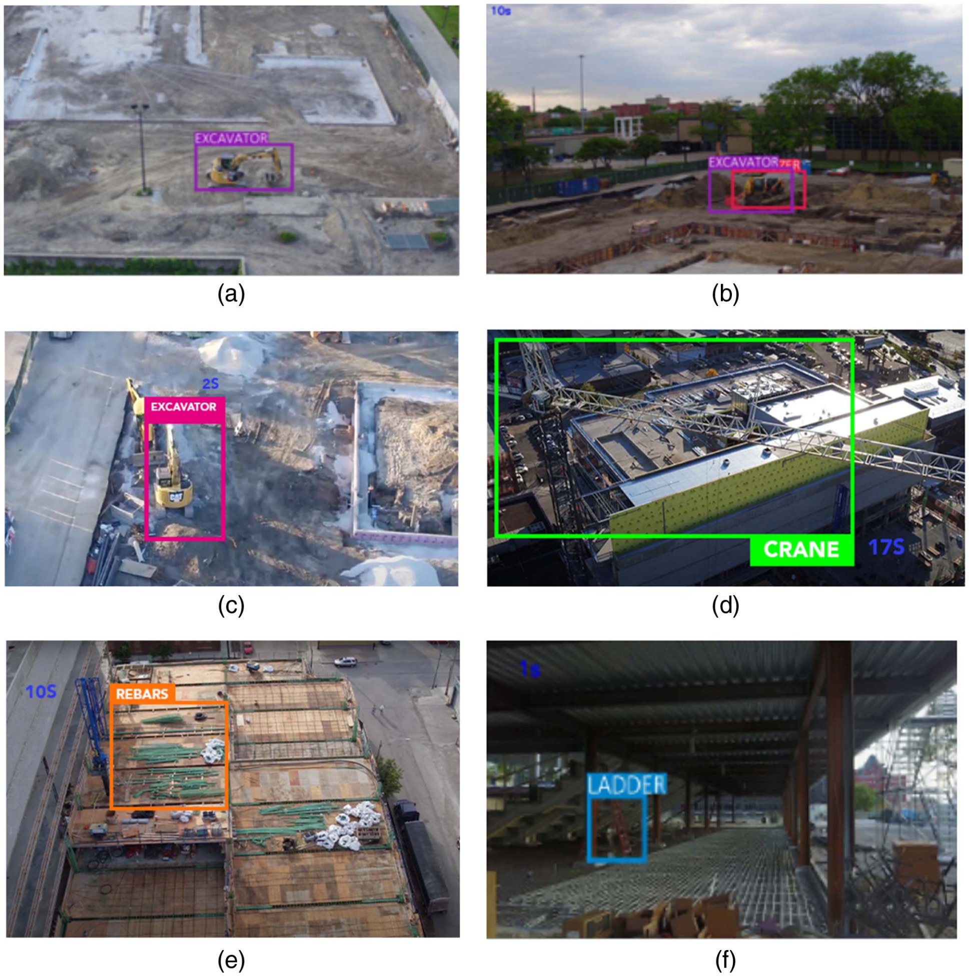
Demonstration Achieved Performance

The presented approach overcame three main challenges over from the reviewed implementations, highlighting advantages of the selected strategy in building object detectors using UAV aerial images. The first is the adaptation of SSD to enable real-time detection of construction resources from UAVs over R-CNNs. The second is learning rate annealing and cyclic learning rate training methodologies, which reduces the human effort of time for collecting and processing data. The approach demands a considerably smaller data set for training without compromising the model performance, thereby using less training and test training resources, facilitating the implementation with other construction classes. Using a smaller data set implies reducing the human effort to collect and process data, making it suitable for the reviewed and presented models. The third is custom CNN design to meet the requirements of images collected from UAV-sensing systems with two introduced concepts: spatial and temporal resolutions. A summary table shows the differences between the presented research and existing implementations (Table 3).

Discussion

Current UAV operations using standard digital camera lens and sensors involve constant movement and traveling along a path within a short time frame that makes it difficult to capture images to a ground area with adequate scale at a given instant in time—a spatial resolution problem. The UAV shifting position relative to the ground impacts the collected images' temporal resolution because it produces continuously changing viewing (sensing) area and viewing angles. Small scales and orientations of IFOV due to multiple positions and elevations in the construction zone airspace translate to major difficulties when implementing deep-learning-based objects (construction resources) detectors to function in real-time. Construction equipment and materials appear semioccluded within the images due to their location, joint positioning, and differing backgrounds.

The UAV images' unfavorable spatial and temporal resolutions cause an exhaustive data-preparation process for aerial images to make them suitable to use as a sample ground truth in the model. There were not relevant aerial image data sets (low-altitude image data sets) for construction projects. Most of the existing images were from the data sets with views taken from the ground, which have limited use for the object detectors. The researchers used web-scraped images. However, these images do not always provide the required viewpoint, and often images appear to be recorded from very low AGLs. Filtering images that have the desired viewing



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 7. Detected construction resources (classes) using predictors of the SSD models and UAV aerial images in real time: (a) excavator; (b) excavator and bulldozer; (c) excavator; (d) crane; (e) rebars; and (f) ladder. (Images by authors.)

Table 3. Performance reported from deep-learning-based detectors for construction resources

Other research efforts	Model	Data	UAV	Real Time	Score
Formwork detection in UAV pictures of construction sites (Jahr et al. 2018)	DetectNet	1,400 images (1 class)	Yes	Yes	44.48% (mAP)
Engineering vehicles detection based on modified faster R-CNN for power grid surveillance (Xiang et al. 2018)	Faster R-CNN	1,552 images (2 classes)	No	No	89.12% (mAP)
Single-label and multilabel classification of construction objects using deep-transfer learning methods (Nath et al. 2019)	VGG16	1,859 images (2 classes)	No	Yes	Accuracy: 85.5%, precision: 75.6%, recall: 95.3%
Detecting construction equipment using a region-based fully convolutional network and transfer learning (Kim et al. 2018)	R-FCN	2,920 images (5 classes)	No	No	94% (mAP)
Efficient single-shot-multibox detector for construction site monitoring (Thakar et al. 2018)	SSD-Inception	Not given	No	Yes	51.24% (mAP), 63.76% (mAP)
Presented approach	SSD 500	1,400 images (6 classes)	Yes	Yes	0.751% (mAP)

angle was challenging and time-consuming. Furthermore, the number of priors being used in the model determines the number of boxes that could be annotated for the classes in each image.

For the selected one-phase deep-learning model (SSD) in this study, the performance was affected when the number of bounding boxes exceeded five or more—this number required further filtering of the scraped data. Real-time applications seldom provided the option for filtering unwanted frames, thereby affecting the results. The degradation in performance can be overcome by increasing the number of priors used, but this strategy has a drawback due to increased inference time—a processing speed reduction. Finally, the selected SSD approach has a low localization loss, but a higher classification loss because the same bounding boxes are used to predict multiple classes. As a result, an increase in the number of classes increased the resultant loss. Therefore, there is a trade-off between processing speed (inference time) and the number of classes detected when using the SSD. For resource detection in construction, this drawback is a severe impediment because—even for few classes—because they might appear grouped in the construction zone at the same time.

The requirement for a large amount of data for training the neural network was a significant impediment during the implementation. This issue is common with deep-learning performances because data sets used directly impact the accuracy. The authors customized a design for the selected deep-learning model to reduce the effect, using the cyclic learning rate technique and proper tuning of the hyperparameters. The effect was obtaining the same or better results from half the amount of data and half the number of epochs that have been traditionally required using the same one-phase deep-learning model. In Fig. 6, it can be observed that test loss converges to 2.92 in 120 epochs, whereas the loss of test loss reaches 2.94 in 120 epochs. The graph trend indicates that the cyclic learning approach converges faster than the constant learning rate, where all other parameters are the same. Although learning-rate annealing results in a lower test loss, its performance is inferior in terms of test loss compared with the cyclic learning rate. From this observation, it can be inferred that training the model on a significantly larger data set can lead to better performance. Even by changing the nature of the web-scraped data, better results can be achieved.

Conclusion

The research presented an analytical conceptualization and a showcase demonstration to detect construction resources within aerial images in real time. The proposed conceptualization consisted of two parts: spatial resolution and temporal resolution. Spatial resolution refers to the IFOV within an image of objects and backgrounds on the ground with an adequate scale at a given instant in time. Temporal resolution refers to the frequency for obtaining imagery of the same particular area of a construction project. The new conceptualization served to analyze the challenges of designing and implementing a deep-learning algorithm for the case object detection in real time on construction sites.

The showcase demonstration is an implementation that detects objects of interest (construction resources) extracted from aerial images produced on UAV sensing systems in real time. The demonstration illustrated the challenges of implementation that resulted from the spatial and temporal resolutions. The authors built the implementation using a customized deep-learning technique (SSD) to extract and encode images from feature representations to predict objects of interest. These predictors are bounding boxes that provide information on the presence of construction resources that inform monitoring and project control tasks.

The implementation could detect six classes of construction resources in a given UAV video with mAP values in the range of 0.78–0.85. These mAP values are considered good results compared with implementations discussed in the literature review, particularly considering the size, aspect ratio, and UAV aerial image density. As discussed in the "Architecture (Phase 2)" section, existing implementations can achieve similar results, but they did not test the unfavorable spatial and temporal resolution conditions of UAV images. They also used considerably more extensive data for the selected classes, and the images were taken from fixed positions to cover the same particular area of construction sites.

The employed SSD300 method avoids using region-proposal techniques and performs detection in a single pass over the image. The approach can run in real time with high accuracy and on hardware with limited capacity, and outperformed its CNN counterparts, such as R-CNN and faster R-CNN, in speed. The use of multiscale feature maps improved object detection at different scales due to the low spatial resolution of aerial images (e.g., registration of construction equipment classes with very low scale values). The existing data for construction resources—any class or materials and equipment—are limited. Such data are nonexistent for aerial fields of view, which is a limiting factor for collection and training sample sets. The limitation impacts deep learning for object-detecting approaches because they require a good amount of data for training the models to make accurate predictions. The higher detection accuracy at different image scales reinforced the choice of the SSD algorithm.

The proposed research can also address different domains and applications in construction management with minor modifications to the architecture. For example, the applications can include detecting equipment at indoor construction sites or for a particular piece of equipment, it could account for different makes, models, or color. The proposed method can be expanded and transferred to construction projects by training multiple classes, making modifications in the neural network layers, and training it with relevant data. However, introducing multiple sets of classes of construction resources will necessitate additional diversified data for training the neural network, thereby enhancing the robustness of the model to improve spatial and temporal resolution conditions. The same approach used in this research can be implemented using less data and computing resources for better or equivalent performance as existing implementations.

Automating the pipeline would help reduce time spent specifying the model input for training, running it, and feeding the output for activity recognition in real time. For example, for data preparation to train the model, tools that automatically solve class imbalances—multiple sizes of objects of interest in the test images—will improve the final training network and efficiency of implementing the approach. Efforts for improvement of implementation should be directed to automation for data-augmentation techniques that allow the creation of syntactic scenarios [e.g., cropping and adding some images of construction resources (equipment) into image frames of empty construction sites]. The augmentation method allows for creating different activities in a single image of a vacant construction site, and the resultant images can then be used to train the model.

The spatial-and-temporal-resolution concept and techniques can be used to shed light on the potential problems using aerial images from UAV and to improve the performance of other researchers and practitioners' implementations from the community, thereby making them more suitable for deployment for activities such as for project control and management in real-time scenarios.

Finally, the research presented herein offers many important opportunities for further study. First, there is a trade-off between processing speed and the number of classes of construction resources that can be detected from a UAV due to the number of features/details identified from the birds-eye view of the image, which limits the number of classes (construction resources) that can be detected within images in real time. Second, new methods to generate data sources that better reflect the changing nature of the construction project environment promise an improvement in training model performance, thereby causing transferability of the approach to other construction projects (e.g., web-scraped data on multiple physical construction environments). Third, bounding boxes as outputs of construction resource detection could improve awareness for activity identification in construction management tasks by reducing object identification uncertainty for planning and monitoring tasks—reducing human effort as an input for human–machine interfacing conditions, which benefits, among other things, safety (particularly for roadway construction).

Data Availability Statement

Some or all data, models, or codes that support the findings of this study are available from the corresponding author upon reasonable request (SSD trained model, images used for training, and construction project UAV videos).

Acknowledgments

This research was partially supported by the National Science Foundation (NSF) Grant No. IIS 1550833.

References

Albeaino, G., M. Gheisari, and B. W. Franz. 2019. “A systematic review of unmanned aerial vehicle application areas and technologies in the AEC domain.” *J. Inf. Technol. Constr.* 24: 381–405.

Arabi, S., A. Haghighat, and A. Sharma. 2020. “A deep-learning-based computer vision solution for construction vehicle detection.” *Comput.-Aided Civ. Infrastruct. Eng.* 35 (7): 753–767. https://doi.org/10.1111/mice.12530.

Asadi, K., A. Kalkunte Suresh, A. Ender, S. Gotad, S. Maniyar, S. Anand, M. Noghabaei, K. Han, E. Lobaton, and T. Wu. 2020. “An integrated UGV-UAV system for construction site data collection.” *Autom. Constr.* 112 (Apr): 103068. https://doi.org/10.1016/j.autcon.2019.103068.

Avola, D., L. Cinque, G. L. Foresti, N. Martinel, D. Pannone, and C. Piciarelli. 2020. “A UAV video dataset for mosaicking and change detection from low-altitude flights.” *IEEE Trans. Syst. Man Cybernetics Syst.* 50 (6): 2139–2149. https://doi.org/10.1109/TSMC.2018.2804766.

Azar, E. R., S. Dickinson, and B. McCabe. 2013. “Server-customer interaction tracker: Computer vision-based system to estimate dirt-loading cycles.” *J. Constr. Eng. Manage.* 139 (7): 785–794. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000652.

Bakirman, T., B. Bayram, B. Akpinar, M. F. Karabulut, O. C. Bayrak, A. Yigitoglu, and D. Z. Seker. 2020. “Implementation of ultra-light UAV systems for cultural heritage documentation.” *J. Cult. Heritage* 44 (Jul–Aug): 174–184. https://doi.org/10.1016/j.culher.2020.01.006.

Barekatain, M., M. Martí, H. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger. 2017. “Okutama-Action: An aerial view video dataset for concurrent human action detection.” In *Proc., 2017 IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*. New York: IEEE.

Brilakis, I., M.-W. Park, and G. Jog. 2011. “Automated vision tracking of project related entities.” *Adv. Eng. Inf.* 25 (4): 713–724. https://doi.org/10.1016/j.aei.2011.01.003.

Deng, J., W. Dong, R. Socher, L. Li, L. Kai, and F.-F. Li. 2009. “ImageNet: A large-scale hierarchical image database.” In *Proc., 2009 IEEE Conf. on Computer Vision and Pattern Recognition*. New York: IEEE.

Deng, L. 2014. “Deep learning: Methods and applications.” *Found. Trends Signal Process.* 7 (3–4): 197–387. https://doi.org/10.1561/2000000039.

Dozzi, S. P., and S. M. AbouRizk. 1993. *Productivity in construction*. Ottawa: Institute for Research in Construction, National Research Council Ottawa.

Duque, L., J. Seo, and J. Wacker. 2018. “Bridge deterioration quantification protocol using UAV.” *J. Bridge Eng.* 23 (10): 04018080. https://doi.org/10.1061/(ASCE)BE.1943-5592.0001289.

Engel, J., J. Sturm, and D. Cremers. 2014. “Scale-aware navigation of a low-cost quadrocopter with a monocular camera.” *Rob. Auton. Syst.* 62 (11): 1646–1656. https://doi.org/10.1016/j.robot.2014.03.012.

Everingham, M., L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2009. “The pascal visual object classes (VOC) challenge.” *Int. J. Comput. Vision* 88 (2): 303–338. https://doi.org/10.1007/s11263-009-0275-4.

Fang, W., L. Ding, B. Zhong, P. E. D. Love, and H. Luo. 2018. “Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach.” *Adv. Eng. Inf.* 37 (Aug): 139–149. https://doi.org/10.1016/j.aei.2018.05.003.

Fang, W., P. E. D. Love, H. Luo, and L. Ding. 2020. “Computer vision for behaviour-based safety in construction: A review and future directions.” *Adv. Eng. Inf.* 43 (Jan): 100980. https://doi.org/10.1016/j.aei.2019.100980.

Ficapal, A., and I. Mutis. 2019. “Framework for the detection, diagnosis, and evaluation of thermal bridges using infrared thermography and unmanned aerial vehicles.” *Buildings* 9 (8): 179. https://doi.org/10.3390/buildings9080179.

Gheisari, M., and B. Esmaeili. 2019. “Applications and requirements of unmanned aerial systems (UASs) for construction safety.” *Saf. Sci.* 118 (Oct): 230–240. https://doi.org/10.1016/j.ssci.2019.05.015.

Girshick, R. 2015. “Fast R-CNN.” In *Proc., 2015 IEEE Int. Conf. on Computer Vision (ICCV)*. New York: IEEE.

Girshick, R., J. Donahue, T. Darrell, and J. Malik. 2014. “Rich feature hierarchies for accurate object detection and semantic segmentation.” In *Proc., 2014 IEEE Conf. on Computer Vision and Pattern Recognition*. New York: IEEE.

Golparvar-Fard, M., A. Heydarian, and J. C. Niebles. 2013. “Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers.” *Adv. Eng. Inf.* 27 (4): 652–663. https://doi.org/10.1016/j.aei.2013.09.001.

Greenwood, W. W., J. P. Lynch, and D. Zekkos. 2019. “Applications of UAVs in Civil Infrastructure.” *J. Infrastruct. Syst.* 25 (2): 04019002. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000464.

Irizarry, J., and D. B. Costa. 2016. “Exploratory study of potential applications of unmanned aerial systems for construction management tasks.” *J. Manage. Eng.* 32 (2): 10. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000422.

Jahr, K., A. Braun, and A. Borrmann. 2018. “Formwork detection in UAV pictures of construction sites.” In Vol. 12 of *Proc., eWork and eBusiness in Architecture, Engineering, and Construction*, edited by R. S. Jan Karlshoj, 265–271. Copenhagen, Denmark: CRC Press. https://doi.org/10.1201/9780429506215-33.

Jinwoo Kim, S. C. 2015. “Robust real-time object detection on construction sites using integral channel features.” In *Proc., Int. Conf. on Construction Engineering and Project Management*. Deakin, Australia: International Centre for Complex Project Management.

Kim, D., M. Liu, S. Lee, and V. R. Kamat. 2019. “Remote proximity monitoring between mobile construction resources using camera-mounted UAVs.” *Autom. Constr.* 99 (Mar): 168–182. https://doi.org/10.1016/j.autcon.2018.12.014.

Kim, H., H. Kim, Y. W. Hong, and H. Byun. 2018. “Detecting construction equipment using a region-based fully convolutional network and transfer learning.” *J. Comput. Civ. Eng.* 32 (2): 04017082. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000731.

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

Kwon, S., J.-W. Park, D. Moon, S. Jung, and H. Park. 2017. “Smart merging method for hybrid point cloud data using UAV and LIDAR in earthwork construction.” *Procedia Eng.* 196: 21–28. https://doi.org/10.1016/j.proeng.2017.07.168.

Li, Y., H. Dong, H. Li, X. Zhang, B. Zhang, and Z. Xiao. 2020. “Multi-block SSD based small object detection for UAV railway scene surveillance.” *Chin. J. Aeronaut.* 33 (6): 1747–1755. https://doi.org/10.1016/j.cja.2020.02.024.

Lin, T., P. Goyal, R. Girshick, K. He, and P. Dollár. 2020. “Focal loss for dense object detection.” *IEEE Trans. Pattern Anal. Mach. Intell.* 42 (2): 318–327. https://doi.org/10.1109/TPAMI.2018.2858826.

Liu, D., J. Chen, D. Hu, and Z. Zhang. 2019. “Dynamic BIM-augmented UAV safety inspection for water diversion project.” *Comput. Ind.* 108 (Jun): 163–177. https://doi.org/10.1016/j.compind.2019.03.004.

Liu, W., D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. 2016. “SSD: Single shot multibox detector.” In *Proc., European Conf. on Computer Vision*, 21–37. Cham, Switzerland: Springer.

Luo, H., C. Xiong, W. Fang, P. E. D. Love, B. Zhang, and X. Ouyang. 2018. “Convolutional neural networks: Computer vision-based workforce activity assessment in construction.” *Autom. Constr.* 94 (Oct): 282–289. https://doi.org/10.1016/j.autcon.2018.06.007.

Memarzadeh, M., A. Heydarian, M. Golparvar-Fard, and J. C. Niebles. 2012. Real-time and automated recognition and 2D tracking of construction workers and equipment from site video streams.” In *Proc., Int. Conf. on Computing in Civil Engineering*. Reston, VA: ASCE.

Nath, N. D., T. Chaspari, and A. H. Behzadan. 2019. "Single- and multi-label classification of construction objects using deep transfer learning methods." *J. Inf. Technol. Constr.* 24: 511–526. https://doi.org/10.36680/j.itcon.2019.028.

National Research Council. 2009. *Advancing the competitiveness and efficiency of the U.S. construction industry.* Washington, DC: National Academies Press.

Redmon, J., and A. Farhadi. 2018. "YOLOv3: An incremental improvement." Preprint, submitted April 8, 2018. http://arxiv.org/abs/1804.02767.

Ren, S., K. He, R. Girshick, and J. Sun. 2015. "Faster R-CNN: Towards real-time object detection with region proposal networks." Preprint, submitted June 4, 2015. http://arxiv.org/abs/1506.01497.

Rezazadeh Azar, E. 2017. "Semantic annotation of videos from equipment-intensive construction operations by shot recognition and probabilistic reasoning." *J. Comput. Civ. Eng.* 31 (5): 04017042. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000693.

Roberts, D., and M. Golparvar-Fard. 2019. "End-to-end vision-based detection, tracking, and activity analysis of earthmoving equipment filmed at ground level." *Autom. Constr.* 105 (Sep): 102811. https://doi.org/10.1016/j.autcon.2019.04.006.

Seo, J., S. Han, S. Lee, and H. Kim. 2015. "Computer vision techniques for construction safety and health monitoring." *Artif. Intell. Eng.* 29 (2): 239–251. https://doi.org/10.1016/j.aei.2015.02.001.

Sherafat, B., C. R. Ahn, R. Akhavian, A. H. Behzadan, M. Golparvar-Fard, H. Kim, Y.-C. Lee, A. Rashidi, and E. R. Azar. 2020. "Automated methods for activity recognition of construction workers and equipment: State-of-the-art review." *J. Constr. Eng. Manage.* 146 (6): 03120002. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001843.

Slaton, T., C. Hernandez, and R. Akhavian. 2020. "Construction activity recognition with recurrent convolutional networks." *Autom. Constr.* 113 (May): 103138. https://doi.org/10.1016/j.autcon.2020.103138.

Szegedy, C., L. Wei, J. Yangqing, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. "Going deeper with convolutions." In *Proc., IEEE Conf. on Computer Vision and Pattern Recognition*. New York: IEEE.

Thakar, V., H. Saini, W. Ahmed, M. M. Soltani, A. Aly, and J. Y. Yu. 2018. "Efficient single-shot multibox detector for construction site monitoring." In *Proc., Int. Smart Cities Conf. (ISC2)*, 1–6. New York: IEEE.

Tzutalin. 2015. "LabelImg." Accessed January 25, 2020. https://github.com/tzutalin/labelImg.

Wu, B., A. Wan, F. Iandola, P. H. Jin, and K. Keutzer. 2017. "SqueezeDet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving." In *Proc., 2017 IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*. New York: IEEE.

Xiang, X., N. Lv, X. Guo, S. Wang, and A. El Saddik. 2018. "Engineering vehicles detection based on modified faster R-CNN for power grid surveillance." *Sensors* 18 (7): 2258. https://doi.org/10.3390/s18072258.

Zhou, J., C.-M. Vong, Q. Liu, and Z. Wang. 2019. "Scale adaptive image cropping for UAV object detection." *Neurocomputing* 366 (Nov): 305–313. https://doi.org/10.1016/j.neucom.2019.07.073.

Zhou, Z., J. Irizarry, and Y. Lu. 2018. "A multidimensional framework for unmanned aerial system applications in construction project management." *J. Manage. Eng.* 34 (3): 04018004. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000597.

Zhu, Z., X. Ren, and Z. Chen. 2017. "Integrated detection and tracking of workforce and equipment from construction job site videos." *Autom. Constr.* 81 (Sep): 161–171. https://doi.org/10.1016/j.autcon.2017.05.005.