Improving Industrial Safety Gear Detection through Re-ID conditioned Detector

Manikandan Ravikiran
Research and Development
Hitachi India Pvt Ltd
Bangalore, India
manikandan@hitachi.co.in

Shibashish Sen
Research and Development
Hitachi India Pvt Ltd
Bangalore, India
shibashish.sen@hitachi.co.in

***Abstract*—Industrial safety gears such as hardhats, vests, gloves and goggles are vital in safety of workers. With the advancement of vision technologies, most industries are moving towards automatic safety monitoring systems for its enforcement. However, most of the industrial safety monitoring systems are plagued by the following problems. To begin with, object detection which is the principal component of this system suffers from the problem of false detections and missed detections which are extremely costly resulting in wrong safety monitoring alerts and safety hazards. Further, while video object detection has seen a large traction through ImagenetDet and MOT17Det challenges, to the best of our knowledge there is no work till date in the context of industrial safety. Finally, unlike existing areas of object detection where there is the availability of large datasets, best of existing research works in detecting industrial safety gears is restricted to mostly hardhats due to lack of large datasets. In this work, we address these previously mentioned challenges by presenting a unified industrial safety system. As part of this developed system, we firstly introduce safety gear detection dataset consisting of 5k images with the previously mentioned classes of safety gears and present exhaustive benchmark on state-of-the-art single frame object detection. Secondly, to address wrong/missed detections we propose to exploit temporal information from contiguous frames by conditioning the object detection in the current frame on results of re-identification of objects computed in prior frames. Finally, we conduct extensive experiments using the developed Re-ID conditioned object detection system with various state-of-the-art object detectors to show that the proposed system produces mAP of 85%, 87%, 92% and 78% with average improvements of 5% mAP across the previously mentioned safety gears under complex conditions of illumination, posture and occlusions.**

***Index Terms*—Object Detection, Safety Gear Detection, Person Re-Identification, Convolution Neural Networks**

I. Introduction

Industries such as manufacturing, construction, etc. are very essential and an integral part of infrastructure development which gives a tremendous boost to a countrys economy. The construction industry has registered enormous growth worldwide in recent years. Although the development of technology is rapid in most of the sectors, construction work is still labor-intensive, high-risk with workers working on complex tasks requiring high-intensity operations [1], which always leads to hazards.

According to the most recent United States' Bureau of Labor Statistics, the number of fatalities in the US has gradually increased by 26% between 2011 and 2016 [2]. Similarly, according to the UK Health and Safety Executive (HSE), 38 construction workers suffered fatal injuries in Great Britain between April 2014 and March 2015, while this figure rose to 45 during the same period the following year. Finally, in India [3], nearly 48,000 workers die in the country due to occupational accidents out of which 24.2% is recorded in the construction industry.

All these statistics put forth the need for construction safety and risk assessment technologies. Multiple types of injuries occur in construction ranging from injuries due to falls, falling material and objects, electrical hazards, etc. with each having consequences that are more serious than the other. Most of these accidents involving eyes, legs, feet, and toes are often fatal [4]. Further many heads and neck injuries are caused by falling from a height or being struck by vehicles and other moving plants and equipment. From 2003 to 2010, 2210 construction workers in the United States died as a result of traumatic brain injuries [5], [6], accounting for 24% of the total number of deaths from construction accidents.

Wearing safety gear such as hardhat, safety vests, gloves, and goggles is an effective protective measure for minimizing the risk of serious injuries. In construction accidents, the hardhats protect workers by resisting penetration by objects, absorbing shock from direct blows to the head by objects and reducing electrical shock hazards. Safety goggles prevent exposure of eyes to molds, fungi or rodent droppings. Safety gloves prevent burns and electrical shocks. Further, the Bureau of Indian standards recommends wearing hardhats, goggles, and gloves in typical construction environments when there is a potential for a head injury from impacts, falling or flying objects, or electrical shock [8]. Despite the vital role of these safety gears in protecting life, it is well-known that the injuries still occur due to lack of wearing any protective equipment [7].

An automated industrial safety monitoring system helps to improve the safety compliance of construction workers by ensuring safety gears are appropriately worn. It is well known that traditional safety management methods, such as risk analysis and safety training, have seldom effect on construction safety [9]. An automatic construction surveillance system will

978-1-7281-4732-1/19/$31.00 ©2019 IEEE

be helpful in real-time monitoring which in turn improves worker safety and reduces labor costs. However, most of the industrial safety monitoring systems are plagued by the following open problems.

- To begin with, object detection which is the principal component of this system, which is used for detecting safety gears suffers from the problem of false/missed detections and miss classifications which are extremely costly resulting in wrong safety monitoring alerts and safety hazards [10].
- Secondly, video stream based processing to reduce detection errors is yet to be adapted for industrial safety, despite main stream adaptation in general AI tasks [11], [12].
- Finally, unlike existing areas of computer vision, where there is the availability of large public datasets, best of existing research works in detecting industrial safety gears is restricted to mostly hardhats due to lack of large datasets.

With previously mentioned points in mind, through this work we make the following contributions:

- Firstly we introduce safety gear detection dataset consisting of 5k images with annotations for four safety gears including helmet, jacket, goggle and glove.
- Second, to address video level wrong/missed detections we propose to exploit temporal information from contiguous frames by propagating boxes through trackers.
- Thirdly, to address video level miss classification we propose to condition class label of detected object in the current frame based on the results of re-identification of objects computed in prior frames. Such a conditioning mechanism helps in propagation of class labels, thereby reducing the miss classification. Further in this work, it's also supplemented with decoupled classification refinement.
- Finally, we conduct extensive experiments to show that the proposed system produces mAP of 85%, 87%, 92% and 78% with average improvements of 5% mAP across the previously mentioned safety gears under complex conditions of illumination, posture and occlusions.

The rest of the paper is organized as follows. In section II we present literature, foll wed by experimental setup, algorithm and dataset used in section III. In section V we present various experiments on the developed approach. Finally, in section VII we conclude with discussion on work done and some implications on possible future works.

II. Related Works

Computer Vision for Industrial Safety Systems: Large number of works exist on computer vision based construction safety monitoring with majority of works falling in detection of hard hats and safety vests with Fang et al. [13] presenting detailed study of Faster-RCNN for task of safety helmet detection under various visual conditions. Son et al. [14] focusing on detection of Industrial workers with varying posture, appearance and backgrounds using Faster RCNN, however the approach is restricted to detection of non hard hat use only. Fang et al. [15] presenting Improved faster-RCNN for improving generalization of worker detection. Du et al. [16] propose using facial features, motion and color information of workers wearing same colored helmets and facing towards camera, which reduces applicability of such a method in actual scenarios and similar works are also proposed by Shrestha et al. [17] which uses edge information. Multiple other works are presented for hard helmet detection ranging from use of Histogram of Oriented Gradient (HOG) with Circle Hough Transform (CHT) [18], HOG feature template of a human object , cascade classifiers [19] and color information [20] While detecting hard hats has seen large body works there are few works detecting safety vests, notable work include on by Mosberger et al. [21] which proposes using combination of segmentation, localization and classification to detecting safety jackets. Then there are works by Park et al. [22] that uses fluorescent color of safety vests by processing local color histograms extracted from the regions of interest. In general, most existing work focus on detecting either hard hats or safety vests only. In our work, we extend to scope to four classes of safety gears namely hard hats, vests, goggles and gloves respectively.

Object Detection: Most of the recent work in object detection has focused on single-frame images dominated by Convolutional neural networks beginning with series of works by Girshick et al. [24]–[26] which were famous for their high accuracy and speed where they replaced scanning window detectors such as Viola-Jones [27], with a region proposal and classifier pipeline. Since 2015, Single Shot Detectors [28]–[34] have taken over which replaced the proposal-plus-classification paradigm with a regression formulation that directly estimates a set of bounding boxes and class labels. Both region proposal style detectors, and single shot detectors, are fast and reasonably accurate.

When developing object detection algorithms for surveillance situations, it is important to note that, the natural input is video (stream of frames). However, most of the standard previously mentioned detectors treat each frame independently, and simply process the input one frame at a time. There has been a comparatively small amount of work on object detectors that explicitly take a sequence of multiple frames as input. Under this we two variants namely have box-based approaches [36] and feature based approaches [35] which are as explained.

Box based approaches operate on the sequence of bounding-boxes produced by object detectors applied independently to multiple sequential frames. Most notable of the work in box-based approaches include work by Han et al. [36] that replaces standard Non-Maximal Suppression (NMS) with one that incorporates bounding boxes from multiple frames. Followed by works of Tripathi et al. [37] describe another box-level technique that processes a sequence of object detector outputs using a recurrent network to improve object predictions. Works by Kang et al. [38], [39] use the output from a single-frame detector to produce spatial-temporal Tubelet that are further processed to generate improved box predictions. Similarly, work of Lu et al. [40] use feature maps from a single-frame detector within detected regions and pass these to a recurrent network that outputs new bounding boxes and class probabilities. All these approaches are related to another class of box-level techniques known as tracking-by-detection [41]–[43] where the basic idea is to associate detections across the output of an object detector applied independently to sequential single-frame images to create tracks that can be used to remove false positives and restore missed detections.

Feature based approaches integrate features from multiple frames, rather than independent application of detector like box-based approach. Feature based approaches including most initial work Zhu et al. [44] uses optical flow to warp feature maps computed from two input frames into correspondence. Following this, again [45] combine two orthogonal ideas with a spatially adaptive feature computation to further improve results. Another key work by Friechtenhofer et al. [46] uses a deep network to combine detection and tracking to improve object detection in videos. More recently, there are works Broad et al. [47] that use recurrent layer for fusing features.

In the video object detection literature, there has been significantly more work on box-level methods owing to its natural sync with human thinking. As such in this work, we focus on box based detectors. More specifically, in this work we propose re-identification conditioned sequential detector with tracker to handle problem of missed and incorrect detections. Furthermore, we address classification errors through combination of decoupled classification refinement and re-identification conditioned sequential detector.

III. Datasets and Experimental Setup

A. Datasets

Since there is no off-the-shelf dataset available, a dataset was created in internal setting under simulated conditions that is typical across various industrial sites. The simulated workers along with their safety gears were annotated to generate the ground truth for training.

Data Collection: Videos were collected under following conditions.

- **Visual Range :** Near and far range of objects up-to 20 feets.
- **Illumination :**Include frames of bright and dark illumination.
- **Poses:** Include varying poses for the worker under consideration. See Figure 1.

Image annotation: For each of the worker in the image frame we annotated following

- **Bounding Boxes:** For each worker wearing/not wearing safety gear we annotated bounding boxes for the worker and his/her head. This was used for training the DCR based object detector used in the system (See section III-B). All the annotation were done using *Labelme* [50] tool.
- **Classification labels:** Further, for each of the bounding box annotated worker we further annotate binary class labels indicating presence or absence of the of Jacket and Glove. Similarly, for each of the annotated head we create two binary class labels indicated presence or absence of helmet and goggle. Overall, we have two detections, with two class label for each.

The final dataset statistics is as shown in Table III-A. We separate the dataset into three splits of training, development and test sets randomly, in the ratio of 80%:10%:10%. Sample images and their ground truth annotations are as shown in the figure 1.

B. Multistage Decoupled Classification Refinement (DCR)

Deep learning object detectors usually involve a backbone network such as VGG or Resnet etc. that is trained on large image classification datasets to yield scale-invariant features. Following this, a localization branch is connected to this backbone. This results in a conflict of covariance and correlation, where correlated features are required for classification and covariant features are needed for detection. Alternatively, fine-tuning end-to-end will force the backbone to gradually learn translation covariant feature, which might potentially downgrade the performance of the classifier. As such based on the works of Cheng et al. [48], we propose to use a Multistage Decoupled Classification Refinement (DCR) detector. Diagrammatic representation of the same is shown in Figure 2 below.

For this work, we use three state-of-the-art object detection models including RetinaNet, Faster-RCNN and Single Shot Multibox Detector (SSD) with previous proposed modification.

C. Evaluation Metric

1) Accuracy: We use PASCAL VOC mAP metric as the measure of accuracy. More details of the same can be found in [49].

2) Speed: In order to cater real time requirements, we calculate the speed of our entire approach. The speed in our case is calculated as the time taken for processing one image as it goes to through our entire pipeline starting from input to getting final detections and class labels. Since we calculate the speed in GPU, the resultant values obtained are very fast.

3) Robustness: Robustness represents the degree of tolerance of an object detection method when applied to testing various images. Industrial sites are usually in open outdoor environments and contain a large amount of workers, equipment and material. Therefore, changes in weather, illumination, individual postures, visual range and occlusions frequently occur on industrial sites. These factors inevitably have a significant impact on the visual features on such work sites. A good algorithm should be robust to such changes and not degrade significantly under varying conditions. Correctness and speed in different situations are indicators reflecting the robustness of the model.

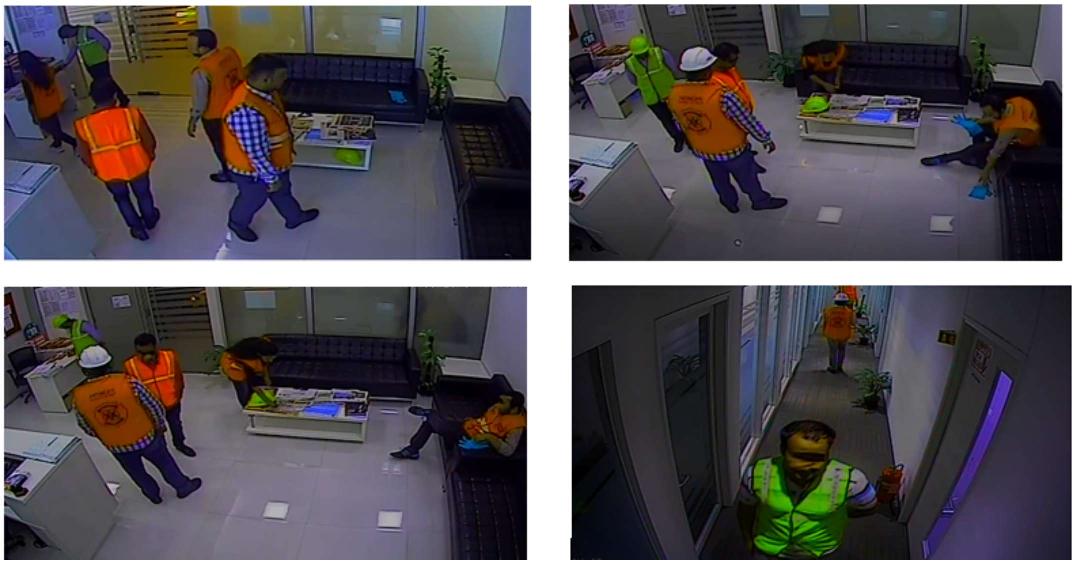

Fig. 1. Example images from dataset showing varying poses, brightness, range and occlusion in workers with safety gears. Best viewed in color.

TABLE I
DATASET CHARACTERISTICS OF SAFETY GEAR DETECTION DATASET.

Categories	No	Values	Total Instances				Number of Frames
			H	J	GL	GO	
Illumination	**1**	**Bright**	800	600	400	400	500
	2	**Dark**	400	400	400	400	500
Posture	**1**	**Standing**	500	600	400	400	500
	2	**Bending**	350	200	100	200	500
	3	**Sitting**	300	100	150	50	500
Range	**1**	**Small**	200	150	200	300	500
	2	**Medium**	150	100	50	50	500
	3	**Large**	100	100	50	100	500
Occlusion	**1**	**No occlusion**	50	50	50	50	500
	2	**Partial Occlusion**	50	50	50	50	500

IV. RE-ID CONDITIONED DETECTION ALGORITHM

A. Background

Given an continuous steam of video of frame $I_t, t = 0, 1, ..N-1, N, N+1, .., T$, our goal is to avoid missed detection, wrong detections and miss classification across the frames. Let D_t be the tracklet of frame t such that $D_t = \{< B_t, C_t >\}$ where B_t, C_t denotes its bounding boxes and class label. A scheme widely adopted in previous work [41]–[43] is sequential detection with tracking, outlined in Algorithm 1.

Given a video frame I_t, an object detector for individual images is first applied to produce per-frame detection result $B_t =$ **DetectOnImage**(I_t) where B_t denotes a set of bounding boxes together with their corresponding category scores. Non-maximum suppression is then applied to remove redundant bounding boxes, resulting in $B_t = NMS(B_t)$. Then the tracking algorithm associates the existing tracklets D_{t-1} to the detection results B_t producing tracklets up to frame I_t as $D_t =$ **AssociateTracklet**(D_{t-1}, B_t) outputting $\{B_t\}$. Additionally, Box propagation is applied, where detected boxes in the existing tracklets D_{t-1} are propagated to the current frame, $B_t^{'} =$ **PropagateBoxKalman**(D_{t-1}). The propagated boxes are concatenated with the per-image detected boxes as $B_t = [B_t^{'}, B_t]$ which is again followed by non-maximum suppression and are associated to the existing tracklets.

Algorithm 1: Sequential Detector with Tracker

```
Input: Video Frames $\{I_t\}_{t=0}^{T}$
Output: : All boxes $\{B_t\}_{t=0}^{T}$
Procedure:
    $B_0 =$ DetectOnImage$(I_0)$
    Initialize the tracklets $D_0$ from $B_0$
     for $t = 1$ to $T$
       $B_t =$ DetectOnImage$(I_t)$
       $B_t^{'} =$ PropagateBoxKalman$(D_{t-1})$
       $B_t = [B_t^{'}, B_t]$ := Box Concatenation
       $B_t = NMS(B_t)$
       $D_t =$ AssociateTracklet$(D_{t-1}, B_t)$
     end for
```

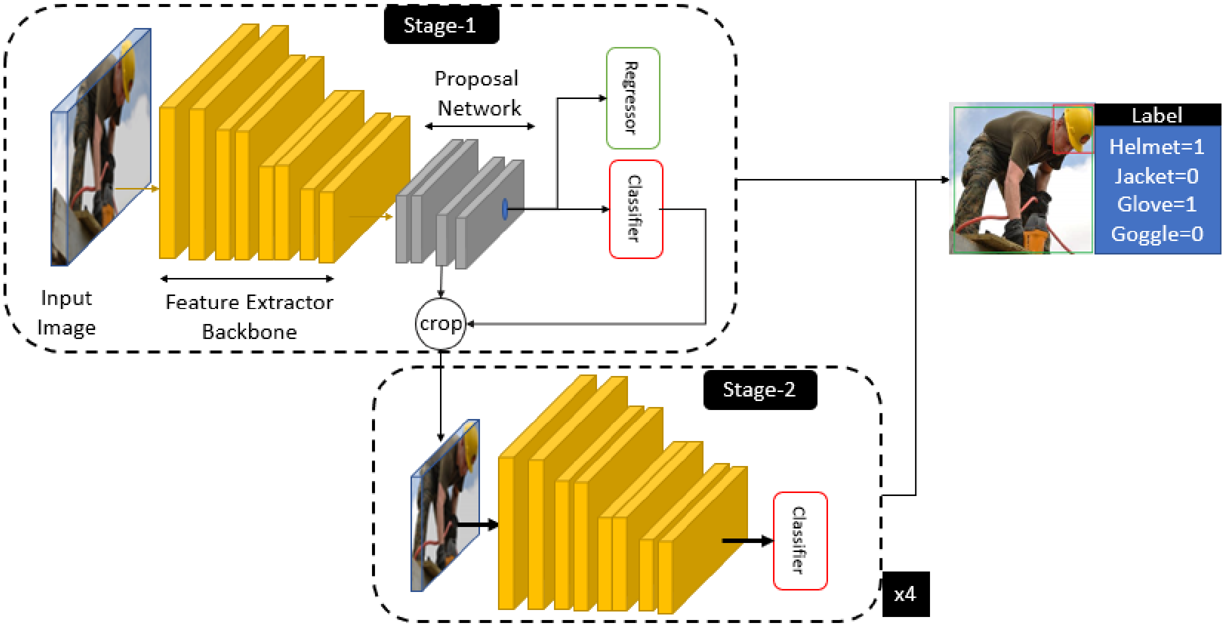

Fig. 2. Decoupled Classification Refinement with Multistage Training based Detector. The classifier stage is repeated four time for each of the safety gear classes.

B. Algorithm

Our target is to solve two problems of video level missed & wrong detection and miss classification.

- **Missed Detection:** Missed detections are handled by including boxes that are relevant, based on information from prior frames and boxes of new detections
- **Wrong Detection:** Wrong detections are handled by coupling multiframe detections through Re-Identification a.k.a Re-Id Conditioning
- **Miss Classification:** Miss classification are handled by propagating class labels. Further this is supplemented by DCR detectors to avoid inherent network level classification errors.

The revised re-identification conditioned detection algorithm is as shown in Algorithm 2. Solution 1 is executed by steps $M_t = \textbf{IOU}(B^{'}_t, B_t)$ and $O_t = [B^{'}_t, B_t]$ and Solution 2 is taken care by $N_t = \textbf{PersonReID}(B^{'}_t, B_t)$ and $B^{'}_t = M_t$ **and** N_t **or** O_t.

Algorithm 2: Re-identification Conditioned Sequential Detector with Tracker

```
Input: Video Frames $\{I_t\}_{t=0}^{T}$
Output: : All boxes $\{B_t\}_{t=0}^{T}$
Procedure:
    $B_0 = \textbf{DetectOnImage}(I_0)$
    Initialize the tracklets $D_0$ from $B_0$
    for $t = 1$ to $T$
        $B_t = \textbf{DetectOnImage}(I_t)$
        $B^{'}_t = \textbf{PropagateBoxKalman}(D_{t-1})$
        $M_t = \textbf{IOU}(B^{'}_t, B_t)$
        $N_t = \textbf{PersonReID}(B^{'}_t, B_t)$
        $O_t = [B^{'}_t, B_t]$
        $B^{'}_t = M_t$ and $N_t$ or $O_t$ := Re-ID Conditioning
        $B_t = NMS(B_t)$
        $D_t = \textbf{AssociateTracklet}(D_{t-1}, B_t)$
    end for
```

V. Experiments

Previously in section IV we presented the re-identification conditioned detector with tracker algorithm. In this section, we present detailed ablation study and results so obtained. The analysis and finding are presented in three parts namely i) Results across detectors ii) Results across various parameters iii) Comparison across baselines and proposed algorithms iv) Error source and pending problems.

A. Experiment 1: Baselines

To begin with we create exhaustive evaluation in the context of worker safety gear detection using DCR detectors (section III-B) and sequential detector with tracker algorithm (Algorithm 1), on the datasets discussed earlier in section III-A. Firstly, we trained three DCR Detectors on the training sets, with each of architectures tuned for various hyper parameters such as learning rate etc. to ensure quicker convergence. The results so obtained are as shown in Table II and III. The details of analysis reveals following.

- **Baseline v/s Proposed Approaches:** Comparing the baseline results in Table II with results obtained using Algorithm 1 from Table III. From Table III we can see, that sequential detectors does perform significantly better than the baseline detectors, by a margin of 2% mAP.

Comparing Tables II and III, we can see that this is mainly due to increase in results of conditions involving occlusion and range. We will revisit analysis in upcoming section (see section VI) for each conditions in section VI.

- **Results across Detectors:** As far as individual detectors go, DCR-SSD overall performs significantly better than both DCR-Faster-RCNN and DCR-RetinaNet for most of the safety gears. More specifically DCR-SSD outperforms DCR-RetinaNet for all the four safety gears across all the conditions of illumination, posture, range and occlusion. This is also true even in the case of DCR-SSD with Algorithm II. DCR-SSD and DCR-Faster-RCNN perform very similarly to DCR-Faster-RCNN across both Tables II and III except DCR-Faster-RCNN performs well across illumination and occlusion by average of 2% mAP. Further this observation is true across Tables II and III. We believe this is because of richer capacity of region proposal network where during ROI pooling, they retain information that is mostly dominated, which in our case the detection themselves rather than the factor such as illumination or occlusion. We leave more detailed analysis of these results for future works.
- **Errors:** As far as source of errors, in individual safety gears are concerned the insight so obtained from error analysis was common across the all the detectors. For Helmet, Jacket and Goggle the drop in mAP is mainly due to classification error. However, for glove missed detection error dominated more than the classification error.

B. Experiment 2: Re-ID Conditioning

Having examined the baselines using different detector alone and with algorithm 1, in this section we asses the performance using proposed re-identification conditioned sequential detection approach. As usual, use multistage DCR detectors of Faster-RCNN, SSD and RetinaNet.

More specifically in this experiment, we modify the step of concatenation of propagated boxes, in the original sequential detection algorithm to include boxes that are relevant from previous frames and the newly predicted boxes. We do this in three stages, where we first detect boxes in the current frame, next we find boxes relevant w.r.t previous frame through Intersection-over-union. Following this, we do re-identification of objects so obtained after IOU step to reduce classification error. Since our core task is to detect if the safety gears are worn, rather than 1-1 mapping of people between frames we conjecture re-identification will significantly help in improving results. Finally, we merge these re-identified boxes with newly obtained detection for the current frame. Thus in the process, we rectify both the classification labels and detected boxes if any. The results so obtained are as shown in Table IV for all the three detectors.

- **Baseline v/s Proposed Approaches:** Firstly comparing Tables IV with IV and II, we see that the results are significantly higher across all the conditions of illumination, posture, range and occlusion.
- **Results across Detectors:** inline with observations from section V-A we can see that DCR-SSD outperforms DCR-RetinaNet across all the results and performs very similar to DCR-Faster-RCNN. Further comparing Tables IV,III and II we can see that mAP improves by average of 10% for conditions involving range and occlusion, which is as expected when detection is coupled with tracking and re-identification. Also we can see, that Re-ID conditioned DCR-SSD outperforms DCR-Faster-RCNN for Helmet, Goggle and Re-ID Conditioned DCR-Faster-RCNN outperforms DCR-SSD for Glove and Jacket by average of 1%. Also, the results of gloves are very similar across the detectors and across the Tables IV,III and II. This is because, the detectors so trained seldom detects gloves or gives wrong detection, as such leading to higher error, compared to other safety gears.
- **Errors:** This is inline with, previously described observation.

VI. Ablation Study

To evaluate the robustness of the developed system, in this section we study impact of parameters such as illumination, posture, range and occlusion on the overall performance across the baselines and the proposed system.

A. Impact of Illumination

It is well known that typical industrial works are extremely long and time consuming, thus requiring workers to work on varying lighting conditions. This is true both in the case of outdoor and indoor industrial workers. Thus the developed method should be robust against lighting variations. Keeping this mind, we evaluate robustness of our approach for varying lighting conditions namely bright and dark conditions which corresponds to morning and evening work time. The results for the same are as shown in Table V across the baseline as well as the proposed re-id conditioned detection algorithm.

From the Table V, we can make following observations:

- **Results across Detectors:** Among all the detectors with various illumination DCR-SSD, produces top results for Helmet and Goggle, while DCR-Faster-RCNN produces best result for Jacket and Glove. DCR-Retina-Net produces results very close to that of rest.
- **Results across the parameter:** Results under bright illumination is higher by an average mAP of 1% across all the experiments. More specifically for the baseline approach, Algorithm 1 and Algorithm 2 brighter results are higher by 4%, 2% and 1% mAP respectively, which shows that the proposed approach does improve results with varying brightness conditions, where the detectors do fail. Results for Glove safety gear, is same across all the experiment with a values of 76% mAP in DCR-RetinaNet. Finally, we can see that the net improvement obtained for darker illumination (7% mAP) is higher than that of bright illumination (4% mAP).
- **Baseline v/s Proposed Approaches & Errors:** Comparing results across baseline, algorithm 1 and algorithm

TABLE II
BASELINE PERFORMANCE (MAP@ THRESHOLD=0.55) USING VARIOUS DETECTORS. H:=HELMET, J:=JACKET,GL:=GLOVE,GO:=GOGGLE

Categories	DCR-Faster-RCNN				DCR-SSD				DCR-RetinaNet			
	H	J	GO	GL	H	J	GO	GL	H	J	GO	GL
Illumination	0.76	0.67	0.93	0.73	0.74	0.67	0.93	0.73	0.70	0.70	0.91	0.73
Posture	0.79	0.67	0.88	0.79	0.79	0.67	0.88	0.78	0.70	0.74	0.89	0.78
Range	0.74	0.67	0.87	0.75	0.74	0.64	0.87	0.77	0.67	0.67	0.87	0.77
Occlusion	0.77	0.65	0.90	0.75	0.75	0.64	0.90	0.74	0.68	0.68	0.9	0.74
Average	**0.77**	**0.67**	**0.90**	**0.75**	**0.75**	**0.65**	**0.90**	**0.75**	**0.68**	**0.70**	**0.90**	**0.76**

TABLE III
PERFORMANCE (MAP@ THRESHOLD=0.55) OF VARIOUS DETECTORS USING SEQUENTIAL DETECTOR WITH TRACKER.H:=HELMET, J:=JACKET,GL:=GLOVE,GO:=GOGGLE

Categories	DCR-Faster-RCNN				DCR-SSD				DCR-RetinaNet			
	H	J	GO	GL	H	J	GO	GL	H	J	GO	GL
Illumination	0.76	0.67	0.93	0.73	0.74	0.67	0.93	0.73	0.7	0.7	0.91	0.76
Posture	0.79	0.67	0.88	0.79	0.79	0.67	0.88	0.78	0.7	0.74	0.89	0.79
Range	0.74	0.68	0.89	0.76	0.76	0.66	0.89	0.78	0.67	0.67	0.87	0.75
Occlusion	0.79	0.65	0.91	0.76	0.76	0.66	0.91	0.74	0.69	0.69	0.91	0.76
Average	**0.77**	**0.67**	**0.90**	**0.76**	**0.76**	**0.67**	**0.90**	**0.76**	**0.69**	**0.70**	**0.90**	**0.77**

TABLE IV
PERFORMANCE (MAP@ THRESHOLD=0.55) OF USING PROPOSED RE-ID CONDITIONED DETECTION ALGORITHM. H:=HELMET, J:=JACKET,GL:=GLOVE,GO:=GOGGLE

Categories	DCR-Faster-RCNN				DCR-SSD				DCR-RetinaNet			
	H	J	GO	GL	H	J	GO	GL	H	J	GO	GL
Illumination	0.81	0.81	0.91	0.77	0.83	0.80	0.92	0.73	0.82	0.81	0.91	0.76
Posture	0.80	0.81	0.87	0.78	0.84	0.80	0.92	0.78	0.82	0.80	0.89	0.79
Range	0.86	0.8	0.89	0.77	0.87	0.78	0.89	0.78	0.81	0.78	0.87	0.75
Occlusion	0.87	0.81	0.91	0.77	0.87	0.81	0.91	0.74	0.83	0.79	0.91	0.76
Average	**0.83**	**0.81**	**0.90**	**0.77**	**0.85**	**0.80**	**0.91**	**0.76**	**0.82**	**0.80**	**0.90**	**0.77**

TABLE V
COMPARISON OF PERFORMANCE (MAP@ THRESHOLD=0.55) ACROSS APPROACHES WITH VARYING ILLUMINATION

Approch		DCR-Faster-RCNN				DCR-SSD				DCR-RetinaNet			
		H	J	GO	GL	H	J	GO	GL	H	J	GO	GL
Baseline	**Bright**	0.78	0.72	0.94	0.73	0.76	0.72	0.94	0.73	0.70	0.74	0.93	0.76
	Dark	0.72	0.63	0.92	0.73	0.72	0.63	0.92	0.73	0.70	0.65	0.9	0.76
	Average	**0.76**	**0.67**	**0.93**	**0.73**	**0.74**	**0.67**	**0.93**	**0.73**	**0.7**	**0.7**	**0.91**	**0.76**
Algorithm 1	**Bright**	0.78	0.72	0.9	0.76	0.76	0.72	0.93	0.73	0.7	0.7	0.91	0.76
	Dark	0.72	0.63	0.9	0.76	0.76	0.63	0.92	0.73	0.69	0.69	0.9	0.76
	Average	**0.76**	**0.67**	**0.93**	**0.73**	**0.74**	**0.67**	**0.93**	**0.73**	**0.7**	**0.7**	**0.91**	**0.76**
Algorithm 2	**Bright**	0.82	0.82	0.91	0.77	0.84	0.8	0.92	0.73	0.82	0.81	0.91	0.76
	Dark	0.81	0.81	0.9	0.77	0.82	0.8	0.91	0.73	0.81	0.8	0.9	0.76
	Average	**0.81**	**0.81**	**0.91**	**0.77**	**0.83**	**0.8**	**0.92**	**0.73**	**0.82**	**0.81**	**0.91**	**0.76**

2, we can see that an average improvement of 9% for Helmet, 13% for Jacket and 3% for glove and goggle showing a minor drop of 1%. We believe that the drop in goggle's performance is an outlier result due to kalman filter parameters and leave further exploration of results for future work.

B. *Impact of Posture*

Typically industrial work consists of large number of complex tasks ranging from carrying bricks to bending bar etc. thus worker constantly work in varying posture ranging from sitting, standing and bending. To accommodate the same, we collected dataset inline as shown from Table I. The dataset is collected in way such that that all the three variations of posture are present and each of the frame consists of two or more postures simultaneously. We calculate results, through manual analysis of each of the detections across categories to obtain following insights.

- **Baseline v/s Proposed Approaches & Errors:** We obtain net improvement of 1% and 14% mAP for Helmet and Jacket classes, with a drop of 1% mAP for glove and goggle. Also we don't see any improvement for Algorithm 1, when compared to baselines.
- **Results across Detectors:** Among all the detectors with various posture conditions DCR-SSD, produces top

TABLE VI
COMPARISON OF PERFORMANCE (MAP@ THRESHOLD=0.55) ACROSS APPROACHES WITH VARYING POSTURE

Approch		DCR-Faster-RCNN				DCR-SSD				DCR-RetinaNet			
		H	J	GO	GL	H	J	GO	GL	H	J	GO	GL
Baseline	**Standing**	0.8	0.67	0.89	0.79	0.8	0.67	0.88	0.78	0.72	0.76	0.9	0.79
	Bending	0.78	0.67	0.88	0.79	0.78	0.67	0.88	0.78	0.69	0.73	0.88	0.79
	Sitting	0.78	0.67	0.88	0.79	0.78	0.67	0.88	0.78	0.69	0.73	0.88	0.79
	Average	**0.79**	**0.67**	**0.88**	**0.79**	**0.79**	**0.67**	**0.88**	**0.78**	**0.7**	**0.74**	**0.89**	**0.79**
Algorithm 1	**Standing**	0.8	0.67	0.89	0.79	0.8	0.67	0.88	0.78	0.72	0.76	0.9	0.79
	Bending	0.78	0.67	0.88	0.79	0.78	0.67	0.88	0.78	0.69	0.73	0.88	0.79
	Sitting	0.78	0.67	0.88	0.79	0.78	0.67	0.88	0.78	0.69	0.73	0.88	0.79
	Average	**0.79**	**0.67**	**0.88**	**0.79**	**0.79**	**0.67**	**0.88**	**0.78**	**0.7**	**0.74**	**0.89**	**0.79**
Algorithm 2	**Average**	0.8	0.82	0.87	0.78	0.84	0.8	0.92	0.78	0.84	0.81	0.9	0.79
	Dark	0.79	0.81	0.87	0.78	0.84	0.8	0.92	0.78	0.81	0.8	0.89	0.79
	Sitting	0.79	0.81	0.87	0.78	0.84	0.8	0.92	0.78	0.81	0.8	0.89	0.79
	Average	**0.8**	**0.81**	**0.87**	**0.78**	**0.84**	**0.8**	**0.92**	**0.78**	**0.82**	**0.8**	**0.89**	**0.79**

results for Helmet and Goggle, while DCR-RetinaNet produces best result for Jacket and Glove. DCR-Faster-RCNN produces results very close to that of RetinaNet, this is unlike the case of illumination. We believe this is because of feature pyramids used as part of the DCR-RetineNet architecture. However since, the net difference being 1%, we leave such an analysis of impact of feature pyramids on detecting varying sized objects to future work.

- **Results across the parameter:** Among all the results of varying posture, we see that results with standing posture is higher than that of bending and sitting postures. Further, we also see that results for bending and sitting are similar to one another. Moreover, this trend is consistent across all the algorithms and across the different detectors. We also see that results of gloves is constant across all the experiments.

C. Impact of Range

Generally in construction and industrial areas the cameras are placed in variety of places. For example if the construction is indoor then the camera is placed in very low level thereby most objects in the scene including construction workers are very close to one another. If its a open construction areas there is a need for catering large visual ranges, especially due to stochastic camera placements. These camera positions significantly affect the performance of the developed method due to relative quality of features so extracted. As such in our case keeping these situations in mind, we present three cases of small, medium and large where small is distance of object from camera in range of 5-10ft distance , medium is in range of 10-30ft and large consists of objects that are 40ft away from the camera. In our case, based on this requirement we collected dataset across two different area that allows for capturing in such a setting. Example images are shown in Figure 1. Experiments with these setting show following observations.

- **Baseline v/s Proposed Approaches & Errors:** Compared to baseline, algorithm 1 and 2 improves results of large range objects by 1% and 4% respectively. In general we can see that the improvement obtained is directly related to distance of object from the camera. We obtain a average improvement of 13% mAP for Helmet, 12% mAP for Jacket, 1% for goggles and gloves each respectively.
- **Results across Detectors:** Among all the detectors with various posture conditions DCR-SSD, produces top results for Helmet and Goggle, while DCR-Faster-RCNN produces best result for Jacket and Glove. DCR-RetinaNet produces results very close to that of rest.
- **Results across varying parameters:** The results of frames with objects with large range is the least and with the small range is the highest. This is consistent across all the experiments. Finally, for objects with large range, we can see that algorithm 2 produces the maximum improvement in SSD with 8% mAP for helmet, 14% for jacket, 2% and 3% each for glove and goggle respectively.

D. Impact of Occlusion

Most of the construction environments are densely populated with large number of worker working on various aspects so involved. Further much of construction areas are filled with multiple equipment's. As such there is need of safety monitoring system that is not susceptible to occlusions, which is not the case of detectors, where most of them are susceptible to occlusion leading to missed detections. As such we create dataset to include occlusion and non-occluded workers. Such an example is also shown in Figure 1. In line with previous experiments, we evaluate across all the algorithms with all the DCR detectors to obtain following observations.

- **Results across Detectors:** As far as detector's performance goes, DCR-Faster-RCNN obtains the best performance for detecting objects with occlusion across the baseline, algorithm 1 and algorithm 2. Followed by DCR-SSD and DCR-RetinaNet.
- **Results across varying parameters:** For detecting non-occluded objects the results of RCNN and SSD is similar and RetinaNet is lower than both by average of 1% mAP. For occluded objects, we see an net improvement of 17% helmet and jacket, 0.02% for goggles and 0.01% for glove respectively and non-occluded objects we see an improvement of 6%, 14% and 1% for Helmet, Jacket and Glove classes. Moreover, we see that the improvement

TABLE VII
COMPARISON OF PERFORMANCE (MAP@ THRESHOLD=0.55) ACROSS APPROACHES WITH VARYING RANGE

Approch		**DCR-Faster-RCNN**				**DCR-SSD**				**DCR-RetinaNet**			
		H	**J**	**GO**	**GL**	**H**	**J**	**GO**	**GL**	**H**	**J**	**GO**	**GL**
Baseline	**Small**	0.7	0.59	0.75	0.64	0.69	0.54	0.75	0.67	0.66	0.59	0.75	0.64
	Medium	0.73	0.68	0.9	0.72	0.74	0.65	0.9	0.74	0.67	0.68	0.92	0.72
	Large	0.77	0.78	0.96	0.87	0.77	0.78	0.96	0.87	0.67	0.78	0.95	0.87
	Average	**0.74**	**0.67**	**0.87**	**0.75**	**0.74**	**0.64**	**0.87**	**0.77**	**0.67**	**0.67**	**0.87**	**0.75**
Algorithm 1	**Small**	0.7	0.59	0.77	0.66	0.72	0.56	0.77	0.7	0.66	0.59	0.75	0.64
	Medium	0.73	0.7	0.92	0.74	0.76	0.65	0.92	0.74	0.67	0.68	0.92	0.72
	Large	0.77	0.77	0.98	0.87	0.78	0.78	0.98	0.89	0.67	0.78	0.95	0.87
	Average	**0.74**	**0.68**	**0.89**	**0.76**	**0.76**	**0.66**	**0.89**	**0.78**	**0.67**	**0.67**	**0.87**	**0.75**
Algorithm 2	**Small**	0.77	0.69	0.77	0.69	0.77	0.68	0.77	0.7	0.69	0.67	0.75	0.64
	Medium	0.88	0.77	0.92	0.75	0.88	0.77	0.92	0.75	0.79	0.75	0.92	0.72
	Large	0.9	0.92	0.98	0.87	0.91	0.87	0.98	0.89	0.89	0.92	0.95	0.87
	Average	**0.86**	**0.8**	**0.89**	**0.77**	**0.87**	**0.78**	**0.89**	**0.78**	**0.81**	**0.78**	**0.87**	**0.75**

obtained with occlusion is higher than that of objects without occlusion.

VII. DISCUSSION AND CONCLUSION

In this work, we present our approach of re-id conditioned detector for safety gear detection. which we began by creating a large dataset of 5k images with varying characteristics as shown in section I. We then created a large scale benchmark in section V-A using DCR detection from section III-B, where we obtained lower results due missed detection and wrong classification. To mitigate this we proposed an extension of typical sequential detector that uses re-identification information from multiple frames through tracker for improving detection and reducing miss classification in section 2. We investigated the proposed approaches with multiple experiments where we first created benchmarks with DCR approaches and sequential detectors without re-identification. Following this we did a broad evaluation keeping robustness of the detectors for various conditions in mind. We first begin with understanding impact of illumination in section VI-A to cater usage of developed system across different environments, where we saw that the proposed approach gives average improvement of 9% for Helmet, 13% for Jacket, 3% for glove with goggle showing a minor drop of 1%. We also saw more improvement in case of darker illumination. Following this in section VI-B we verified performance improvement for various postures to see that the net improvement is significantly lower for bending and sitting postures. We also investigated impact of range in section VI-C and impact of occlusion in section VI-D to see that the proposed approach improves the results in these cases as well. Further across all the experiments we see improvements when using proposed approach and limited to no improvement when using baseline and Agorithm 1. At the same time, we can see that results for Faster-RCNN and SSD are quite similar across the experiments leading to choice of using the said approach as needed. While one would argue, that Faster-RCNN alone is sufficient we think that the proposed approach with SSD has better memory and performance. Overall we addressed the issue of missed and wrong detections produced by detectors with re-identification conditioned detector.

While we addressed the re-identification conditioned detector, there are multiple result which we didn't explore in details. To begin with, in Table V we didn't analyze in detail the source behind the drop of performance for goggles and in Table VI we conjectured higher performance of RetinaNet due to feature pyramids, which need further experimentation. Further across all the experiments, gloves had the least results owing to wrong detection, which needs to be addressed. Further, instead of DCR based detectors, we can explore cascaded detectors which are robust to various environmental changes.

REFERENCES

[1] Schneider, Steffen and Pam Susi. Ergonomics and construction: a review of potential hazards in new construction. American Industrial Hygiene Association journal 55 7 (1994): 635-49 .

[2] Bureau of Labor Statistics, Construction, NAICS 23, (2017).

[3] 48,000 die due to occupational accidents yearly: Study, The times of India, 2017. In https://timesofindia.indiatimes.com/business/india-business/48000-die-due-to-occupational-accidents-yearly-study/articleshow/61725283.cms

[4] Jeong, B.Y. (1998). Occupational deaths and injuries in the construction industry. Applied ergonomics, 29 5, 355-60 .

[5] Konda, S., Tiesman, H.M., Reichard, A.A. (2016). Fatal traumatic brain injuries in the construction industry, 2003-2010. American journal of industrial medicine, 59 3, 212-20.

[6] Colantonio, A., McVittie, D., Lewko, J. H. Yin, J. Traumatic brain injuries in the construction industry. Brain injury 23 11, 873-8 (2009).

[7] Dolan, E. Kriz, P. K. Protective Equipment, (2017).

[8] AERB SAFETY GUIDELINES, PERSONAL PROTECTIVE EQUIPMENT , Bureau of Indian Standards. In https://tinyurl.com/y4xmwthh

[9] Naticchia, B., Vaccarini, M. Carbonari, A. A monitoring system for real-time interference control on large construction sites. Automation in Construction 29, 148-160 (2013).

[10] Cai, Z., Vasconcelos, N. (2018). Cascade R-CNN: Delving Into High Quality Object Detection. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6154-6162.

[11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015.

[12] Leal-Taix, L., Milan, A., Reid, I.D., Roth, S., Schindler, K. (2015). MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. ArXiv, abs/1504.01942.

[13] Qi Fang, Heng Li, Xiaochun Luo, Lieyun Ding, Hanbin Luo, Timothy M. Rose, Wangpeng An, Detecting non-hardhat-use by a deep learning method from far-field surveillance videos, Automation in Construction, Volume 85, 2018, Pages 1-9

[14] Detection of construction workers under varying poses and changing background in image sequences via very deep residual networks Hyojoo Son, Hyunchul Choi, Hyeonwoo Seong, Changwan Kim Pages 27-38, Automation in Construction 99, 27 (2019).

TABLE VIII
COMPARISON OF PERFORMANCE (mAP@ THRESHOLD=0.55) ACROSS APPROACHES WITH VARYING OCCLUSION

Approch		**DCR-Faster-RCNN**				**DCR-SSD**				**DCR-RetinaNet**			
		H	**J**	**GO**	**GL**	**H**	**J**	**GO**	**GL**	**H**	**J**	**GO**	**GL**
Baseline	**Occlusion**	0.72	0.61	0.82	0.69	0.67	0.61	0.82	0.69	0.62	0.62	0.82	0.69
	No-Occlusion	0.83	0.68	0.96	0.8	0.83	0.68	0.96	0.8	0.74	0.7	0.96	0.8
	Average	**0.77**	**0.64**	**0.9**	**0.75**	**0.75**	**0.64**	**0.9**	**0.74**	**0.68**	**0.68**	**0.9**	**0.75**
Algorithm 1	**Occlusion**	0.76	0.62	0.84	0.69	0.68	0.62	0.84	0.69	0.62	0.63	0.84	0.69
	No-Occlusion	0.84	0.68	0.96	0.81	0.84	0.68	0.96	0.8	0.76	0.71	0.96	0.81
	Average	**0.79**	**0.65**	**0.91**	**0.76**	**0.76**	**0.66**	**0.91**	**0.74**	**0.69**	**0.69**	**0.91**	**0.76**
Algorithm 2	**Occlusion**	0.86	0.79	0.84	0.72	0.87	0.79	0.84	0.69	0.81	0.77	0.84	0.69
	No-Occlusion	0.88	0.84	0.96	0.82	0.87	0.84	0.96	0.8	0.84	0.82	0.96	0.81
	Average	**0.87**	**0.81**	**0.91**	**0.77**	**0.87**	**0.81**	**0.91**	**0.74**	**0.83**	**0.79**	**0.91**	**0.76**

[15] Fang W., Ding L., Zhong B., Love P.E., Luo H. Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach. Adv. Eng. Inform. 2018;37:139149. doi: 10.1016/j.aei.2018.05.003

[16] S. Du, M. Shehata, and W. Badawy, Hard hat detection in video sequences based on face features, motion and color information, 3rd International Conference on Computer Research and Development, 2011

[17] Hard-Hat Detection for Construction Safety Visualization Kishor Shrestha, Pramen P. Shrestha,Dinesh Bajracharya and Evangelos A. Yfantis, Journal of Construction Engineering, 2015

[18] Automatic Detection of Helmet Uses for Construction Safety, IEEE/WIC/ACM International Conference on Web Intelligence Workshops,2016

[19] Z. Zhu, M.W. Park, and N. Elsafty, Automated monitoring of hardhats wearing for onsite safety enhancement,International Construction Specialty Conference of the Canadian Society for Civil Engineering (ICSC), 2015.

[20] Automated Hardhat Detection for Construction Safety Applications BE Mneymneh, M Abbas, H Khoury Procedia Engineering 196, 895-902

[21] M.W. Park, N. Elsafty, and Z. Zhu, Hardhat-Wearing Detection for Enhancing On-Site Safety of Construction workers, Journal of Construction Engineering and Management 141, 04015024 (2015).

[22] R. Mosberger, H. Andreasson, and A. Lilienthal, Sensors 14, A Customized Vision System for Tracking Humans Wearing Reflective Safety Clothing from Industrial Vehicles and Machinery, 17952 (2014).

[23] M.-W. Park and I. Brilakis, Construction worker detection in video frames for initializing vision trackers, Automation in Construction, (2012).

[24] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the Conference on Computer Vision and Pattern Recognition, pages 580587, 2014.

[25] Ross Girshick. Fast r-cnn. In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), pages 14401448, 2015.

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards realtime object detection with region proposal networks. In Advances in Neural Information Processing Systems, pages 9199, 2015.

[27] Paul Viola and Michael Jones. Robust real-time face detection. International Journal of Computer Vision, 57(2):137154, 2004.

[28] Pierre Sermanet, David Eigen, Xiang Zhang, Michal Mathieu, Robert Fergus, and Yann Lecun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In Proceedings of the International Conference on Learning Representations (ICLR), April 2014.

[29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 779788, 2016.

[30] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, ChengYang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In European Conference on Computer Vision, pages 2137. Springer, 2016.

[31] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[32] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. DSSD : Deconvolutional Single Shot Detector. arXiv:1701.06659, 2017.

[33] Yi Li, Kaiming He, Jian Sun, et al. R-FCN: Object detection via region-based fully convolutional networks. In Advances in Neural Information Processing Systems, pages 379387, 2016

[34] Bichen Wu, Forrest Iandola, Peter H. Jin, and Kurt Keutzer. SqueezeDet: Unified, Small, Low Power Fully Convolutional Neural Networks for Real-Time Object Detection for Autonomous Driving. In CVPR Workshops, 2017.

[35] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In ICCV. IEEE, 2017

[36] Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang. Seqnms for video object detection. arXiv:1602.08465, 2016

[37] Subarna Tripathi, Zachary C Lipton, Serge Belongie, and Truong Nguyen. Context matters: Refining object detection in video with recurrent neural networks. arXiv:1607.04648, 2016.

[38] Kai Kang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Object detection from video tubelets with convolutional neural networks. In Computer Vision and Pattern Recognition, 2016.

[39] Kai Kang, Hongsheng Li, Tong Xiao, Wanli Ouyang, Junjie Yan, Xihui Liu, and Xiaogang Wang. Object detection in videos with tubelet proposal networks. In Computer Vision and Pattern Recognition, 2017.

[40] Yongyi Lu, Cewu Lu, and Chi-Keung Tang. Online video object detection using association lstm. In ICCV, 2017.

[41] M Andriluca, S. Roth, and B. Schiele. People-tracking-by-detection and peopledetection-by-tracking. In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[42] J. Berclaz, F. Fleuret, and P. Fua. Robust people tracking with global trajectory optimization. In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2006.

[43] Michael Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool. Online multi-person tracking-by-detection from a single, uncalibrated camera. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(9), 2010

[44] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided featur aggregation for video object detection. In ICCV. IEEE, 2017.

[45] Xizhou Zhu, Jifeng Dai, Lu Yuan, and Yichen Wei. Towards high performance video object detection. In CVPR. IEEE, 2018.

[46] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In ICCV, 2017.

[47] Broad, A., Jones, M.N., Lee, T. (2018). Recurrent Multi-frame Single Shot Detector for Video Object Detection. BMVC.

[48] Cheng, B., Wei, Y., Shi, H., Feris, R.S., Xiong, J., Huang, T.S. (2018). Revisiting RCNN: On Awakening the Classification Power of Faster RCNN. ArXiv, abs/1803.06799.

[49] Everingham, M., Gool, L.V., Williams, C.K., Winn, J.M., Zisserman, A. (2009). The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision, 88, 303-338.

[50] Ketaro Wada, labelme: Image Polygonal Annotation with Python, 2016