Object Detection and Occluded Face Recognition: A Study of Multistage Decoupling, Re-ID and Reference Conditioning

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

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

***Abstract*—Object detection and face recognition are an integral part of most surveillance systems in recent times especially in the areas of manufacturing, construction surveillance, etc. In this work, we address these problems by proposing practical improvements to existing state-of-the-art methods in multiple phases with a focus on a plurality of application-level requirements. As the first contribution, we propose to improve the performance efficiency of fine-grained object detection through Multistage Decoupling and Re-ID conditioning with former reducing miss classification by mitigating the correlation-covariance issue and latter reducing miss/wrong detection through re-identification conditioned box propagation. As a second contribution, we focus on improving face recognition under cases of occlusions. Occlusion in faces results in non-isotropic clusters causing high errors in face recognition, which is further increased due to lack of availability of large datasets for novel environments. To handle a lack of availability of large occluded face datasets and enhancing face representations from existing state-of-the-art face recognition models to a setting involving occluded faces with small datasets, we propose to learn reference conditioned projections, that projects the extracted representations into the lower-dimensional manifold and are isotropic. Finally, we show the advantages of these contributions through quantitative evaluations of multiple state-of-the-art datasets and approaches across multiple application domains, with the first one leading to improved performance of object detection and second contribution paving way for highly accurate face recognition. In the due process, we also present new datasets, show comprehensive experiments and practical advice from our extensive empirical results for those interested in getting the most out of this work, for developing real-world surveillance systems.**

***Index Terms*—Convolutional Neural Network, Deep Learning, Object Detection, Face Recognition**

I. Introduction

Visual recognition technologies such as object detection, i.e. predicting the exact location of a given object in an image with a bounding box and face recognition i.e. identification of people from images have graduated from proof of concept level to being productized. Starting from the seminal work by [1], proposing the first high performance deep convolutional neural network (CNN) for image classification, the algorithms and systems have very far, with better performing networks for image classification, such as the GoogLeNet [2], VGGNet [3] and ResNet [4], as well as networks for object detection, such as Fast-RCNN, Faster-RCNN [5], R-FCN [6], SSD and its variants [7], [8] , YOLO and its variants [9], RetinaNet [10] and very recently Mask-RCNN [11]. A similar track could see in face recognition with state-of-the-art approaches including [7] and [8] also follow such a strategy. Lately, all the CNN based approaches have shown excellent performance in various face recognition tasks, notable among them include [9], [10] which achieves an accuracy-rate of 99.78% and 99.63% on the facial dataset of Labeled Faces in the Wild (LFW) [11]. These high performing object detection and face recognition modules now find applications in upcoming autonomous cars, airport, industrial and manufacturing surveillance, etc.

II. Challenges

Even after the huge success of deep learning with high accuracies, a plethora of approaches fail to come through in stage of novel applications ex: in the areas of construction, industrial surveillance, this is despite all the additional bells and whistles. More specifically they exhibit the following challenges

1) To begin with, object detection which is the principal component of any surveillance application, suffers from the problem of false/missed detections and miss classifications which are extremely costly resulting in safety hazards [12].
2) Secondly, video stream based processing to reduce detection errors is yet to be adapted for applications like industrial and construction safety, despite mainstream adaptation, in general, AI tasks [13], [14].
3) Finally, unlike existing areas of computer vision, where there is the availability of large public datasets, the best of existing research work in the applied areas of surveillance is restricted due to lack of large datasets.

On the other hand, compared to object detection general face recognition has reached near 100% accuracy across a plurality of benchmarks, however to date, it remains difficult

* Both authors contributed equally. Order is decided based on coin toss.

to obtain satisfactory accuracy on faces varying in pose, illumination, and occlusion, among which facial occlusion has always been considered as an extremely challenging research problem. While general face recognition has garnered a lot of works over the past five years. Interests in recognition of faces with occlusions have been fairly less, with limited focus on generalization study of state-of-the-art face recognition models for occluded faces. More specifically following are some of the challenges in occluded face recognition

4) To begin with, there is neither the availability of large datasets consisting of multiple identities with facial occlusion, for training existing deep CNN face recognition models from scratch nor is there a dataset sample that is sufficient enough for using the classical domain adaptation techniques involving finetuning, mapping functions, etc.
5) Secondly on the applied side say, in the case of face recognition in industrial scenes, there is a need to handle occlusion of various types involving thicker and darker eye protection glasses (welding glasses) that cover most parts of the face. However, existing face recognition algorithms show the problem of non-isotropic clustering especially with the introduction of occlusion.
6) Finally, due to previously mentioned reasons there exist no benchmarks that validate the state-of-the-art approaches proposed to date.

III. Contribution

Based on the above, through this work, we make the following contributions.

- Firstly to tackle **#3** and **#4** we introduce i) Safety gear detection dataset consisting of 5k images with annotations for four safety gears including helmet, jacket, goggle and glove and ii) Synthetically created occluded face datasets based on popular public datasets. Further, we present a large scale benchmark study on these using the current state-of-the-art approaches.
- Second, to address **#1** and **#2** we propose to condition the class label of the detected object in the current frame based on the results of the re-identification of objects computed in prior frames. Such a conditioning mechanism helps in the propagation of class labels, thereby reducing the miss classification. Further in this work, it's also supplemented with decoupled classification refinement. Thereby extending single frame detectors into video domain.
- Third, to tackle **#5** we propose a reference conditioned low-rank projection algorithm.
- Finally, we present comprehensive experimentation and benchmarking involving 100+ experiments where we achieve improvements in performance across all of the previously mentioned datasets.

IV. Related Work

Computer Vision for Industrial Safety Systems: Construction safety monitoring has seen a large number of works mainly focusing on helmets and jackets with Fang et al. [15] and Son et al. [16] presenting works based on Faster-RCNN and improving its generalization. Then there are works by Du et al. [18] using motion and color information which reduces the applicability of such a method in actual scenarios and similar works are also proposed by Shrestha et al. [19] which uses edge information. Then there are works using Histogram of Oriented Gradient (HOG) with Circle Hough Transform (CHT) [20], HOG feature template of a human object add cite here, cascade classifiers [21] and color information [22]. On the other side jackets have seen some works including those of Morsberger et al. [23] which proposes using combination of segmentation, localization and classification, then Park et al. [24] that uses fluorescent color of safety vests by processing local color histograms 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: Object detection is typically dominated by CNN's beginning with series of works by Girshick et al. on region proposal networks [25]–[27] then Single Shot Detectors [29]–[35] Both region proposal style detectors, and single-shot detectors, are fast and reasonably accurate. However surveillance situation has streaming of frames, thus treating frame-level input is of limited use. Yet there are few works on video object detection broadly under box-based approaches [37] and feature-based approaches [36] which are as explained. Box based approaches act on the sequence of bounding boxes from single frame detectors with some sort of linking strategy, notable approaches include those by Han et al. [37] that replaces standard Non-Maximal Suppression (NMS) with one that incorporates bounding boxes from multiple frames and Tripathi et al. [38] that uses an RNN. Then there are tubelets and feature map-RNN combination by Kang et al. [39], [40] & Lu et al. [41]. All these approaches are related to another class of box-level techniques known as tracking-by-detection [42]–[44] 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. Then there are few approaches that directly link at feature level initial work Zhu et al. [45] uses optical flow to warp feature maps Friechtenhofer et al. [62] uses a deep network to combine detection and tracking to improve object detection and [46] combine two orthogonal ideas with a spatially adaptive feature computation to further improve results. Overall 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 a re-identification conditioned sequential detector with the tracker to handle the problem of missed and incorrect detections. Furthermore, we address classification errors through a combination of decoupled classification refinement and re-identification conditioned sequential detector.

Occluded Face Recognition: Face recognition has two

decades of works with and most recently with [47] and then followed by works of [48] which proposed an end-to-end Siamese architecture trained with a contrastive loss function. Multiple followup works build on similar lines, however with a large amount of data [49]. Most notable among these includes DeepFace [50] and FaceNet [51] which used between 100 million and 200 million face images of about 8 million different people for training using a triplet loss function and achieved an accuracy of 99.63% on the LFW benchmark. While general face recognition has garnered a plethora of works over the past five years*. Interests in recognition of faces with occlusions have been fairly limited and typically handled using two different types of methods namely local features from non-occluded regions or reconstruction based ideas. For the former Gabor wavelet features, PCA and SVM were used in [52] to detect the occluded regions and LBP descriptors were used to match the non-occluded regions and the sparse representation-based classification (SRC) proposed in [53] has received a lot of attention in the latter works. Moreover, all these approaches still require a large amount of occluded face data sets for training. Unlike these approaches, we focus on adapting representations from state-of-the-art face recognition models trained on large data sets of general face recognition to occluded face recognition through a small representative set of occluded faces.

Projecting Representations: Learning to project representations from source to target domain as seen lots of works in computer vision. Notable works include [54] for joint representation learning and [55] proposing a Bishifting Auto-Encoder network. Then there most recent works of [56] and [57] which uses projection for surveillance face recognition. Our work is similar to previous works, in the sense that we too consider front faces from public datasets such as LFW as the source domain and images from another dataset with different identities that are occluded by goggles as the target domain. However, unlike these approaches, we focus on learning the projection model with a small representative set of occluded faces.

The rest of the paper is organized as follows. In section V we present the developed datasets with experimental setup used in this work. In section VI we brief on decoupled classification refinement, with VIII and VII presenting proposed algorithms. Sections IX & X presents experiments and results. Finally, in section XII, we conclude with discussions and possible implications for future works. Please note that this paper is a representative version of original works [58] & [59] , which explains the ideas, algorithms and experiments in more detail.

V. Datasets

A. Occluded Face Recognition

Typically public benchmark datasets are either web scraped or laboratory collected with different ages, poses, illumination, expressions, facial hair, and other occlusions. Yet to the best of our knowledge there exist no public datasets with goggle occluded faces and those which exist either is not explicitly collected for such a setting or requires spurious licensing. Hence, we create synthetic goggled face images of the public datasets, by augmenting glasses on the faces through keypoint identification. Multiple augmented public datasets that have been used in this work are shown in Table I along with the size of the datasets and the extent of variation. Figure 1 shows examples of goggled versions of the datasets so created.

TABLE I
Dataset statistics of various face datasets

Dataset	No of identities	No of images
AT&T	40	400
ESSEX	394	7900
FEI	200	2800
GT	50	750
CFP	311	5000
LFW	1600	13000
GOFD	5	1000

Moreover, to evaluate our approach in a practical setting we also collected an in house **Goggle Occluded Face dataset (GOFD)**. We place the following constraints as part of the dataset collection setting

- Goggled Frontal Face, with varying occlusion levels involving transparent and opaque goggles.
- Maximum of 45 degrees of horizontal variation in face pose.
- Images are collected from cameras placed at a moderate inclination upto 7 feet.

*1) **Reference and Test Images**:* All the goggled datasets are split into two parts namely a reference a.k.a gallery set and test a.k.a probing set. The reference set consists of 3 samples of images for each identity in the entire dataset. Moreover, we don't produce any additional reference images through augmentation during testing. The reference images were sampled randomly to avoid any bias on the results.

*2) **Support Images**:* As seen above, we use only 3 reference image per identity and the largest of the benchmark datasets [60] consists of only 1600 identities. Moreover, for projection approaches to work well, there is a need for large datasets. As such in this work, we use a support set which we concatenate with the smaller reference sets in reference conditioning step of our algorithm (See section VIII. In this work, we use the LFW dataset (See section V) as the support set. While the description of these sets, looks complex as we can see (see VIII) the algorithm is intuitive and straightforward.

The dataset statistics including reference, test and sample images for GOFD are as shown in Table I and Figure 1 respectively.

B. Safety Gear Detection

Since there is no off-the-shelf dataset available, a dataset was created in an 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. The dataset was collected with varied illumination, posture and visual ranges

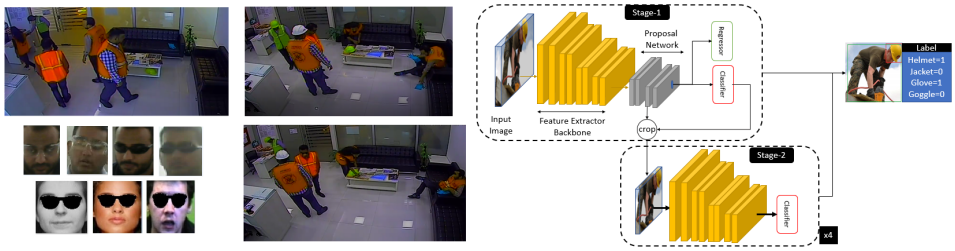

Fig. 1. Example images from datasets and network training strategy used in this work.

with annotations for boxes and class labels. The final dataset statistics are as shown in Table VI. 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 figure 1.

VI. 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. [61], we propose to use a Multistage Decoupled Classification Refinement (DCR) detector. Diagrammatic representation of the same is shown in Figure 1.

VII. 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 [42]–[44] 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 $B_t = \textbf{DetectOnImage}(I_t)$ 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 $D_t = \textbf{AssociateTracklet}(D_{t-1}, B_t)$ outputting $\{B_t\}$. Additionally, Box propagation is applied, where detected boxes in the existing tracklets D_{t-1} are propagated to the current frame, $B_t^{'} = \textbf{PropagateBoxKalman}(D_{t-1})$. 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:
$B_0 = \textbf{DetectOnImage}(I_0)$
Initialize the tracklets D_0 from B_0
for $t = 1$ *to* T
$B_t = \textbf{DetectOnImage}(I_t)$
$B_t^{'} = \textbf{PropagateBoxKalman}(D_{t-1})$
$B_t = [B_t^{'}, B_t]$:= *Box Concatenation*
$B_t = NMS(B_t)$
$D_t = \textbf{AssociateTracklet}(D_{t-1}, B_t)$
end for

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.

VIII. Reference Conditioned Low Rank Projection

A. Background

Typically across multiple computer vision problems the resulting drop due change in environment is argued as a problem of difference in data domains a.k.a generalization

TABLE II
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

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

error [54]. While there are a plethora of approaches proposed [55] for domain adaptation, most of these are data-intensive and for many environments face data is extremely scarce owing to the issue of privacy and labor involved in data collection. As such these approaches are not suitable for our problem in this work we propose a simple adaptation technique for face representations extracted from a pretrained network that can adapt with small samples (3 samples/face identity) from the target domain (goggled faces) based on following intuitions.

- Projecting data into a lower-dimensional subspace removes noisy dimensions and makes it isotropic.
- Projection approaches aim to find the directions of maximum variance in high-dimensional data and projects it onto a new subspace with equal or fewer dimensions than the original one. As such, it would capture information even from samples that have the smallest representation among the population used for learning projections.

Overall, projecting the representations to lower-dimensional manifold conditioned on reference images i.e. considering the information from reference image is preserved should improve the performance of the face representation for goggle occluded faces.

B. *Algorithm*

Let $S(i)\epsilon\Re^P$ and $T(i)\epsilon\Re^P$ be representations of the large face data sets that don't contain any goggle occluded faces and small sample of goggle occluded reference faces respectively obtained from pretrained models.

Let $X(i) = \{x_1, x_2, ...x_M, x_{M+1}, ...x_{M+K}\} = S||T$ be the matrix of dimension (M+K)xP consisting of M+K representations so obtained through vertical concatenation of S and T. Originally from section VIII-A $T(i)$ is non-isotropic. We model low rank projection as a q-rank linear model denoted by its rank-q affine hyper-plane as.

$$f(\lambda) = \mu + V_q\lambda \tag{1}$$

where μ represents mean of T of dimension $\Re^p$ and V_q is a PxQ matrix.

Fitting such a model is done by objective

$$\underset{\mu,\lambda_i,V_q}{min} \sum_{1}^{M+K} ||x_i - \mu - V_q\lambda_i||^2 \tag{2}$$

We partially optimize this using Equation 1, leading to minimizing V_q through

$$\underset{V_q}{min} \sum_{1}^{M+K} ||(x_i - \bar{x}) - V_q V_q^T(x_i - \bar{x})||^2 \tag{3}$$

which we solve using Singular Value Decomposition.

Algorithm 3: Reference Conditioned Low Rank Projection

Input: Face representations $S\epsilon\Re^d$, $T\epsilon\Re^d$, $X\epsilon\Re^d$
Output: Processed Representations $X^{'}\epsilon\Re^d$
Procedure:

- **Reference Conditioning:** Concatenate representations from source and target set $X = S||T$
- **Optimization:** Minimize $\underset{\mu,\lambda_i,V_q}{min} \sum_1^{M+K} ||x_i - \mu - V_q\lambda_i||^2$
- **Projecting Representations:** $X^{'} = V_q X$

TABLE III
RESULTS (RANK-1 ACCURACY %) OF BASELINE EXPERIMENTS USING VGGFACE, ARCFACE AND LIGHTCNN ON PUBLIC BENCHAMRK DATASETS WITH OCCLUDED REFERENCE IMAGES

Dataset	VGGFace	LightCNN	MobileFaceNet ArcFace	ResNet ArcFace
ATT	85.13	99.32	97.98	99.67
ESSEX	90.71	96.25	93.62	98.72
FEI	75.23	92.1	84.28	99.29
GT	85.47	90.13	89.25	99.85
CFP	32.02	63.53	48.87	87.26
LFW	22.42	54.92	49.08	88.14

TABLE IV
RESULTS (RANK-1 ACCURACY %) OF BASELINE EXPERIMENTS USING VGGFACE, ARCFACE AND LIGHTCNN ON PUBLIC BENCHAMRK DATASETS WITH NON-OCCLUDED REFERENCE IMAGES

Dataset	VGGFace	LightCNN	MobileFaceNet ArcFace	ResNet ArcFace
ATT	53.51	91.23	91.99	99.33
ESSEX	62.03	90.8	84.56	98.28
FEI	56.70	93.6	84.72	99.63
GT	86.05	91.48	86.11	99.69
CFP	17.52	60.12	41.42	86.7
LFW	8.09	53.06	47.36	87.26

IX. OCCLUDED FACE RECOGNITION EXPERIMENTS

A. Experiment 1: Baselines:

To begin with, we first created an exhaustive baseline to evaluate generalization performance using representations from pretrained models for goggled public face datasets. For this experiment, all the pretrained models are used as it is, without any modification to obtain the following results.

Based on the results from Table III, we summarize as follows.

- Most of the representation previously trained on huge datasets, tend to perform very badly across the other datasets with goggle occluded faces due to inherent dataset bias.
- The face representation so generated is non-isotropic leading to bad performance in face recognition, especially those that involve simple distance-based metrics such as cosine and euclidean.

With these observations in mind, in section VIII we present experiments on RLCP.

B. Experiment 2:

Following baselines, experiments are performed on RLCP where we first execute reference conditioning and optimization using 3. After this, each of the test images is re-projected using the learned projection model of that particular dataset. For example, for the AT&T dataset, we first do reference conditioning by concatenating the face representations of reference images (See Table I) and images from the support set which is the LFW dataset (See section V), which in turn is used to learn the projection model, used for projecting representation in test time. Finally, a distance metric between the projected reference and test representations is calculated to obtain a correct face identity. Results so obtained are as shown in Table III. Comparing Tables III and III we can see that

- RLCP improves performance, as visible comparing results over baselines and maximum improvement are visible on the pretrained VGGFace model and less when representations are from the pretrained ArcFace model.
- Moreover, we can see a maximum improvement of 1.2% for AT&T, 1.3% for Essex dataset, 2.1% for the CFP dataset, 4.2% for FEI and 0.92 % for LFW datasets on VGGFace and 0.8% for GT dataset on LightCNN.
- Further, we can see that there is a limited improvement in results, using representations from ResNet-Arcface except 0.1% improvement on the results of FEI dataset. This is mainly, because by default the results of Arcface are high compared to that of other pretrained models, leaving limited room for improvements.

TABLE V
RESULTS (RANK-1) OF BASELINE EXPERIMENTS ON GOFD DATASET WITH OCCLUDED AND NON-OCCLUDED FACES.

Dataset	VGGFace	LightCNN	MobileFaceNet ArcFace	ResNet ArcFace
GOFD-OCR	86.81	99.51	85.08	98.57
GOFD-NOCR	69.76	91.82	62.76	99.75

TABLE VI
RESULTS (RANK-1 ACCURACY %) OF EXPERIMENTS USING REFERENCE CONDITIONED LOW RANK PROJECTION WITH OCCLUDED REFERENCE IMAGES.

Dataset	VGGFace	LightCNN	MobileFaceNet ArcFace	ResNet ArcFace
ATT	86.31	99.32	97.98	99.67
ESSEX	92.00	96.25	96.32	98.72
FEI	79.49	92.3	85.11	99.38
GT	85.47	90.97	89.25	99.85
CFP	34.1	63.53	48.87	87.26
LFW	23.34	54.92	49.08	88.14

C. Experiment 3:

Previous experiments we used one reference image for each face identity in the dataset with occlusion. While this setup is ideal, in most real-time applications due to lack of availability of goggle occluded reference image we need to use a standard face image without any occlusion, that is collected as part of the identification process or part of the database of people or so on. Considering this we experimentally verify our developed approach under this setting, to obtain results as shown in Table IV and VIII respectively. In Table IV, we find the baselines and in Table VIII we show improvement, with both cases using non-occluded reference face images. The reference image selected is in line with that of experiment X-A. Comparing Table IV, VIII and results from experiment X-A we can infer following:

- The results of experiment 2 is lesser than that of experiment 1 by an average of 10%. This is true in both the cases of baselines and the improvements so obtained. We speculate that the drop is because of the resulting miss-alignment in sub-spaces.
- Compared to baselines under the same setting we see maximum improvement of 2.4% for AT&T using Light-CNN,2.3% for Essex using Light-CNN, 0.1 % for FEI on

TABLE VII
RESULTS (RANK-1 ACCURACY %) OF EXPERIMENTS USING REFERENCE CONDITIONED LOW RANK PROJECTION ON GOFD DATASET WITH OCCLUDED AND NON-OCCLUDED FACES.

Dataset	VGGFace	LightCNN	MobileFaceNet ArcFace	ResNet ArcFace
GOFD-OCR	91.13	100	86.1	98.81
GOFD-NOCR	77.89	93.02	70.26	99.75

TABLE VIII
RESULTS (RANK-1 ACCURACY %) OF EXPERIMENTS USING REFERENCE CONDITIONED LOW RANK PROJECTION WITH NON-OCCLUDED REFERENCE IMAGES.

Dataset	VGGFace	LightCNN	MobileFaceNet ArcFace	ResNet ArcFace
ATT	54.76	93.63	92.78	100
ESSEX	62.03	93.16	84.93	98.28
FEI	56.70	93.6	84.81	99.7
GT	86.05	92.82	86.78	99.85
CFP	17.52	60.12	41.42	86.7
LFW	8.09	53.06	47.36	87.26

MobileNetArcFace, 1.4% on LightCNN pretrained face recognition models.

- In line with experiment 1, we can see limited improvement in results of ArcFace and maximum improvement in the case of VGGFace. Overall the proposed approach has an average improvement of 1.29% over baseline which is 0.52% higher than that obtained in experiment X-A.

D. Experiment 4:

While we obtained results on synthetically occluded versions of these datasets in section X-A and IX-C, real datasets are often complex with varying noise information such as reflections, etc. In that sense, we further compare the results on a realistic **GOFD** dataset. The results so obtained are as shown in Table V and VII. On comparing results we can see that

- The proposed approach improves results of all the pretrained models with an average of 1.51% with occluded reference faces and 5.6% with non-occluded faces.
- We obtained a maximum improvement of 8.13% and 4.32% with non-goggled reference and goggled reference images using VGG-Face representations.
- Again we obtain the least improvement on Arc-face as it produces high results.

X. RE-ID CONDITIONED DETECTION EXPERIMENTS

Previously in section VII we presented the re-identification conditioned detector with tracker algorithm. In this section, we present a detailed experiment on Re-ID conditioning with ablation study and results. The analysis and findings 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 an exhaustive evaluation in the context of safety gear detection using DCR detectors (section VI) and sequential detector with tracker algorithm (Algorithm 1), on the datasets discussed earlier in section VI. Firstly, we trained three DCR Detectors on the training sets, with each of the architectures tuned for various hyperparameters such as learning rate, etc. to ensure quicker convergence. The results so obtained are as shown in Table IX and X. The details of the analysis reveal the following.

- **Baseline v/s Proposed Approaches:** Comparing the baseline results in Table IX with results obtained using Algorithm 1 from Table X. From Table X we can see, that sequential detector does perform significantly better than the baseline detectors, by a margin of 2% mAP. Comparing Tables IX and X, we can see that this is mainly due to the increase in results of conditions involving occlusion and range. We will revisit analysis in upcoming section (see section XI) for each conditions in section XI.
- **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 IX. DCR-SSD and DCR-Faster-RCNN perform very similarly to DCR-Faster-RCNN across both Tables IX and X except DCR-Faster-RCNN performs well across illumination and occlusion by an average of 2% mAP. Further this observation is true across Tables IX and X. We believe this is because of the 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 a more detailed analysis of these results for future works.
- **Errors:** As far as the source of errors, in individual safety gears, are concerned, the insight so obtained from error analysis was common across 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

TABLE IX
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 X
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 XI
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**

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 XI for all the three detectors.

- **Baseline v/s Proposed Approaches:** Firstly comparing Tables XI with XI and IX, we see that the results are significantly higher across all the conditions of illumination, posture, range and occlusion.
- **Results across Detectors:** In line with observations from section X-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 XI,X and IX we can see that mAP improves by an 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 an average of 1%. Also, the results of gloves are very similar across the detectors and across the Tables XI,X and IX. This is because, the detectors so trained seldom detect gloves or gives the wrong detection, as such leading to higher error, compared to other safety gears.
- **Errors:** This is in line with, previously described observation.

XI. ABLATION STUDY

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

A. Impact of Illumination & Posture

Typically industrial environments have long works hours under varying lighting conditions (day, night, outdoor, indoor) with complex tasks involving carrying bricks, bending bars resulting in varying illumination and posture conditions. To test the sanity of the developed idea, we initially had collected datasets with a similar setting. Here we present experimental results and findings so obtained.

From experiments, we can make the following observations

- **Baseline v/s Proposed Approaches & Errors:** In case of posture, we obtain a net improvement of 1% and

TABLE XII
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**

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. In case of illumination comparing results across baseline, algorithm 1 and algorithm 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%.

- **Results across Detectors:** Among all the detectors with various posture conditions DCR-SSD, produces top results for Helmet and Goggle, while DCR-RetinaNet produces the 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 the impact of feature pyramids on detecting varying sized objects to future work. However, for under varying illumination, DCR-SSD, produces top results for Helmet and Goggle, while DCR-Faster-RCNN produces the best result for Jacket and Glove.
- **Results across the parameter:** Among all the results of varying posture, we see that results with standing posture are 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 the results of gloves are constant across all the experiments. Results under bright illumination are 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.

B. Impact of Range & Occlusion

Detectors are extremely sensitive to the size of objects, this especially true in case of indoor cameras and open cameras catering large visual ranges, especially due to stochastic placements. Further much of surveillance areas involve occlusion due to people and other objects. All these factors cause missed or wrong detections. To verify the robustness of the proposed approach, we study the impacts of the previously mentioned aspects. Experiments on this show following observations.

- **Baseline v/s Proposed Approaches & Errors:** Compared to baseline, algorithms 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 the distance of the object from the camera. We obtain an average improvement of 13% mAP for Helmet, 12% mAP for Jacket, 1% for goggles and gloves each respectively.
- **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. A similar trend can ve observed for ranges.
- **Results across varying parameters:** For detecting non-occluded objects the results of RCNN and SSD is similar and RetinaNet is lower than both by an average of 1% mAP. For occluded objects, we see a 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 obtained with occlusion is higher than that of objects without occlusion. The results of frames with objects with a large range are the least and the small range is the highest. This is consistent across all the experiments. Finally, for objects with a large range, we can see that algorithm 2 produces the maximum improvement in SSD with 8% mAP for the helmet, 14% for the jacket, 2% and 3% each for glove and goggle respectively.

XII. DISCUSSIONS & CONCLUSION

In this work we present our work on RCLP, Re-ID Condition with state of the art algorithms. We created large benchmark datasets for occluded face recognition and safety gear detection with exhaustive baselines. Following which we investigated the proposed approaches for improvements in results, errors, etc. In the process, we did exhaustive ablation study with varying parametric conditions of RLCP

TABLE XIV
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**

TABLE XV
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**

and Re-ID Conditioning. As far as occluded face recognition is concerned, we investigated the proposed approaches with multiple experiments in line with that of earlier benchmarks where we obtained an average improvement of 0.77% and maximum improvement of 4.25% on the FEI dataset which we can see from Table III. We further investigated the effect of using non-occluded reference images to show an average improvement of 1.29% and a maximum of 2.46% across the synthetically created occluded face datasets. To verify the sanity of the proposed approach on realistic use-cases, we further verified the proposed approach on the GOFD dataset where we obtained a maximum improvement of 8.13% and 4.32% respectively when using non-occluded and occluded faces respectively. We further investigated the Re-DI conditioning 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 of illumination, posture, range, and occlusion. While we addressed the aforementioned issues through low-rank projection, we think there is still a possibility of improving the results, especially with a large gap existing between VGGFace and ArcFace. In our experiments, we could see that the Rank-5 accuracy metric is almost 100% which suggests combining re-ranking and low-rank projection. Also, instead of the distance metric, a more robust approach of voting classifiers could be explored. Additionally, across all the experiments, gloves had the least results owing to the wrong detection, which needs to be addressed. Finally, instead of DCR based detectors, we can explore cascaded detectors which are robust to various environmental changes.

REFERENCES

[1] D. S. Bartoo, Financial services innovation: opportunities for transformation through facial recognition and digital wallet patents. (Dissertations Theses - Gradworks, Ann Arbor, 2013)

[2] M. Ketcham, N. Fagfae, The algorithm for financial transactions on smartphones using two-factor authentication based on passwords and face recognition. International Symposium on Natural Language Processing. (Springer, Cham, New York, 2016), pp. 223–231

[3] J. Xiao, Research on application of face recognition in area of public security. Comput. Sci. 43(11A), 127–132 (2016)

[4] D. Chawla, M. C. Trivedi. A comparative study on face detection techniques for security surveillance (Springer, Singapore, 2018), pp. 531–541

[5] Q. Zhao, M. Ye, The application and implementation of face recognition in authentication system for distance education. International Conference on NETWORKING and Digital Society. (IEEE, Los Alamitos, 2010), pp. 487–489

[6] D. Yang, A. Alsadoon, P. W. C. Prasad, et al., An emotion recognition model based on facial recognition in virtual learning environment. Procedia Comput. Sci. 125, 2–10 (2018)

[7] Y. Sun, D. Liang, X. Wang, et al., DeepID3: face recognition with very deep neural networks. Comput. Sci. abs/1502.00873 (2015)

[8] M. A. Hasnat, J. Bohn, J. Milgram, et al., vonMises-Fisher mixturemodel-based deep learning: application to face verification. arXiv: 1706.04264 (2017)

[9] R. Ranjan, C. D. Castillo, R. Chellappa, L2-constrained softmax loss for discriminative face verification. arXiv preprint arXiv: 1703.09507 (2017)

[10] F. Schroff, D. Kalenichenko, J. Philbin, et al., FaceNet: a unified embedding for face recognition and clustering[J]. Computer vision and pattern recognition, 815–823 (2015)

[11] L. J. Karam, T. Zhu, Quality labeled faces in the wild (QLFW): a database for studying face recognition in real-world environments[J]. Proceedings of SPIE - The International Society for Optical Engineering. 9394:93940B-93940B-10 (2015)

[12] 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.

[13] 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.

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

[15] 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

[16] 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).

[17] 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;37:139–149. doi: 10.1016/j.aei.2018.05.003

[18] 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

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

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

[21] 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.

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

[23] 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).

[24] 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).

[25] 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 580–587, 2014.

[26] Ross Girshick. Fast r-cnn. In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), pages 1440–1448, 2015.

[27] 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 91–99, 2015.

[28] Paul Viola and Michael Jones. Robust real-time face detection. International Journal of Computer Vision, 57(2):137–154, 2004.

[29] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël 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.

[30] 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 779–788, 2016.

[31] 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 21–37. Springer, 2016.

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

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

[34] 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 379–387, 2016

[35] 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.

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

[37] 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

[38] 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.

[39] 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.

[40] 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.

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

[42] 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.

[43] 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.

[44] 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

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

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

[47] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, recognition: A convolutional neural-network ap- proach," Neural Networks, IEEE Transactions on, vol. 8, no. 1, pp. 98-113, 1997.

[48] S. Chopra, R. Hadsell, and Y. LeCun, a sim- ilarity metric discriminatively, with application to face verification," in 2005 IEEE Computer Society Con- ference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 539-546, IEEE, 2005.

[49] A. Krizhevsky, I. Sutskever, and G. E. Hinton, - genet classification with deep convolutional neural networks," in Advances in neural information processing systems, pp. 1097-1105, 2012.

[50] Y. Taigman, M. Yang, M. Ranzato, and L.Wolf, Deep- face: Closing the gap to human-level performance in face verification," in IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 1701-1708, IEEE, 2014.

[51] F. Schro, D. Kalenichenko, and J. Philbin, : A unified embedding for face recognition and cluster- ing," in Proceedings of the IEEE Conference on Com- puter Vision and Pattern Recognition, pp. 815-823, 2015.

[52] R. Min, A. Hadid, and J. L. Dugelay, Improving the recognition of faces occluded by facial accessories," in 2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011), pp. 442-447, 2011.

[53] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, Robust face recognition via sparse represen- tation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 210-227, 2009.

[54] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474, 2014.

[55] Kan, M., Shan, S., Chen, X. (2015). Bi-Shifting Auto-Encoder for Unsupervised Domain Adaptation. 2015 IEEE International Conference on Computer Vision (ICCV), 3846-3854.
[56] Xie, X., Cao, Z., Xiao, Y., Zhu, M., Lu, H. (2015). Blurred image recognition using domain adaptation. 2015 IEEE International Conference on Image Processing (ICIP), 532-536.
[57] Banerjee, S., Das, S. (2016). Domain Adaptation with Soft-margin multiple feature-kernel learning beats Deep Learning for surveillance face recognition. ArXiv, abs/1610.01374.
[58] Shibashish Sen, Manikandan Ravikiran Enhancing Image Representations for Occluded Face Recognition via Reference Conditioned Low-Rank projection, 48th Annual IEEE AIPR 2019: Cognition, Collaboration, and Cloud Washington, D.C. October 15-17, 2019
[59] Manikandan, Ravikiran, Shibashish Sen, Improving Industrial Safety Gear Detection through Re-ID conditioned Detector, 48th Annual IEEE AIPR 2019: Cognition, Collaboration, and Cloud Washington, D.C. October 15-17, 2019
[60] Learned-Miller, E.G., Huang, G., Roy Chowdhury, A., Li, H., Hua, G. (2016). Labeled Faces in the Wild: A Survey.
[61] 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.
[62] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In ICCV, 2017.