Computers, Materials & Continua
DOI:10.32604/cmc.2022.024272
Article

Tech Science Press

Human Faces Detection and Tracking for Crowd Management in Hajj and Umrah

Riad Alharbey[1], Ameen Banjar[1], Yahia Said[2,3,*], Mohamed Atri[4], Abdulrahman Alshdadi[1] and Mohamed Abid[5]

[1]Department of Information Systems and Technology, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia
[2]Electrical Engineering Department, College of Engineering, Northern Border University, Arar, Saudi Arabia
[3]Laboratory of Electronics and Microelectronics (LR99ES30), Faculty of Sciences of Monastir, University of Monastir, Tunisia
[4]College of Computer Sciences, King Khalid University, Abha, Saudi Arabia
[5]CES Laboratory, ENIS, University of Sfax, Tunisia
*Corresponding Author: Yahia Said. Email: said.yahia1@gmail.com
Received: 11 October 2021; Accepted: 14 December 2021

Abstract: Hajj and Umrah are two main religious duties for Muslims. To help faithfuls to perform their religious duties comfortably in overcrowded areas, a crowd management system is a must to control the entering and exiting for each place. Since the number of people is very high, an intelligent crowd management system can be developed to reduce human effort and accelerate the management process. In this work, we propose a crowd management process based on detecting, tracking, and counting human faces using Artificial Intelligence techniques. Human detection and counting will be performed to calculate the number of existing visitors and face detection and tracking will be used to identify all the humans for security purposes. The proposed crowd management system is composed form three main parts which are: (1) detecting human faces, (2) assigning each detected face with a numerical identifier, (3) storing the identity of each face in a database for further identification and tracking. The main contribution of this work focuses on the detection and tracking model which is based on an improved object detection model. The improved Yolo v4 was used for face detection and tracking. It has been very effective in detecting small objects in high-resolution images. The novelty contained in this method was the integration of the adaptive attention mechanism to improve the performance of the model for the desired task. Channel wise attention mechanism was applied to the output layers while both channel wise and spatial attention was integrated in the building blocks. The main idea from the adaptive attention mechanisms is to make the model focus more on the target and ignore false positive proposals. We demonstrated the efficiency of the proposed method through expensive experimentation on a publicly available dataset. The wider faces dataset was used for the train and the evaluation of the proposed detection and tracking model. The proposed model has achieved good results with 91.2% of mAP and a processing speed of 18 FPS on the Nvidia GTX 960 GPU.

Keywords: Crowd management; Hajj and Umrah; face detection; object tracking; convolutional neural networks (CNN); adaptive attention mechanisms

1 Introduction

For Muslims, doing regional duties such as Hajj and Umrah is critical. But such duties are known by their extremely crowded spaces due to the big number of people. Hajj and Umrah are considered one of the largest human gatherings, where millions of people gather in a specific time and place, whether in the Two Holy Mosques (Makkah and Madinah), in addition to millions of pilgrims in the holy sites. The Kingdom works to maintain the safety and security of pilgrims and protect them from the dangers of crowding and gathering during the annual Hajj and Umrah seasons, in line with the Kingdom's vision 2030 to harness the potentials and capabilities to serve the guests of Rahman.

Hajj and Umrah are a major focus of Vision 2030 as the Kingdom seeks to increase the number of Umrah performers to 30 million by 2030. To help the faithful to perform their religious duties comfortably in overcrowded areas, a crowd management system is a must to control the entering and exiting for each place. Since the number of people is very high, an intelligent crowd management system can be developed to reduce human effort and accelerate the management process. A crowd management process is based on detecting, tracking, and counting human faces using Artificial Intelligence techniques. Human faces detection and counting will be performed to calculate the number of existing visitors and face recognition will be used to identify all the humans for security and health purposes.

First and foremost, effective crowd management helps to ensure the safety of those at Hajj and Umrah, from the guests to the security staff, and the workers. When the Hajj takes place in Makkah, everyone in the venue should be able to perform their duties without worrying about their safety. The consequences of a poorly managed crowd can be disastrous, people can be injured and lives can be lost. Effective crowd management can help minimize the risk of overcrowding occurrences.

The main importance of the crowd management system comes out when manipulating the crowd and ensuring comfortable movement and a safe environment. In effect, controlling the entering and exiting of authorized people can enable the manipulation of the crowd flow through an automatic system. Also, the crowd management system allows detecting dangerous situation because of overcrowded areas and ensure a quick intervention of the special teams to fix the problem. Automatic crowd management helps to reduce human effort and accelerate the process. Watching many surveillance systems is a hard task that needs a focus to detect dangerous situations.

In Hajj and Umrah situation, people are continuously moving. So, there is a possibility to count a person twice or more and that may cause a problem. To overcome this problem, we propose to add a face recognition framework to the crowd management system to assign each person with an ID and track it. This method will allow controlling the entering of only authorized people to avoid overcrowded areas and to eliminate the multiple counts of the same person to get precise statistics. Besides, focusing on faces detection allows enhancing the counting process since it is impossible to detect the entire body of the person in an overcrowded area.

The proposed crowd management system is mainly based on image processing tasks and data storage tasks. For image processing, the collected images are analyzed and human faces are detected and identified. Then each face's identifier (ID) is stored in a database for further recognition and

eliminating the possibility of multiple counting. The recent advances in image processing techniques [1] have boosted the state-of-the-art to a new level for many tasks such as image recognition [2], scene recognition [3], object detection [4], traffic signs detection [5], face identification [6], image inpainting [7], and medical image retrieval [8]. The success of those techniques comes from the use of very deep neural networks [9] with the ability to learn directly from input data and through a prediction methodology that mimics the biological brain.

In this work, we propose to use the Yolo v4 object detection framework [10] as our baseline and applied many improvements. Starting with the backbone, we proposed a better model with high accuracy and good processing speed. Then, we added adaptive attention mechanisms to make the model focus on the target to detect. For backbone, we applied the Cross-Stage-Partial-connections on the ResNeXt model. The ResneXt model [9] is a very powerful model that achieved state-of-the-art image recognition on the ImageNet dataset with a top-1 error of 21.2%. Cross-Stage-Partial-connections (CSP) [11] was proposed to reduce the computation cost of deep neural networks without decreasing the accuracy. It was very effective and allow a reduction of 20% of the computations. The CSP guarantee to achieve real-time processing using very deep neural networks on low-performance computers equipped with low-end graphics processing units (GPU).

The Yolo v4 has collected many techniques at the same framework. It started by proposing novel data augmentation techniques and designed better loss functions. Then, spatial pyramid pooling [12] and path aggregation [13] were applied at the detection stage. Those additions have improved the detection, accuracy of the Yolo v4 compared to the older Yolo version and state-of-the-art object detection frameworks. Combining many improvement techniques allows achieving better performances. In this work, we propose to add an adaptive attention mechanism to the Yolo v4 at the features extraction and detection stages to make more focus on the detected target and ignore false positive predictions. Besides, we propose a new set of anchors that fit for face detection at long distances. The proposed model was trained and evaluated on the wider face dataset [14].

The rest of the paper is organized as follows: section 2 will present an overview of related works with a discussion on the limitation of existing works. The proposed approach will be presented and detailed in section 3. In section 4, we present the experiment and report the achieved results while presenting a deep discussion on the efficiency of the proposed approach. Conclusions and future works will be presented in section 5.

2 Related Works

Crowd management systems are very important systems for controlling overcrowded spaces. Hajj and Umrah are the best situations to test the crowd management system because of the huge number of existing people and the complexity of the environment.

Generally, a crowd management system is based on detecting, tracking, and counting existing humans in a defined space. Many works have been proposed in the context of human detection for a variety of applications.

Ayachi et al. [15] proposed a pedestrian detection system for advanced driver assistance systems. The proposed system was based on lightweight separable convolution blocks, which are designed for possible embedded implementation. The detection was performed through a linear regression method to accelerate the processing time. In [16] a pedestrian detection system was proposed based on the Yolo v2 model [17] with a lightweight backbone. The SqueezeNet model [18] was used as a backbone to achieve better results and low model size.

For crowd management, humans must be detected and counted to control the flow. Lamba et al. [19] presented a survey on the most recent advances for crowd management and monitoring. In [20] a crowd management system was proposed to predict overcrowded urban areas when natural disasters happen such as an earthquake, typhoon, and national festivals. In such a situation the behavior of humans may differ from daily situations. The proposed system was based on an autoencoder with custom convolutional long-short term memory (LSTM) layers [21]. The proposed approach was used to detect crowd flow and can predict the flow density through the same model using a Multitask Convolutional LSTM Encoder-Decoder network. The achieved results were very impressive when the proposed approach was evaluated in different events such as the earthquake, New Year's Day, and the Tokyo marathon.

Das et al. [22] proposed a Convolutional Neural Network with a special attention mechanism for crowd counting to be used for crowd management. The proposed approach aims to separate standing people and sitting people and classify them into different categories. First, a base Convolutional Neural Network was used to predict the number of people in different categories. Second, a global density map was generated based on crowd counting. Finally, a refinement process was performed through a weighted linear regression layer to split the density map into standing and sitting density maps. A custom dataset was built for evaluating the proposed approach and an extensive experiment was performed to prove the efficiency of the method. Reported results show that the proposed approach achieved an accuracy of 86.49% and a mean absolute error of 4.80 and 4.15 for standing and sitting categories respectively which outperform existing state-of-the-art methods.

Seema et al. [23] propose to combine traffic analyses and crowd management for smart cities surveillance. The proposed system was based on the single-shot multi-box detection model (SSD) [24]. The SSD was used to count the objects of interest in the video. Based on counting vehicles and pedestrians, the traffic flow was managed. The vehicle's density was used to manage traffic signals. In addition, a license plate detection system was integrated to detect vehicles violating any traffic regulations. The surveillance videos were analyzed to predict the crowd statistics for managing the crowd in emergency cases safely.

A crowd counting and density estimation was proposed in [25] for crowd management. The proposed approach was based on the combination of a Convolutional Neural Network with two output signals, the first used for crowd counting and the second is used for density map estimation. For better prediction, only human heads were considered in the count. The proposed approach was evaluated on a custom-made dataset with 107 images containing 45000 annotated humans. The images have a range of between 58 and 2200 humans in each image. The achieved results show the efficiency of the proposed approach.

3 Proposed Approach

First, will present an overview on the proposed crowd management system and define its components. Second, we will move on to introduce the proposed backbone and the architecture of the building blocks. Also, we will present the applied technique to reduce the computations. Third, we will present the design of the Yolo v4 and its main parts. Finally, we will present the proposed adaptive attention mechanisms to improve the performances of the Yolo v4 for human face detection.

The proposed approach for crowd management in Hajj and Umrah was combined from an offline process and an online process. In the offline process, we train an object detection model for human face detection. In the online process, each detected face will be assigned with a numerical ID and stored for further identification and tracking. The pipeline of the proposed approach is illustrated in Fig. 1.

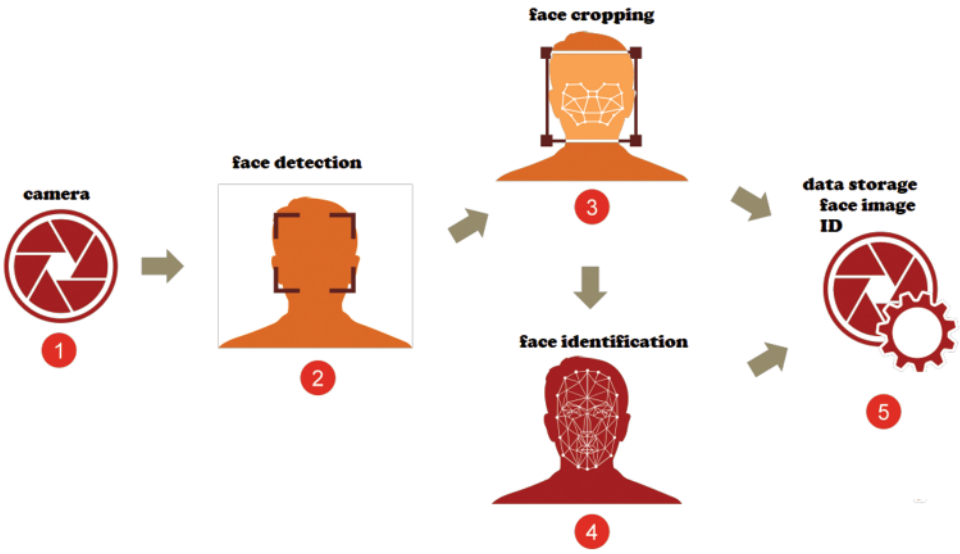

Figure 1: Proposed pipeline for human face detection and identification for crowd management system

In the first step, RGB cameras were used to collect data in a specific area. In the second step, a face detection model was used to detect human faces by processing the data provided by the cameras. In the third step, the detected face is cropped for further processing. In the fourth step, each cropped face is compared to the stored data and identified. If the cropped face was already identified and stored in the database, it is assigned with its old identifier. But if the face was not already identified, it will be assigned with a new identifier (ID). In the final step, the cropped face and its ID are stored in a database for further identification and tracking. This procedure eliminates the need for the identification of all humans present in an overcrowded space manually which reduces the human effort and facilitates the management of the crowds.

The challenging part of the proposed pipeline is face detection. Due to the challenging conditions such as the small size of the target, occlusion, geometric deformation, and so on, it is very necessary to build a robust face detection system that can overcome those challenges. So, we propose to use a powerful object detection framework with a very deep Convolutional Neural Network as a backbone. Also, we applied many techniques to enhance the performance and the processing speed.

The Yolo v4 was used as an object detection framework. It was designed to achieve real-time processing while getting high detection accuracy. The proposed improvement by Yolo v4 has enhanced the accuracy by 10% and the processing speed by 12% compared to the Yolo v3 [26]. Besides, the model training becomes easier and requires less computation, and can be performed on a single GPU.

In Yolo v4, they start by enhancing the backbone by applying a bag of freebies and a bag of specials. In this work, we proposed to replace the original darknet model with a deeper and more accurate model. For that, we proposed the use of the ResNeXt model [27]. The model was an extension of the ResNet model. The main innovation was to use a set of transformers with the same topology inside the same block (Fig. 2).

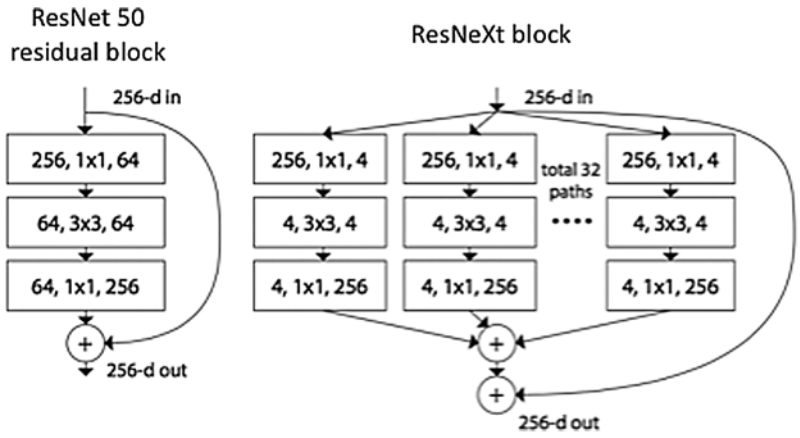

Figure 2: Building block of the ResNeXt model compared to ResNet model

The size of the transformers was called cardinality. Empirical experimentations on the ImageNet dataset [28] have proved that expanding the cardinality is more efficient than increasing the network depth or width to get better accuracy. In ResNeXt, an aggregation transformation was proposed to consider each neuron and network which is called network-in-neuron. The main idea was to project an input into a low dimension output and transform it. Then all the transformed outputs are aggregated to a single output. The use of a set of low dimension transformations made the network more flexible. Two main rules were considered to ensure that the computation complexity of the blocks is roughly the same: (1) in case of generating spatial maps of the same size, then the same hyper-parameters were shared in the block, and (2) for downsampling a spatial map by a factor of two, the width of the block must be doubled. Those rules were very important for maintaining the computation complexity and eliminated the design of each block separately. So, they design one block and repeated it until building the network while just changing the hyper-parameters at each level. The ResNeXt-101 has ranked in second place in the ImageNet large scale visual recognition challenge (ILSVRC) 2016 [28] with a top-1 error of 21.2%

As the ResNeXt-101 has a high computation complexity, we proposed to apply the Cross-Stage-Partial-connections (CSP) [11] to reduce the computation complexity without damaging the accuracy. The CSP has proved that 20% of the computation can be reduced without any loss in accuracy. The Cross-Stage-Partial-connections were inspired by the DenseNet [29] that comes with the idea of connecting all previous layers to the actual layer and propagate the bottom features to the top of the network cheaper and more effectively.

For CSP architecture, there are two main concepts. Fig. 3 illustrated the application of the CSP on the ResNeXt blocks. The first is the partial dense block and the second is the partial transition layer. For the partial dense block, the output of the previous block is divided into two parts where a part goes through the actual block and the other goes directly to the partial transition layer. The main purposes of this concept are the following: (i) doubling the gradient path and eliminating gradient vanishing. (ii) balancing the computation of each block bypassing only the half channels through the block while the other half passes to the end directly. (iii) reducing memory usage by saving half of the memory of part that passes without computations. For the partial transition layer, a hierarchical feature fusion mechanism was designed to prevent different layers from learning the same gradient information by splitting the gradient flow. Due to balancing the computation and reducing the memory usage, CSP

allows reducing the computation while guarding or enhancing the accuracy by enhancing the gradient pass.

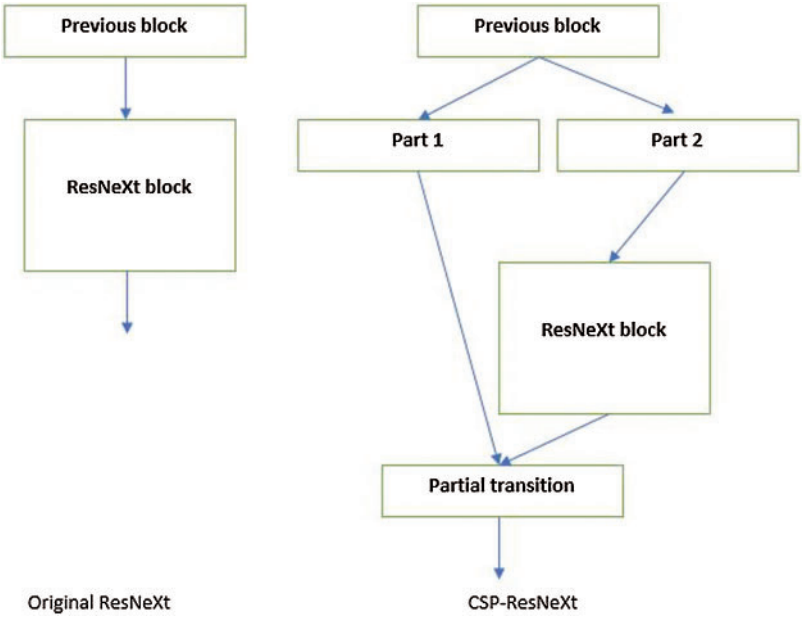

Figure 3: Applying the cross-stage-partial-connections on the ResNeXt model

To design the face detection system, we propose to integrate the proposed backbone to the Yolo v4 object detection framework [10]. The Yolo v4 has introduced the combination of almost all innovation techniques such as SPPNet [12], path aggregation [13], FPN [30], attention mechanism [31], and novel data augmentation techniques. This combination has widely enhanced the accuracy and the processing speed while reducing the computation overhead.

The Yolo v4 is composed of 4 main stages which are: the input, the backbone, the neck, and the prediction. Fig. 4 presents the architecture of the Yolo v4. The input is the image or the video that will be processed to detect existing objects. The backbone is a deep neural network used to extract relevant features for further processing. The neck is used for region proposals that may contain the target object. The prediction is the stage of score assignment for each detected object.

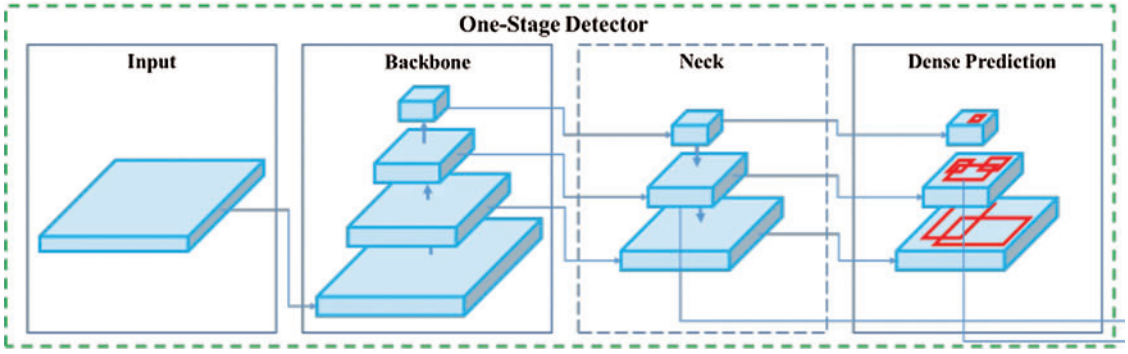

Figure 4: Yolo v4 architecture

The optimization techniques applied to the Yolo v4 were divided into two categories. All techniques applied to get better accuracy without increasing the inference speed were called bags of freebies. All techniques that enhance the accuracy and influence the inference speed were called bag of specials.

Generally, an object detection model is trained offline which allows to development of more efficient training techniques that result in achieving better accuracy without damaging the inference speed. Data augmentation techniques were the most used strategies to enhance accuracy. In effect, the main purpose of data augmentation techniques is to increase data variability to meet real-world conditions. Thus, improved the generalization power of the model and make it more robust when tested with new data. For real scene images, geometric and photometric distortions are one of the challenges to handle. So, applying a data augmentation technique that mimics those challenges was a very effective solution. Random scaling, cropping, translation, rotation, and flipping were the most used techniques to deal with geometric distortion while adjusting the contrast, hue, saturation, and noise were effectively deployed for dealing with photometric distortion. In Yolo v4, more data augmentation techniques were proposed mixing images by multiplying and superimposing with different coefficient ratios, and then adjusting the label with these superimposed ratios. Also, a CutMix technique [32] was used by implementing the ground truth of a random image into another image. Furthermore, a style transfer generative adversarial model (GAN) was used as a data augmentation technique. This was useful to reduce the texture bias learned by the detection model.

Usually, the training data is randomly collected and there is a problem of data imbalance between classes. The focal loss [33] was proposed to handle this problem by making the model focus on hard samples. There are more problems related to data collection and labeling such as expressing the relationship of the degree of association between different classes. To overcome this issue, a label smoothing [34] was proposed which convert hard labels to soft label for the training process.

Training a neural network model is based on optimizing a loss function using gradient descent algorithms. So, the loss function is a critical component for the performance of the model. Generally, cross-entropy and its variant are used for classification problems, and mean square error is used for regression problems which are used to predict the parameters of the bounding box. For a direct estimation of the bounding box parameters, each parameter must be treated as an independent variable. But such a method does not consider the integrity of the target object. To solve the problem, the intersection over union (IoU) was used as a loss function [35]. The newest variant CIoU loss [36] was proposed to consider the overlapping area, the distance between the center points of the ground truth and the predicted bounding boxes, and the aspect ratio.

For the bag of special, many techniques were proposed to enhance the receptive field and increase the capability of features integration. The SPP was integrated into Yolo v4 to enlarge the receptive field. Since the SPP was originally designed to generate a vector in output and this cannot be applied for dense prediction using convolution layers, it was modified by concatenating different outputs to a tensor and used as input to the next layer. Besides, an attention model was deployed to enhance the accuracy. The spatial attention module (SAM) [37] was modified for use in Yolo v4 by replacing the pooling operation followed by convolution with only a convolution layer. The attention model results in small additional computation and does not affect the inference speed. Another important focus in the design of the neural network model is the activation function. Since the first success of the convolutional neural network for the computer vision task, the rectified linear unit (ReLU) was the first activation function that solves the problem of gradient vanishing. Then, many activation functions were proposed to enhance the performance such as the PReLU and Leaky ReLU, Scaled Exponential

Linear Unit (SELU), swish and Mish. The Mish activation function was adopted by Yolo v4 since it achieved state-of-the-art performances and it is a continuously differentiable activation function.

For the final prediction process, the non-maximum suppression (NMS) technique is applied to select the best-fit bounding box from a set of bounding boxes that predicts the same object. The original NMS does not consider the object context. So, for Yolo v4 the DIoU NMS [36] was used to consider the difference between center points as context information.

The Yolo v4 was designed for general object detection and does not work well for the detection of tiny human faces. So, we propose to add an adaptive attention mechanism to enhance the focus of the model on human faces and allow their detection in high-resolution images. Generally, channel-wise attention is solving the problem of what to focus on and spatial attention is used to solve the problem of where to focus. So, we propose to combine both attention mechanisms for better performances.

Traditional attention mechanisms are designed through pooling layers across the channel dimensions. However, we designed an adaptive spatial attention mechanism through a fully convolutional layer. As pooling layers are parameterless, using convolutional layers instead enhances the learning capability of the model without any additional computations. Fig. 5 presents the proposed spatial attention mechanisms. It was designed in a way to generate an attention map that recalibrates the features from different spatial locations. The 1×1 convolution layer was used to squeeze the features across the channel dimension. Besides, it prevents the influence of the backpropagation on the backbone directly. The 3×3 convolution is used to excite a local area reaction to enhance the efficacity. Since the adaptive spatial attention mechanism is composed of 1×1 and 3×3 convolution, the relative position and the receptive fields of the spatial attention map are similar to the output of the backbone. So, the pixels of the spatial attention network weigh the pixels of output feature maps at the same location. The adaptive spatial attention mechanism was integrated into the backbone in a plug-in manner.

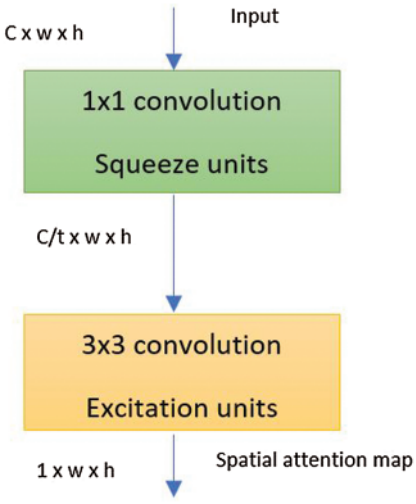

Figure 5: Adaptive spatial attention mechanism. (t is the compression ratio)

The adaptive channel-wise attention was designed by a squeeze and excitation structure followed by a domain attention network. Fig. 6 illustrates the proposed adaptive channel-wise attention mechanism. Mainly, different pooling layers are important for the attention mechanism. The main idea behind this is for an input feature map, applying a global average pooling allows the identification

of the object extend, and applying a global max pooling tends to identify the location of the object which are two main features for the detection task. Using global max pooling instead of max pooling is more useful for the detection of small objects.

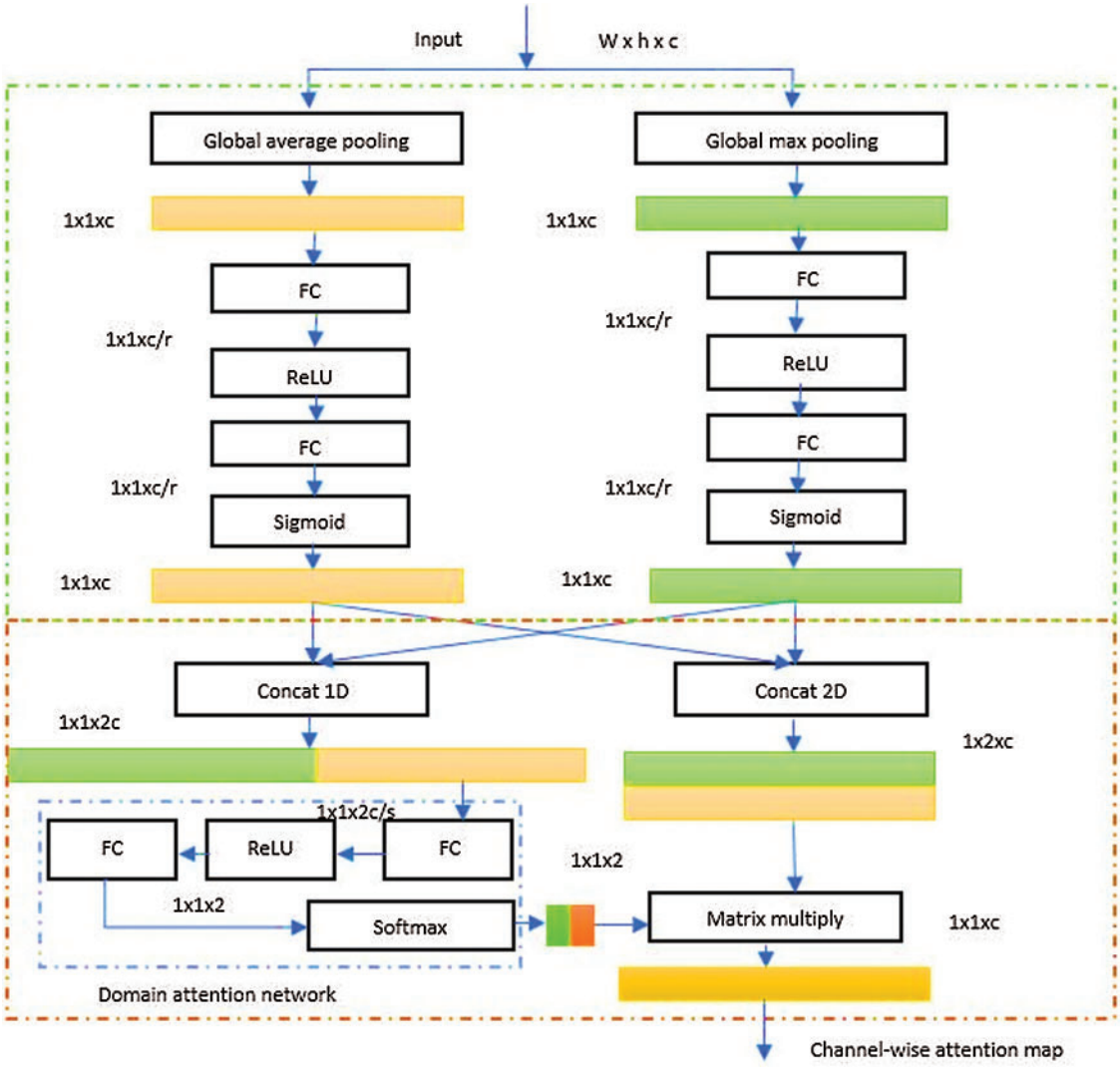

Figure 6: Adaptive channel-wise attention mechanism. (s and r are the compression ratios)

Considering the mentioned above, many works have designed a channel-wise attention mechanism based on the combination of global max pooling and global average pooling. Then, weigh both paths equally. But in reality, objects have different scales and aspect ratios and an equal weight may work well for some objects but the bad result will be achieved for others. In this work, we deal with human faces at different aspect ratios which we will focus on. The adaptive channel-wise attention mechanism will be designed to handle the difference of aspect ratios to achieve the ultimate results. As mentioned earlier, we added a domain attention network to the pooling structure which is the main novelty of the proposed attention mechanism compared to existing ones. The proposed domain attention network was designed concerning three main rules. First, it must be fully data-driven where intermediate

features and outputs can be adapted to the input. Second, the network must be powerful to weigh raw vectors.

Finally, the network must be lightweight to avoid many additional computations and reduce the overall complexity. The domain attention network is composed form three fully connected layers and a hidden layer. The output of the network is weight tensor sensitive to the target domain. This vector is used to recalibrate the raw channel-wise generated by the previous pooling structure. The adaptive channel-wise attention mechanism was integrated into the detection stage and the backbone at the ResNeXt blocks where low semantic features can be detected.

The integration of the proposed adaptive attention mechanisms was performed in a way to maintain the backbone structure to take advantage of the pre-trained weight and we make several changes on the detection stage. The adaptive spatial attention mechanism was integrated into the ResNeXt building blocks (ResX). Also, the adaptive channel-wise attention mechanism was integrated into the ResX blocks after applying the adaptive spatial attention mechanism. In the detection stage, only the adaptive channel-wise attention mechanism was applied. Since the top layers contain rich sematic features and less positional information, it was important to implement the channel-wise attention but the spatial attention has no impact. Fig. 7 present the design of the proposed implementation of the adaptive attention mechanisms on the Yolo v4. For clearance, the presented structure of the Yolo v4 was simplified and only important parts were illustrated.

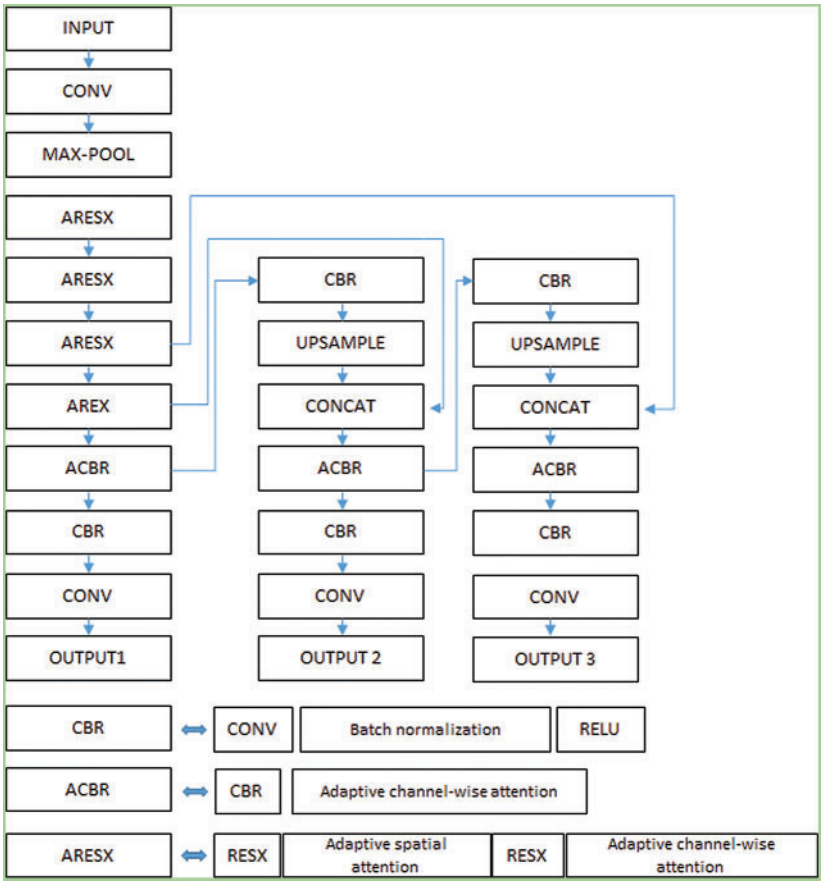

Figure 7: The proposed implementation of adaptive attention mechanisms in the Yolo v4

The proposed adaptive attention mechanisms were designed to be implemented in a plug-in manner. Due to the lack of positional features in the top layers and the small size of feature maps, channel wise attention mechanism has been integrated. Subsequently, spatial and channel-wise attention mechanisms have been integrated in the bottom building blocks of the backbone because of the lack of semantic features at those layers. This configuration enabled a quick initialization of the model using pre-trained weight which accelerate the training process and guarantee high performances.

Considering the mentioned analyses, the design of the proposed model was based on the Yolo v4 with ResNeXt model as backbone with additional adaptive attention mechanisms. As shown in Fig. 7, the channel-wise attention is implemented in the ACBL block and the ARESX building block while the spatial attention mechanism is only implemented in the ARESX building block.

4 Experiments and Results

4.1 Dataset

For training and evaluation, the wider face dataset [14] was used. The data was collected from the internet using search engines such as Google and Bing. Then, the data were manually filtered and annotated. The dataset contains 32203 images with a total of 393703 annotated faces. The dataset is very challenging due to the diversity of the collected data and the capturing conditions such as occlusions, pose variation, and geometric deformation. Most of the images in the dataset were collected from events in overcrowded areas. That makes it very useful for the studied task and will help to achieve better performance. The dataset was randomly divided into training, validation and testing sets where 40% of the data was used for training, 10% for validation and 50% for testing. Pascal VOC evaluation metric was adopted for the evaluation of performances.

4.2 Implementation Details

All the experiments were carried out on a desktop running the Ubuntu 20.04 LTS equipped with an Intel i7 CPU, 32 GB of RAM, an Nvidia GTX 960 GPU. TensorFlow Deep Learning framework was used for the development of the proposed model with support of CUDA acceleration and cuDNN library. The OpenCV library was used for images manipulation and display.

Model training was performed using the Adam optimizer which is a gradient descent variant that optimizes the learning rate alongside the parameters and accelerates the convergence process. The model was trained for 40 epochs with an initial learning rate of 0.001. The size of the input images was fixed to 320×320 for both training and testing to achieve high performance and to respect real-time constraints. The batch size was fixed to 4 due to the limited memory of the used GPU. Backbone was initialized using the pre-trained weights on the ImageNet dataset. The model has trained alternatively by training the detection stage and freezing the backbone then training the complete model. The compression ratios were fixed as follow: $r = s = 16$ and $t = 32$. The training was performed for 110k iterations and lasted for two days. An early stop condition was established if the loss is not reduced for 10 K iterations.

4.3 Evaluation and Comparison

The performance of the model was evaluated based on different metrics such as mean Average Precision (mAP), processing speed (FPS), and floating-point operations (FLOPS). The proposed model was evaluated using the standard parameters of Yolo v4. The loss optimization curves are

presented in Fig. 8. The loss was reduced from 65.34 to 1.05 which proved the efficiency of the learning algorithm based on Adam optimizer for searching the minimum. Besides, the use of the pre-trained weight has been very important for reaching the convergence. The proposed model has achieved an mAP of 92.1% while running with a speed of 18 FPS. The achieved results outperform most state-of the-art models tested on the wider faces dataset.

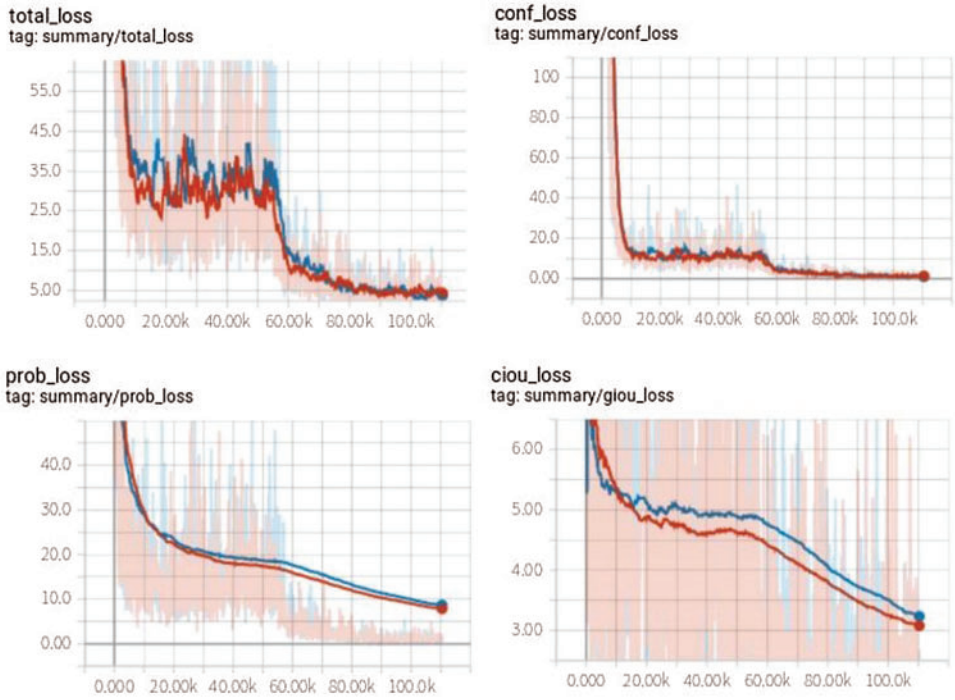

Figure 8: Different loss curves of the proposed model compared to the original Yolo v4. Bleu for the Yolo v4 and red for the proposed model

To further improve the efficiency of the proposed method, we compared against the state-of-the-art works on the same dataset. Tab. 1 presents a comparison against the most recent works on the used dataset. The proposed method widely achieved the existing works with a big margin in terms of both precision and speed. Even methods with good precision struggles from slow processing and fast methods achieve low precision. However, the proposed method has achieved a good trade-off between speed and precision.

4.4 Ablation Study

An ablation study was conducted to evaluate the effectiveness of the adaptive attention mechanisms. To show the impact of the proposed improvement, we evaluated the performance of the original Yolo v4 on the same dataset. Tab. 2 presents the achieved results for the original Yolo v4 and the improved version with adaptive attention mechanisms. The original Yolo v4 has fewer GFLOPS that is due to the additional attention mechanism implemented in the improved version. The proposed version has better precision and a similar processing speed to the original one. As shown in Tab. 2, the proposed adaptive attention mechanisms have a positive impact on precision and do not explode the computation complexity.

Table 1: Comparison against most recent works on the wider faces dataset

Model	mAP (%)	Speed (FPS)
Faster RCNN [38]	88.7	4
Zhang et al. [39]	89.1	10
HOANG [40]	75.4	12
Yolo-faces [41]	69.3	38
Retinaface [42]	61.55	22
Yolo v4 (ours)	92.1	18

Table 2: Achieved results compared to the original Yolo v4

Model	mAP (%)	Speed (FPS)	Model size (MB)	GFLOPS
Yolo v4 (original)	90.8	19	268.5	118.6
Yolo v4 (ours)	92.1	18	270.3	120.2

4.5 Implementation Demo

The proposed face detection model was integrated into a crowd management system based on detecting and counting human faces to estimate crowd density and facilitate their management. A demo of human faces detection in hajj is presented in Fig. 9. The proposed system has proved its efficiency for detecting tiny faces at a complex background and degraded conditions such as occlusion and deformation. The generalization power of the detection model was very high since it was not trained on Hajj and Umrah images. Using the adaptive attention mechanisms has a great impact on the overall performance.

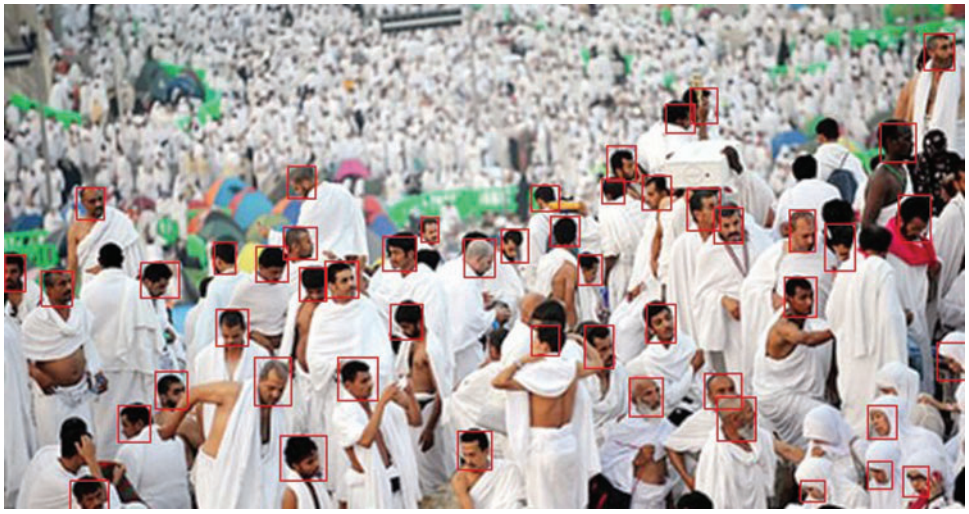

Figure 9: Demo of the human faces detection in Hajj

5 Conclusion

Due to the importance of Hajj and Umrah for Muslims, it is very critical to do their duties in comfortable situations. Crowd management systems are a good solution to manage the crowd to avoid dangerous situations. In this paper, we proposed a crowd management system based on detecting, tracking, and counting human faces. It is more efficient to detect human faces instead of detecting the whole body in a crowded area due to challenging conditions such as occlusion and deformation. The proposed face detection method was based on the Yolo v4 object detection framework with a ResNeXt backbone and additional adaptive attention mechanisms. Extensive experimentation has proved the efficiency of the proposed adaptive attention mechanism. We proposed two kinds of attention to taking advantage of all the features. The adaptive spatial attention was used to solve the problem of object position and the adaptive channel-wise attention was used to resolve the problem of what object to focus on. Compared to many existing works, the proposed method achieved a good balance between precision and speed.

Funding Statement: This work was funded by the University of Jeddah, Jeddah, Saudi Arabia, under Grant No. (UJ-21-ICL-4). The authors, therefore, acknowledge with thanks the University of Jeddah technical and financial support.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. C. Pandey, "Mind, machine, and image processing," in *Deep Learning for Image Processing Applications*, 1st ed., vol. 31, Amsterdam, Netherlands: IOS Press, pp. 1–26, 2017.

[2] A. Mouna, R. Ayachi, Y. Said, E. Pissaloux and M. Atri, "Indoor image recognition and classification via deep convolutional neural network," in *Proc. Int. Conf. on the Sciences of Electronics, Technologies of Information and Telecommunications*, Hammamet, Tunisia, Springer, Cham, pp. 364–371, 2018.

[3] A. Mouna, R. Ayachi, Y. Said and M. Atri, "Deep learning-based application for indoor scene recognition," *Neural Processing Letters*, vol. 51, no. 3, pp. 1–11, 2020.

[4] A. Riadh, Y. Said and M. Atri, "A convolutional neural network to perform object detection and identification in visual large-scale data," *Big Data*, vol. 9, no. 1, pp. 41–52, 2021.

[5] A. Riadh, M. Afif, Y. Said and M. Atri, "Traffic signs detection for real-world application of an advanced driving assisting system using deep learning," *Neural Processing Letters*, vol. 51, no. 1, pp. 837–851, 2020.

[6] Y. Said, M. Barr and H. E. Ahmed, "Design of a face recognition system based on convolutional neural network (CNN)," *Engineering, Technology & Applied Science Research*, vol. 10, no. 3, pp. 5608–5612, 2020.

[7] S. Ozturk, "Image inpainting based compact hash code learning using modified U-net," in *Proc. Int. Symp. on Multidisciplinary Studies and Innovative Technologies*, Istanbul, Turkey, pp. 1–5, 2020.

[8] Ş Öztürk, "Class-driven content-based medical image retrieval using hash codes of deep features," *Biomedical Signal Processing and Control*, vol. 68, no. 102601, pp. 1–9, 2021.

[9] L. Weibo, Z. Wang, X. Liu, N. Zeng, Y. Liu *et al.*, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.

[10] B. Alexey, C. Y. Wang and H. Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[11] W. Chien-Yao, H. Y. M. Liao, Y. H. Wu, P. Y. Chen, J. W. Hsieh *et al.*, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, Seattle, WA, USA, pp. 390–391, 2020.

[12] H. Kaiming, X. Zhang, S. Ren and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.
[13] L. Shu, L. Qi, H. Qin, J. Shi and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 8759–8768. 2018.
[14] Y. Shuo, P. Luo, C. C. Loy and X. Tang, "Wider face: A face detection benchmark," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 5525–5533. 2016.
[15] A. Riadh, Y. Said and A. B. Abdelaali, "Pedestrian detection based on light-weighted separable convolution for advanced driver assistance systems," *Neural Processing Letters*, vol. 52, no. 3, pp. 2655–2668, 2020.
[16] Y. F. Said and M. Barr, "Pedestrian detection for advanced driver assistance systems using deep learning algorithms," *IJCSNS*, vol. 19, no. 10, pp. 9–14, 2019.
[17] R. Joseph and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 7263–7271. 2017.
[18] I. Forrest, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally *et al.*, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size," ArXiv preprint arXiv:1602.07360, 2016.
[19] L. Sonu and N. Nain, "Crowd monitoring and classification: A survey," *Advances in Computer and Computational Sciences*, vol. 1, pp. 21–31, 2017.
[20] J. Renhe, X. Song, D. Huang, X. Song, T. Xia *et al.*, "Deep urban event: A system for predicting citywide crowd dynamics at big events," in *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, Anchorage, AK, USA, pp. 2114–2122, 2019.
[21] S. Xingjian, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong *et al.*, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *Advances in Neural Information Processing Systems*, vol. 28, pp. 802–810, 2015.
[22] D. S. Sarathi, S. Mt Rashid and M. E. Ali, "CCCNet: An attention based deep learning framework for categorized counting of crowd in different body states," in *Proc. Int. IEEE Joint Conf. on Neural Networks (IJCNN)*, Glasgow, UK, pp. 1–8, 2020.
[23] S. Seema, S. Goutham, S. Vasudev and R. R. Putane, "Deep learning models for analysis of traffic and crowd management from surveillance videos," in *Progress in Computing, Analytics and Networking*, vol. 1119, pp. 83–93, 2020.
[24] L. Wei, D. Anguelov, D. Erhan, C. Szegedy, S. Reed *et al.*, "Ssd: Single shot multibox detector," in *Proc. European Conf. on Computer Vision*, Amsterdam, The Netherlands, pp. 21–37, 2016.
[25] H. Yaocong, H. Chang, F. Nian, Y. Wang and T. Li, "Dense crowd counting from still images with convolutional neural networks," *Journal of Visual Communication and Image Representation*, vol. 38, pp. 530–539, 2016.
[26] R. Joseph and A. Farhadi, "Yolov3: An incremental improvement," ArXiv preprint arXiv:1804.02767, 2018.
[27] X. Saining, R. Girshick, P. Dollár, Z. Tu and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 1492–1500, 2017.
[28] R. Olga, J. Deng, H. Su, J. Krause, S. Satheesh *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
[29] H. Gao, Z. Liu, L. V. D. Maaten and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 4700–4708, 2017.
[30] L. Tsung-Yi, P. Dollár, R. Girshick, K. He, B. Hariharan *et al.*, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 2117–2125, 2017.
[31] V. Ashish, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[32] S. Yun, D. Han, S. Joon, S. Chun, J. Choe *et al.*, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, Seoul, South Korea, pp. 6023–6032, 2019.
[33] L. Tsung-Yi, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 2980–2988, 2017.
[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 2818–2826, 2016.
[35] J. Yu, Y. Jiang, Z. Wang, Z. Cao and T. Huang, "UnitBox: An advanced object detection network," in *Proc. ACM Int. Conf. on Multimedia*, Amsterdam, The Netherlands, pp. 516–520, 2016.
[36] Z. Zhaohui, P. Wang, W. Liu, J. Li, R. Ye *et al.*, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. on Artificial Intelligence*, vol. 34, no. 07, pp. 12993–13000, 2020.
[37] S. Woo, J. Park, J. Y. Lee and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 3–19, 2018.
[38] Z. Xiaoxing, X. Peng, Y. Wang and Y. Qiao, "Finding hard faces with better proposals and classifier," *Machine Vision and Applications*, vol. 31, no. 7, pp. 1–15, 2020.
[39] Z. Zhishuai, W. Shen, S. Qiao, Y. Wang, B. Wang *et al.*, "Robust face detection via learning small faces on hard images," in *Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision*, Snowmass, CO, USA, pp. 1361–1370, 2020.
[40] H. T. Minh, G. P. Nam, J. Cho and I. J. Kim, "Deface: Deep efficient face network for small scale variations," *IEEE Access*, vol. 8, pp. 142423–142433, 2020.
[41] W. Chen, H. Huang, S. Peng, C. Zhou and C. Zhang, "YOLO-Face: A real-time face detector," *The Visual Computer*, vol. 37, no. 4, pp. 805–813, 2021.
[42] J. Deng, J. Guo, E. Ververas, I. Kotsia and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 5203–5212, 2020.