Received 7 March 2023, accepted 2 April 2023, date of publication 12 April 2023, date of current version 19 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3266514

RESEARCH ARTICLE

A Deep Learning-Based Efficient Firearms Monitoring Technique for Building Secure Smart Cities

RAJDEEP CHATTERJEE[1,2], (Member, IEEE), ANKITA CHATTERJEE[2], MANAS RANJAN PRADHAN[3], (Member, IEEE), BISWARANJAN ACHARYA[4], (Senior Member, IEEE), AND TANUPRIYA CHOUDHURY[5,6], (Senior Member, IEEE)

[1] School of Computer Engineering, Kalinga Institute of Industrial Technology (Deemed to be University), Bhubaneswar, Odisha 751024, India
[2] Amygdala AI, Khorda, Odisha 751024, India
[3] School of Computing, Skyline University College, Sharjah, United Arab Emirates
[4] Department of Computer Engineering-Artificial Intelligence and Big Data Analytics, Marwadi University, Rajkot, Gujarat 360003, India
[5] Informatics Cluster, School of Computer Sciences, University of Petroleum and Energy Studies (UPES), Dehradun, Uttarakhand 248007, India
[6] CSE Department, Graphic Era Hill University, Dehradun, Uttarakhand 248002, India

Corresponding authors: Rajdeep Chatterjee (cse.rajdeep@gmail.com) and Biswaranjan Acharya (biswaacharya@ieee.org)

ABSTRACT Violence, in any form, is a disgrace to our civilized world. Nevertheless, even in modern times, violence is an integral part of our society and causes the deaths of many innocent lives. One of the conventional means of violence is using a firearm. Firearm-related deaths are currently a global phenomenon. It is a threat to society and a challenge to law enforcement agencies. A significant portion of such crimes happen in semi-urban areas or cities. Governments and private organizations use CCTV-based surveillance extensively today for prevention and monitoring. However, human-based monitoring requires a significant amount of person-hours as a resource and is prone to mistakes. On the other hand, automated smart surveillance for violent activities is more suitable for scale and reliability. The paper's main focus is to showcase that deep learning-based techniques can be used in combination to detect firearms (particularly guns). This paper uses different detection techniques, such as Faster Region-Based Convolutional Neural Networks (Faster RCNN) and the latest EfficientDet-based architectures for detecting guns and human faces. An ensemble (stacked) scheme has improved the detection performance to identify human faces and guns at the post-processing level using Non-Maximum Suppression, Non-Maximum Weighted, and Weighted Box Fusion techniques. This paper has empirically discussed the comparative results of various detection techniques and their ensembles. It helps the police gather quick intelligence about the incident and take preventive measures at the earliest. Also, the same technique can be used to identify social media videos for gun-based content detection. Here, the Weighted Box Fusion-based Ensemble Detection Scheme provides mean average precisions 77.02%, 16.40%, 29.73% for the mAP$_0$.5, mAP$_{0.75}$ and mAP$_{[0.500.95]}$, respectively. The results achieve the best performance among all the experimented alternatives. The model has been rigorously tested with unknown test images and movie clips. The obtained ensemble schemes are satisfactory and consistently improve over primary models.

INDEX TERMS Computer vision, deep learning, ensemble, firearms, smart cities.

I. INTRODUCTION

The world is witnessing a COVID-19 pandemic, which is causing thousands of deaths outside of national borders [1], [2]. People give little attention to other dominant reasons for death in our society. Road accidents and murders are two very prominent causes of the untimely deaths of innocent lives every year [3], [4], [5]. In the 19th century, guns played a crucial role in the American Civil War and subsequently became part of the regular infantry and law enforcement agencies of

The associate editor coordinating the review of this manuscript and approving it for publication was Sathish Kumar.

many countries. However, a gun is also used by the rioters and criminals for their own gain or to terrorize the commoners [6], [7]. People associate gun violence with mass shootings following World War II [8]. In 2015, a total of 36,252 U.S. citizens were killed from gun violence [9]. It continues to be a peril for the United States national health; thus, it becomes a major threat to public health. A massive 1.4 million people were killed worldwide by gun violence between 2012 and 2016 [10]. Some of the most prominent mass shootings in history are: the 1987 Hoddle Street Massacre [11] in Hoddle Street, Clifton Hill, Melbourne; the 2019 Christchurch mosque shootings in Christchurch [12], which resulted in deaths; the 2017 Quebec City mosque shooting [13] in Quebec City; The deadliest mass shooting by a lone individual in modern history occurred in Europe with the 2011 Norway attacks [14] in Norway, in which 92 people were killed. Mass shootings are also a grim reality in Israel. In the 1972 Lod Airport Massacre [15], which killed 26 and injured 80, In India, gun violence has been gaining momentum in recent years. One of the most infamous incidents is the 26/11 Mumbai [16] attacks, which killed 58 people and injured 104 others at the Chhatrapati Shivaji Maharaj Terminus and Taj Hotel in Mumbai, India.

A. RESEARCH GAP

If in a single incident, there is more than one victim due to firearms (gun violence), it is considered as a mass shooting. Mass shootings can be defined in various ways. However, as per FBI (U.S.) defines a ''mass murder'' as ''four or more murdered during an event with 36 no cooling-off period between the murders.'' Again in US Public Law 112–265112th Congress (2013), it is passed that the phrase ''mass killings'' suggests three or more killings in a single incident in a place of public use [17]. Nevertheless, one can define a mass shooting as three or more people being killed or wounded in a single place for a specific period, excluding the shooters. Mass shootings at public places such as schools, university campuses, hotels, and workplaces are modern-day phenomena. Persons under the influence of alcohol, drugs, or a specific ideology perpetrate this type of heinous crime. Detecting people with guns in public places is really a challenge.

B. CONTRIBUTION

Smart surveillance is an indispensable part of smart cities. It makes a city more secure, and thus, the smart city becomes socially and economically sustainable. From a minor burglary to the assassination, different types of guns have been used. 24 × 7 human monitoring [18], [19] for all the buildings and public/private properties is not feasible due to skilled human resource constraints. It is also not a cost-effective approach. Therefore, an automated deep learning-based firearms detection technique is efficient for on-site monitoring in building secure and sustainable smart cities. Besides, face detection also helps the authority know about the perpetrator or the persons around the gunner for police intelligence. A CCTV-based automated face and gun monitoring system is one of the smart infrastructures in the future of sustainable smart cities.

This paper has proposed an ensemble detection strategy for human faces and guns (revolver, pistol, handgun, etc.) in a given image (or video frame). We have used Faster Region-based Convolutional Neural Networks (here on, FRCNN) [20] architecture with ResNet50 [21], [22], [23] and VGG16 [24] as backbones. Also, the EfficientDet [25] architecture with EfficientNet-B0 [26] as the backbone has been implemented for a comparison. Different combinations of detectors have been explored in stacked ensemble configuration after the models have been built as a post-processing phase for detection. Three distinct types of combining techniques have been employed. Non-Maximum Suppression (NMS), Non-Maximum Weighted (NMW), and Weighted Box Fusion (WBF) are used to obtain the final bounding box for an object from all the overlapping boxes. Multiple boxes are generated due to multiple detectors for the same image. The novelty of our work is to empirically demonstrate that the ensemble of Faster RCNN and the latest EfficientDet architectures provide an improved object detection scheme using the same trained models as the performance of the individual model. In the paper's title, the term efficient indicates that the detection results can be improved with the existing trained models without further training in the proposed scheme.

The paper has aimed to contribute as follows:

- ⋆ Automated detection of human faces and different types of guns together in the wild
- ⋆ Introducing a deep learning-based framework to improve the performance of object detection through ensemble
- ⋆ Thus, securing smart cities using intelligent surveillance

C. ORGANIZATION

The paper is divided into six sections. The related studies have been discussed in Section II. In Section III, the proposed ensemble scheme has been described. The following Section IV tells about the dataset and the experimental preparations. The results are explained in Section V. Finally, we conclude the paper in Section VI.

II. RELATED WORKS

In the real world, people often use deep learning techniques like image segmentation, classification, and detection to find answers. There is a lot of literature about different deep learning methods for recognizing human faces. Few research works have been done about deep learning-based gun detection along with face recognition. In this section, we have discussed about some of the research works that focus on guns.

The authors of [27] say that their model breaks the gun (or other weapon) down into its parts and shows how they all work together. Now, a straightforward deep neural network can find the weapon with ease. The final product has been produced by combining all of its results. The AR-15 is the sole type of rifle that is the subject of the study. Once more, they have not given a peer comparison of their semantic neural network model. In [28], the authors have utilized the transfer learning technique with a Convolutional Neural Network pre-trained model for gun detection utilizing X-ray luggage imaging. Transfer learning is advantageous since it performs well even with insufficient training data. In order to fine-tune the current issue, the pre-trained model must first be constructed with adequate data samples and then reused with the same weights. As the baggage is treated as a static background, the work is constrained. In other words, rather than being formed in the wild, the classification model is created in a controlled setting. The authors of [29] have developed a Faster RCNN for gun (pistol) identification based on VGG16. The main goal is to sound an alert if the model spots guns five times in a row in the film. Although they used several datasets, they did not compare the various detection methods.

Two well-known detection architectures for distinguishing between several sorts of weapons (not just one type of gun) have been employed in this research. The post-processing methods like NMS, NMW, and WBF are used to build improved detection models. Even though the model parameters have not been trained or changed any further, the results show that the ensemble techniques are better than the individual architectures. As a result, the suggested ensemble technique produces better detection performance while saving time.

III. PROPOSED ENSEMBLE SCHEME FOR OBJECT DETECTION

A. FASTER REGION-BASED CONVOLUTIONAL NEURAL NETWORKS

Faster RCNN (FRCNN) consists of two steps. The first step is called a Region Proposal Network (RPN) [30] which introduces candidate object bounding boxes. RPN is the replacement for the previously used selective search method. In other words, RPN identifies region boxes (commonly known as anchors) and proposes the boxes that most likely cover the objects. The second step extracts features through RoI pooling from each candidate box. RoI Pooling divides the input feature map into a fixed-sized corresponding region and then applies Max-Pooling to every region. Thus, the output of RoI pooling always has the same dimensions, regardless of the size of the input. Subsequently, it executes the classification and bounding box regression jointly using Eq. 1. FRCNN provides faster inference by sharing the convolutional features of both the RPN and the Fast RCNN. The mechanism helps the unified network find possible objects in a given image.

$$L(P_i, t_i) = \underbrace{\frac{1}{N_{cls}} \sum^{i} L_{cls}(P_i, P_i^*)}_{\text{object / no object}} + \lambda \underbrace{\frac{1}{N_{reg}} \sum^{i} P_i^* L_{reg}(t_i, t_i^*)}_{\text{box regressor}} \tag{1}$$

P_i is the predicted probability, P_i^* indicates 1 for positive anchor and 0 for negative anchor, N_{cls} is the number of anchors in the mini-batch, λ is positive constant (here, 1.0), N_{reg} is the number of total anchors (RPN, $\approx$ 300), t_i and t_i^* suggest the predicted and the ground truth bounding boxes, respectively.

B. EfficientDet

Another object detection strategy we used in this work was EfficientDet. EfficentNet-B0 has been used as the backbone network in EfficientDet. This network is made up of two major components: (a) Bidirectional Feature Pyramid Network (BiFPN), which allows bidirectional fast multi-scale feature fusion citewu2020single. (b) A new compound scaling method jointly scales up [31] backbone, feature network, the box/class network, and the resolution. The BiFPN is used as the feature network. The object class and box network weights are shared across all levels of features. The backbone network EfficientNet gives a remarkable performance in image classification by jointly scaling up all dimensions of network width, depth, and input resolution. We have used the EfficientNet-B0 variant of the model as it has the lowest number of trainable parameters. This object detection method has been combined with the scaling up of the method by using the coefficient ϕ (phi) to jointly scale up to all dimensions of the backbone network, the BiFPN network, the class/box network, and resolution.

C. MEAN AVERAGE PRECISION

Intersection over Union (IoU) is a Jaccard Index-based metric for evaluating the intersection area between two bounding boxes. It uses the actual bounding box (ground truth) and a predicted bounding box. It measures whether the detection of an object is valid (true positive) or not (false positive). Real Positive (TP) indicates a correct detection where IoU is higher than and equals ($\geq$) the given threshold. Similarly, the False Positive (FP) suggests an incorrect detection where IoU is less than ($<$) the threshold value. The common practice is to set the threshold to 50% or 75% or an average range of $50-95\%$ (with 5 step size).

For the readers who are new to this field, it is essential to note that the mAP$_{0.5}$ [32] is calculated using the IoU@0.5, indicating the intersection area between the predicted and actual (ground truth) bounding boxes is 0.5 or above. It is easy to realize using a diagram (see Fig. 1). The green-colored background box is the ground truth (the annotated object), and the red-colored bounding box corresponds to the predicted one.

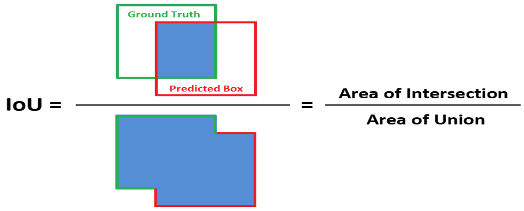


FIGURE 1. Intersection over Union (IoU).

The Mean Average Precision (*mAP*) [33], [34] is the arithmetic mean of the average precision (AP) [35] values for n number of classes (that is, object type). It is given as Eq. 2.

$$mAP = \frac{1}{n}\sum_{n} AP_n \tag{2}$$

D. NON-MAXIMUM SUPPRESSION

Suppose multiple detection techniques have been applied to a single image. The possibility exists that multiple overlapping bounding boxes are generated for the same object in a given image. NMS is an efficient solution to address the said issue. It works as long as the Jaccard overlap between any two predicted bounding boxes exceeds a given threshold; those boxes are considered the detections for the same object. The one box with higher confidence becomes the final prediction. It provides only the largest bounding box, which adequately covers the object [36], [37] while ignoring all the non-maximum boxes.

E. NON-MAXIMUM WEIGHTED

NMS discards the low confidence bounding boxes. Some times these boxes contain latent information regarding an object. Such low confidence boxes can improve the overall coverage of the object and give better detection than greedy NMS. It is explained with an example in [38]. Authors have argued that if two boxes have similar confidences, obtaining the average box is more effective than considering the maximum one.

F. WEIGHTED BOXES FUSION

Weighted Box Fusion is an ensembling technique to combine the predicted bounding boxes from different object detection models. Authors [39] have shown that WBF boosts the performance in detection. Unlike the NMS variants, WBF considers predictions from all the boxes. Similarly, NMW does not consider information about how many predicted boxes are in a group. NMW also uses the IoU value to compute weighted boxes. These are a few drawbacks on the NMW side. As it uses all the box information, not excluding a few predicted boxes, it even gives a better-predicted box from all or most inaccurate bounding boxes.

G. ENSEMBLE DETECTION SCHEME

In our ensemble (object) detection scheme, the training architecture of a detector is given images that have been labeled with the ground truth. FRCNN and EfficientDet have been implemented as the primary object detectors. Each detector can use a variety of deep image classifiers as a backbone. Some of the popular backbone algorithms are VGG16 and ResNet50. However, so far, EfficientDet uses the EfficientNet family of classifiers as its backbone. In this paper, pre-trained ImagNet weights [40], [41] are used to tune our own backbone classifiers. Another important concept in object detection is the region proposal. In simple words, how does a learning architecture understand that the possible object of interest resides in an image/frame? Based on the literature, we have used the Region Proposal Network (RPN) and Bidirectional Feature Pyramid Network (BiFPN) for proposing such regions of interest (ROIs) in FRCNN and EfficientDet, respectively.

In Algorithm 1, a greedy ensemble detection scheme has been described using pseudo-code. The implementations have been done for the different detection models separately. Once the training is complete, the final models are used to infer the image for object detection. Bounding box combining techniques (NMS, NMW, and WBF) have been applied before the mAP computation to avoid multiple overlapping detections of the same object. The primary models are implemented in every possible combination to evaluate the net mAP from the test image set (see Algorithm 2). We use the concept of *power_set*, where the size of the set is calculated based on the total number of primary detection models. In our case, the primary models are 3, and the correct size of the *power_set* is 7. However, the actual possible combinations are only 4: {ResNet50, EffDet}, {VGG16, EffDet}, {VGG16, ResNet50}, and {VGG16, ResNet50, EffDet}. The remaining are redundant among the first three combinations. In our study, all the primary detectors and all the valid ensemble combinations are explained with their results. Based on the highest obtained mAP value, a set of models is selected for the final ensemble scheme. Both of the given algorithms are simple and self-explanatory.

Once the trained model has been generated, it is used to infer new unknown test images for evaluating the model's performance. The quality of a model is examined based on "Intersection over Union" (IoU) and its admissible threshold value (usually it sets to 0.5). However, $IoU@0.75$ and $IoU@[0.5 : 0.95]$ have also been evaluated to find a trained model's robustness. In $IoU@[0.5 : 0.95]$, the mean average precision (mAP) is repeatedly calculated from $IoU@0.5$ to $IoU@0.95$ with a 0.5 step size. Then, the average has been considered a robust performance metric.

IV. DATASET PREPARATION AND EXPERIMENTAL SETUP

A. DATASET

The dataset of 3698 images with 4703 annotated objects has been prepared for human face and gun detection.[1] The major images have been taken from the Internet Movie Firearms Database [42] and the WIDER FACE dataset [43]. The dataset contains images ranging in resolution from 1851 × 2190 to 259 × 194. However, the raw input images are resized to 224 × 224 before being fed to a detection model. It contains mostly JPEG formatted images, but a few PNG-formatted images are also present. All non-JPEG images are transformed accordingly. There are various types of images in the prepared dataset; some have many faces and gun objects, some have many faces, but single gun objects, and some contain either faces or gun objects.

[1] Face and Gun Dataset:https://github.com/cserajdeep/Face-and-Gun-Dataset

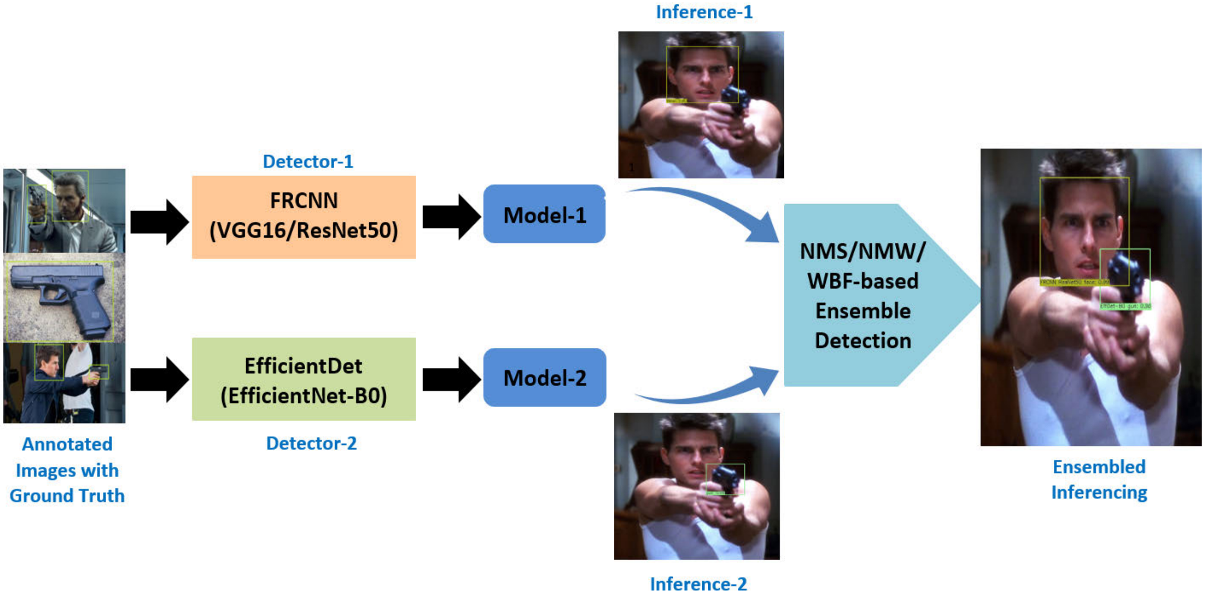

Inference-1
Detector-1
FRCNN (VGG16/ResNet50)
Model-1
NMS/NMW/ WBF-based Ensemble Detection
EfficientDet (EfficientNet-B0)
Model-2
Annotated Images with Ground Truth
Detector-2
Inference-2
Ensembled Inferencing

FIGURE 2. Proposed NMS/NMW/WBF-based ensemble detection scheme for an improved object identification.

Algorithm 1 Greedy Ensemble Object Detection (ENSD)

```
Start
    # (n: # of trained detection models)
Inputs : $m_i$, $i = 1, 2, \dots, n$
         ImgSet, test image-set
    # ensemble
Output : ENSD, set of the selected models
$M \leftarrow \{m_1, m_2, \dots, m_n\}$ and $net\_mAP \leftarrow 0$
$ensemble \leftarrow \{\}$ and $best\_mAP \leftarrow 0$
$discards \leftarrow \{\}$
    # exclude the null element
$power\_set \leftarrow P(M) - \{\emptyset\}$
    # -1 due to the null element
$set\_size \leftarrow sizeof(M)$ and $iter \leftarrow 2^{set\_size} - 1$
for $t := 1$ to $iter$ do
    $mod\_set \leftarrow power\_set(t)$
    if $mod\_set \notin discards$ then
        # calculating the mean average precision
        $net\_mAP \leftarrow compute\_mAP(mod\_set, ImgSet)$
        if $net\_mAP > best\_mAP$ then
            $best\_mAP \leftarrow net\_mAP$
            $ensemble \leftarrow mod\_set$
        else
            $discards \leftarrow mod\_set$
        end if
    end if
end for
$ENSD \leftarrow ensemble$
return $ENSD$
```

Algorithm 2 $compute\_mAP(mod\_set, ImgSet)$

```
Inputs : $img_i, \in ImgSet$
         $sm_j \in mod\_set$
Output : $net\_mAP$
$total\_mods \leftarrow length(mod\_set)$
$boxes \leftarrow \{\}$
for $t := 1$ to $total\_mods$ do
    # the model detects the bounding boxes for an ob-
    # ject
    $boxes \leftarrow inference(img_{\forall i}, sm_t) \cup boxes$
end for
    # provides one final box from all overlapping boxes
$final\_boxes \leftarrow NMS\_or\_NMW\_or\_WBF(boxes)$
$net\_mAP \leftarrow mAP(final\_boxes, ground\_truth\_boxes)$
return $net\_mAP$
```

B. IMAGE PRE-PROCESSING

Some of the prepared dataset images are in PNG format, which has 4 channels (RGB, and Alpha). The PNG-formatted images have been converted to 3 channels JPG formatted images. Besides, we have discarded the gray input images from training. The data augmenter is used for horizontal flips, vertical flips, and 90° rotation of the input images. The Microsoft open-source Visual Object Tagging Tool[2] (VoTT) has been used for image annotation.

C. EXPERIMENTAL SETUP

The models are trained and validated using 2000 epochs and 200 steps per epoch to build them.[3] On average, it takes two and a half days in our system (system configuration discussed below). Total 84 separate images have been used for evaluating the average detection time and calculating

[2] VoTT:https://github.com/microsoft/vott

[3] All models have been built using Tensorflow https://www.tensorflow.org and Keras https://keras.io/

mAP[4] for the objects. The overall mAPs are also computed for each of the detection models, that are, FRCNN[5] (VGG16 and Resnet-50) and EfficientDet-B0.[6] The codes are modified based on our requirements. A stacked ensemble technique has been employed to combine the detection outcomes of multiple detectors. As each detector gives us faces and guns for the same image, combining multiple bounding boxes is a problem. NMS[7] is used to retain the largest bounding box for an object, and NMW/WBF[8] are used to obtain the weighted averaged bounding box from all the predicted boxes. A summary of the details has been given in Table 1 (where ENSD-# indicates the ensemble detection scheme).

TABLE 1. Configuration of different object detection schemes.

Acronym	Detection Type	Backbone	Proposal Network
FRCNN VGG16	Faster RCNN	VGG16	Region Proposal Network (RPN)
FRCNN ResNet50	Faster RCNN	ResNet50	RPN
EffDet-B0	Efficient Detector	EfficientNet-B0	Bidirectional Feature Pyramid Network (BiFPN)
ENSD-1	Faster RCNN, EffDet	ResNet50, EfficientNet-B0	RPN, BiFPN
ENSD-2	Faster RCNN, EffDet	VGG16, ResNet50 EfficientNet-B0	RPN, RPN, BiFPN
ENSD-3	Faster RCNN, EffDet	VGG16, EfficientNet-B0	RPN, BiFPN
ENSD-4	Faster RCNN	VGG16 and ResNet50	RPN, RPN

Furthermore, the proposed scheme of detection have been validated on a few Hollywood movie clips taken from famous action movies. It is observed that the proposed ensemble is also working well on these videos.

D. SYSTEM CONFIGURATION

The work has been implemented using Python 3.6 and Tensorflow-GPU 1.14 on an Intel(R) Core(TM) $i7-9750H$ CPU (9^{th} Gen.) 2.60GHz, 16GB RAM and 6 GB NVIDIA GeForce RTX 2060 with 64 bits Windows 10 Home operating system.

Multiple types of detectors are used in ensemble either using hard NMS or NMW or WBF at the post-processing step. A visualization of the proposed NMS/NMW/WBF-based ensemble (object) detection has been shown in Fig. 2.

V. RESULTS DISCUSSION AND ANALYSIS

As discussed in section III-G, different object detection techniques have been implemented and compared based on the mAP. Furthermore, the mAPs are computed at 0.5, 0.75, and [0.5 : 0.95] to evaluate the model's performance.

A. EVALUATION OF ENSEMBLE MODELS

The primary model have taken a significant time to achieve best results (see Table 2). A comparison of the results based on the calculated mAP (%) is given in Table 3. The results suggest that the ensemble detection schemes outperform all three primary models based on mAP metrics. Another observation is that post-detection combining techniques play an important role in the final identification of an object. The obtained $mAP_{0.50}$ values are 77.02, 71.97, 63.59 and 62.11 for the ENSD-2+WBF, ENSD-3+NMW, ENSD-1+NMS, and primary models, respectively. The performance of all the used object detection techniques is poor at $mAP_{0.75}$; the possible reason could be less training. The results of $mAP_{[0.5:0.95]}$ is similar to the findings at $mAP_{0.50}$. However, the overall results indicate a positive and consistent trend for the proposed ensemble detection schemes. The best of the best performances have been obtained with the ENSD-2 and WBF combination. It provides us with 77.02, 15.49 and 29.73 mean average precision values at $mAP_{0.50}$, $mAP_{0.75}$ and $mAP_{[0.5:0.95]}$. The results give us the following observations:

$$\text{WBF} \succ \text{NMW} \succ \text{NMS} \succ \text{Primary Models}$$

TABLE 2. Training time taken by the primary models.

Detection Technique	Training Time (Hrs.)
Faster RCNN VGG16	36.50
Faster RCNN ResNet50	48.30
EfficientDet-B0	30.70

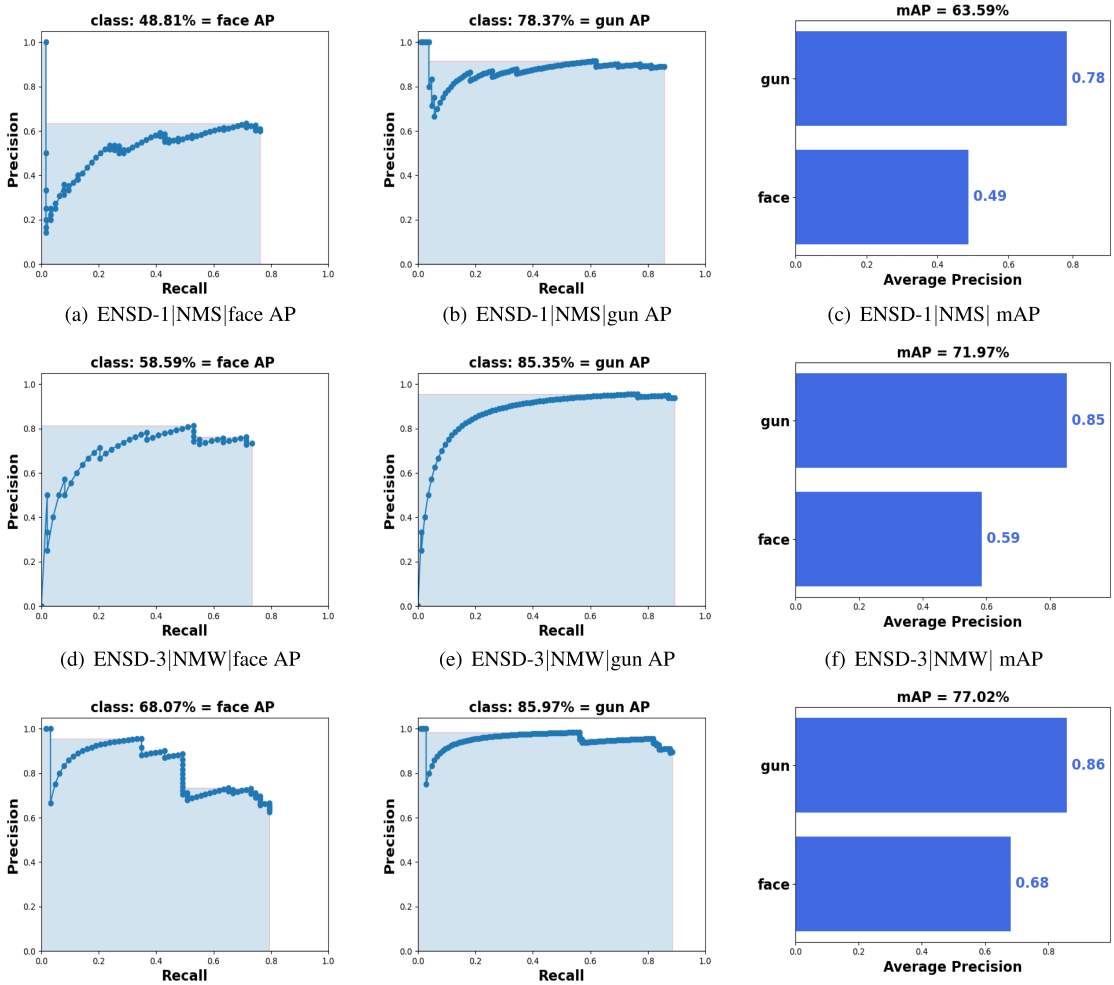
TABLE 3. Comparative mAP (%) results obtained using different object detection schemes (higher is good).

Ensemble	Detection Technique	$\mathbf{mAP_{0.50}}$	$\mathbf{mAP_{0.75}}$	$\mathbf{mAP_{[0.50:0.95]}}$
Primary Models	**FRCNN ResNet50**	**62.11**	**11.58**	**22.74**
	FRCNN VGG16	58.70	05.54	16.87
	EffDet-B0	57.62	11.56	22.57
NMS	**ENSD-1**	**63.59**	**12.78**	**23.52**
	ENSD-2	61.63	8.49	20.73
	ENSD-3	61.77	9.27	21.51
	ENSD-4	58.34	8.06	19.70
NMW	**ENSD-1**	65.49	**14.49**	25.43
	ENSD-2	65.74	12.55	24.38
	ENSD-3	**71.97**	12.12	**25.96**
	ENSD-4	62.50	11.16	23.03
WBF	**ENSD-1**	72.36	**16.40**	28.67
	ENSD-2	**77.02**	15.49*	**29.73**
	ENSD-3	76.44	13.89	29.22
	ENSD-4	69.41	12.18	25.32
Best	***ENSD-2+WBF***	***77.02***	***15.49***	***29.73***

†ENSD-2 has been considered best performing as it gives 2 highest mAP results out of 3 over ENSD-1 which has only one best mAP score.

The class-wise (that is, face and gun) performances at $mAP_{0.5}$ for the said best-performing detectors for NMS, NMW, and WBF are shown in Fig. 3 (a)-(i). The average precision plots for *face* class have been shown in Fig. 3 (a), (d) and (g). Similarly, the average precision plots for *gun* class have been depicted in Fig. 3 (b), (e) and (h). Again, the overall results for each best performing ensemble detectors are given in Fig. 3 (c), (f), and (i) for ENSD-1+NMS,

[4] mAP:https://github.com/Cartucho/mAP
[5] FRCNN:https://github.com/kbardool/keras-frcnn
[6] EfficientDet:https://github.com/xuannianz/EfficientDet
[7] NMS:https://github.com/bruceyang2012/nms_python
[8] NMW/WBF:https://github.com/ZFTurbo/Weighted-Boxes-Fusion



(a) ENSD-1|NMS|face AP (b) ENSD-1|NMS|gun AP (c) ENSD-1|NMS| mAP

(d) ENSD-3|NMW|face AP (e) ENSD-3|NMW|gun AP (f) ENSD-3|NMW| mAP

(g) ENSD-2|WBF|face AP (h) ENSD-2|WBF|gun AP (i) ENSD-2|WBF| mAP

FIGURE 3. Comparative analysis of best performing ensembles [model | ensemble | metric] on mean Average Precision ($mAP_{0.5}$).

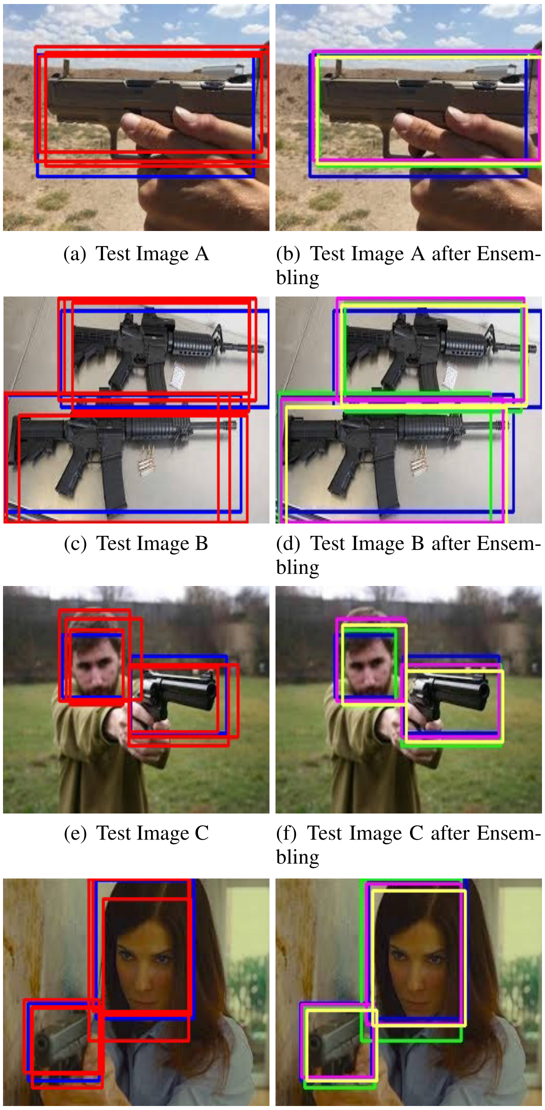
ENSD3+NMW and ENSD-2+WBF, respectively. The performance of the proposed scheme has been well established, as the ENSD-2+WBF achieves a mAP of 77.02 over the best performing mAP of 62.11 of the primary model, FRCNN ResNet50 (other used alternatives). The study is also validated at $mAP_{0.75}$ and $mAP_{[0.5:0.95]}$. The improvements in mAPs are 14.91, 3.91 and 6.99.

B. EVALUATION OF BOUNDING BOX ENSEMBLE

Usually, if the decision classes are numerical, the majority-voting or weighted average is suitable for combining the ensemble outputs. As in our case, the output itself is objects within an image; the methods mentioned earlier do not fit. Even if the majority-voting strategy for object identification has been implemented, it is difficult to detect the appropriate bounding boxes with the highest coverage area. Therefore, NMS, NMW, and WBF are used to retain a single box from multiple bounding boxes around an object at post-processing. The use of NMS at processing is not new. However, the ensemble of FRCNN and EffcientDet object detection architectures with NMW and WBF combining techniques is the novelty of this paper, specifically in this domain. Some test case images before and after the implementation of ensemble combining techniques are shown in Fig. 4. The Fig. 4 (a), (c), (e), and (g) are the test images obtained after use of the ensemble detection scheme. It shows the ground truth (annotated) boxes in blue and the predicted boxes in red. Similarly, Fig. 4 (b), (d), (f), and (h) are the final images with the final bounding boxes are in green, yellow, and pink from NMS, NMW and, WBF, respectively. However, the blue-colored ground truth bounding box is also added for easy understanding. The images are given in two sets; otherwise, the clarity of visualizing the comparison may be diluted.

C. EXPERIMENT BASED ON REAL-WORLD MOVIE FRAMES

In this paper, multiple relevant examples have been examined with our proposed ensemble detection scheme. The proposed



(a) Test Image A (b) Test Image A after Ensembling

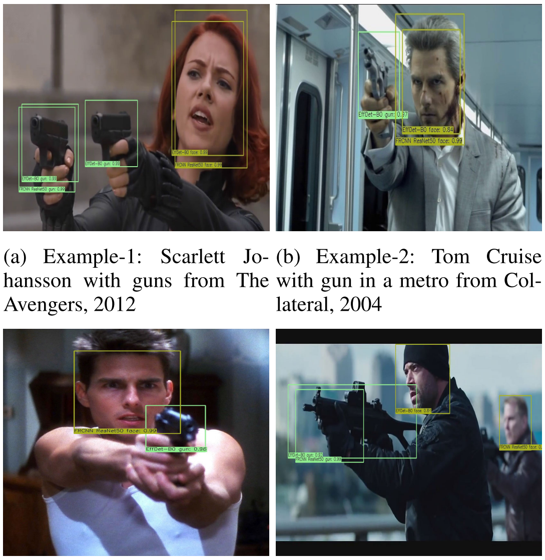
(c) Test Image B (d) Test Image B after Ensembling

(e) Test Image C (f) Test Image C after Ensembling

(g) Test Image D (h) Test Image D after Ensembling

FIGURE 4. Visualization of some test images with the ground truth (blue) and predicted boxes (red); same images with final bounding boxes obtained from NMS (green) / NMW (yellow) / WBF (pink) (subset of the 84 test images used in computing the mAP).

approach has been validated in different Hollywood action movie clips. The Avengers[9] (2012), Collateral[10] (2004), Mission Impossible[11] (1996) and Deadpool[12] (2016) movie clips are taken from the YouTube. Some frames are shown in Fig 5 (a)-(d) before applying the ensemble technique for combining both the detections. From the example frames, it is clear that in reality, if one detector fails to identify objects, the possibility of detection increases using an ensemble approach of multiple types of detectors. Multiple uses of the same detectors in the fixed experimental configuration may provide similar outcomes; thus, using multiple types of detectors to bring diversity in the detection process. It is observed that the trained FRCNN ResNet50 model provides better face detection outcomes, and EffDet-B0 performs well in gun detection based on our experimental setup (here, only two detectors have been shown to avoid congestion based on multiple overlapping boxes). The purpose of this depiction is to realize the benefit of an ensemble detection scheme.



(a) Example-1: Scarlett Johansson with guns from The Avengers, 2012

(b) Example-2: Tom Cruise with gun in a metro from Collateral, 2004

(c) Example-3: Tom Cruise from Mission Impossible, 1996

(d) Example-4: Two unknown actors with guns from Deadpool, 2016

FIGURE 5. Combining of FRCNN ResNet50 and EffDet-B0 for ensemble detection scheme-1 (ENSD-1).

D. EVALUATION BASED ON FRAMES PER SECONDS

The Frames Per Second (FPS) have been computed for all the detectors used in this paper. The obtained FPS are 14, 3, 10 for the EffDet-B0, FRCNN-based ResNet50, and VGG16. The test time for the ensemble detection schemes is some of the time taken by its detectors. Thus, the computed FPS is 3, 2, 6, and 3 for ensemble detectors ENSD-1, ENSD-2, ENSD-3, and ENSD-4, respectively. There is a trade-off between the mAP and the FPS. In Fig. 6, one can find the FPS plots for all seven used detectors as green bars.

E. EVALUATION OF TEST TIME

Furthermore, the test time has been computed for three primary models and four ensemble models using the total 84 test images. Particularly, the test time values for ensemble models exclude the time taken by a combining technique (that is, NMS, NMW, and WBF) as their contribution is in milliseconds (< 20 ms.). The obtained results are shown in the same dual plots Fig. 6 using orange bars. The test time of EffDet-B0 for a single image is better (0.06 sec.) than the ensemble scheme ENSD-3 (0.17 sec.).

[9] The Avengers (2012):https://www.youtube.com/watch?v=HWUsqlUejts (2:49–2:52 mins.)

[10] Collateral (2004):https://www.youtube.com/watch?v=EMS4lYA-hEo

[11] Internet Movie Firearms Database (Mission Impossible, 1996):http://www.imfdb.org/wiki/Mission:_Impossible_(1996)

[12] Deadpool (2016):https://www.youtube.com/watch?v=tLmStxxzhkI

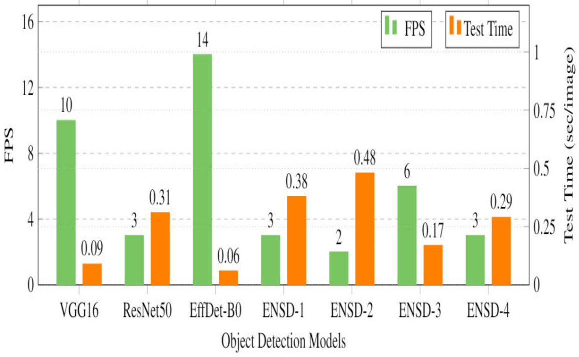


FIGURE 6. **Frames per second (FPS) and Test Time (sec/img) obtained from different used detectors (Table 3).**

The proposed work has shown empirical superiority over non-ensemble methods. However, we feel that some of the points need to be addressed in the continuation, which will be the next work. Some of those issues are:

- ⋆ Dataset preparation, such as different types of image augmentation, has not been examined
- ⋆ Detection of objects (faces and guns) in a low-light scenario is not tested.
- ⋆ Currently, the frame rates are not very encouraging

VI. CONCLUSION

In this paper, rigorous empirical results show how a deep learning-based framework can be used to find faces and guns. It has been seen that the WBF-based ensemble object detection scheme makes it easier to find faces and guns. But the FRCNN architecture makes a model file that is 108 MB in size, while EffDet-B0 only makes a 16 MB model file. So, using EffDet with different sized EfficientNets as the backbone of the lightweight ensemble could be helpful. Again, FRCNN affects the FPS values in ensembles. It can be addressed by using a primary detection architecture with a faster frame rate. Even if the individual models are weak, their ensemble performs better due to the inherent diversity. The work can augment CCTV-based monitoring and transform it into an intelligent surveillance system. The prime aim of the work is to make society safer for our family and friends.

In the future, the work can be deployed for use in real-time testing. There is scope for research on its reliability and scalability in the real world. This can also be combined with other violent and non-violent activities as well as detection of other weapons (such as knives, swords, etc.) to make a more robust intelligent surveillance mechanism. Thus, we can help society become safer.

REFERENCES

[1] A. Remuzzi and G. Remuzzi, "COVID-19 and Italy: What next?" *Lancet*, vol. 395, no. 10231, pp. 1225–1228, Apr. 2020.

[2] C. Sohrabi, Z. Alsafi, N. O'Neill, M. Khan, A. Kerwan, A. Al-Jabir, C. Iosifidis, and R. Agha, "World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19)," *Int. J. Surg.*, vol. 76, pp. 71–76, Apr. 2020.

[3] C. Cabrera-Arnau, R. Prieto Curiel, and S. R. Bishop, "Uncovering the behaviour of road accidents in urban areas," *Roy. Soc. Open Sci.*, vol. 7, no. 4, Apr. 2020, Art. no. 191739.

[4] D. A. Brent, M. J. Miller, R. Loeber, E. P. Mulvey, and B. Birmaher, "Ending the silence on gun violence," *J. Amer. Acad. Child Adolescent Psychiatry*, vol. 52, no. 4, pp. 333–338, Apr. 2013.

[5] J. A. Fox and M. J. DeLateur, "Mass shootings in America: Moving beyond newtown," *Homicide Stud.*, vol. 18, no. 1, pp. 125–145, Feb. 2014.

[6] S. Jaffe, "Decisions to be made on U.S. gun violence research funds," *Lancet*, vol. 395, no. 10222, pp. 403–404, Feb. 2020.

[7] M. E. Smith, T. L. Sharpe, J. Richardson, R. Pahwa, D. Smith, and J. DeVylder, "The impact of exposure to gun violence fatality on mental health outcomes in four urban U.S. settings," *Social Sci. Med.*, vol. 246, Feb. 2020, Art. no. 112587.

[8] T. M. Stein, *Mass Shootings*, D. E. Hogan and J. L. Burstein, Eds., 2nd ed., ch. 37, pp. 444–451.

[9] (Aug. 2019). *America's Gun Culture in Charts*. Accessed: Feb. 6, 2023. [Online]. Available: https://www.bbc.com/news/world-us-canada-41488081

[10] (Sep. 2018). *Gun Violence*. Accessed: Feb. 6, 2023. [Online]. Available: https://www.amnesty.org/en/what-we-do/arms-control/gun-violence/

[11] (Sep. 2017). *Hoddle Street Massacre*. Accessed: Feb. 6, 2023. [Online]. Available: https://www.abc.net.au/news/2017-08-09/hoddle-street-massacre-30-years-on/8786766/

[12] (Mar. 2019). *Christchurch Massacre*. Accessed: Feb. 6, 2023. [Online]. Available: https://www.bbc.com/news/world-asia-47578798

[13] (Jan. 2017). *Quebec City Shoot-Out*. Accessed: Feb. 6, 2023. [Online]. Available: https://www.bbc.com/news/world-us-canada-38793071

[14] (Jul. 2011). *Oslo and Utøya Island Massacre*. Accessed: Feb. 6, 2023. [Online]. Available: https://www.theguardian.com/world/2011/jul/23/norway-attacks

[15] (May 2014). *Lod Airport Massacre Shoot-Out*. Accessed: Feb. 6, 2023. [Online]. Available: https://www.bbc.com/news/av/magazine-27468978/i-survived-the-israeli-airport-massacre

[16] (Nov. 2018). *26/11 Mumbai Attack*. Accessed: Feb. 6, 2023. [Online]. Available: https://mumbaimirror.indiatimes.com/mumbai/other/10-years-of-2008-mumbai-terror-attacks-all-you-need-to-know-about-the-26/11-siege-that-shook-mumbai/articleshow/66795739.cms

[17] (Aug. 2012). *What Exactly is a Mass Shooting?* Accessed: Feb. 6, 2023. [Online]. Available: https://www.motherjones.com/crime-justice/2012/08/what-is-a-mass-shooting/

[18] H. Kruegle, *CCTV Surveillance: Video Practices and Technology*. Amsterdam, The Netherlands: Elsevier, 2011.

[19] D. Trottier, "Crowdsourcing CCTV surveillance on the internet," *Inf., Commun. Soc.*, vol. 17, no. 5, pp. 609–626, May 2014.

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[21] S. Targ, D. Almeida, and K. Lyman, "ResNet in ResNet: Generalizing residual architectures," 2016, *arXiv:1603.08029*.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[23] Z. Wu, C. Shen, and A. Van Den Hengel, "Wider or deeper: Revisiting the ResNet model for visual recognition," *Pattern Recognit.*, vol. 90, pp. 119–133, Jun. 2019.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[25] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10781–10790.

[26] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*.

[27] A. Egiazarov, V. Mavroeidis, F. M. Zennaro, and K. Vishi, "Firearm detection and segmentation using an ensemble of semantic neural networks," in *Proc. Eur. Intell. Secur. Informat. Conf. (EISIC)*, Nov. 2019, pp. 70–77.

[28] S. Akçay, M. E. Kundegorski, M. Devereux, and T. P. Breckon, "Transfer learning using convolutional neural networks for object classification within X-ray baggage security imagery," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1057–1061.

[29] R. Olmos, S. Tabik, and F. Herrera, "Automatic handgun detection alarm in videos using deep learning," *Neurocomputing*, vol. 275, pp. 66–72, Jan. 2018.

[30] (Feb. 2017). *Region Proposal Network*. Accessed: Jun. 20, 2022. [Online]. Available: https://blog.deepsense.ai/region-of-interest-pooling-explained/

[31] (Nov. 2019). *EfficientNet*. Accessed: Jun. 20, 2022. [Online]. Available: https://keras.io/api/applications/efficientnet/

[32] R. Padilla, S. L. Netto, and E. A. B. da Silva, "A survey on performance metrics for object-detection algorithms," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Jul. 2020, pp. 237–242.

[33] P. Henderson and V. Ferrari, "End-to-end training of object class detectors for mean average precision," in *Proc. Asian Conf. Comput. Vis.* Taipei, Taiwan: Springer, 2016, pp. 198–213.

[34] J. Revaud, J. Almázan, R. Rezende, and C. D. Souza, "Learning with average precision: Training image retrieval with a listwise loss," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5107–5116.

[35] (Jul. 2020). *Average Precision*. Accessed: Jun. 25, 2022. [Online]. Available: https://github.com/rafaelpadilla/Object-Detection-Metrics

[36] M. B. Blaschko, J. Kannala, and E. Rahtu, "Non maximal suppression in cascaded ranking models," in *Proc. Scand. Conf. Image Anal.* Berlin, Germany: Springer, 2013, pp. 408–419.

[37] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4507–4515.

[38] H. Zhou, Z. Li, C. Ning, and J. Tang, "CAD: Scale invariant framework for real-time object detection," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 760–768.

[39] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," 2019, *arXiv:1910.13302*.

[40] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," 2016, *arXiv:1604.02201*.

[41] M. Huh, P. Agrawal, and A. A. Efros, "What makes ImageNet good for transfer learning?" 2016, *arXiv:1608.08614*.

[42] (Mar. 2008). *Internet Movie Firearms Database*. Accessed: May 17, 2021. [Online]. Available: http://www.imfdb.org/wiki/Main_Page

[43] (Mar. 2017). *Wider Face Dataset*. Accessed: May 20, 2021. [Online]. Available: http://shuoyang1213.me/WIDERFACE/

RAJDEEP CHATTERJEE (Member, IEEE) received the B.E. degree in computer science and engineering from The University of Burdwan, in 2008, and the M.Tech. and Ph.D. degrees in computer science and engineering from the Kalinga Institute of Industrial Technology (Deemed to be University), India, in 2011 and 2020, respectively. He also qualified for GATE-2008 with an All India Rank 1410.

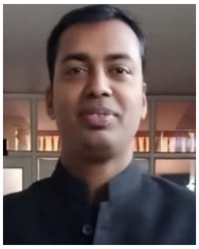
He is currently an Associate Professor with the School of Computer Engineering, Kalinga Institute of Industrial Technology (Deemed to be University). He has published more than 50 research papers in reputed conferences and journals. He regularly reviews research articles from journals, such as IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, *Computers in Biology and Medicine*, and IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE. He has initiated the international conference series on computational intelligence and networks (https://www.cineconf.org). He has co-founded Amygdala AI (https://www.amygdalaai.org). It is a volunteer-driven global open research community. His research interests include machine learning, deep learning, brain–computer interface, and computer vision.

ANKITA CHATTERJEE received the B.Tech. degree in computer science and engineering from WBUT, India, in 2016, and the M.Tech. degree in computer science and engineering from the Kalinga Institute of Industrial Technology (Deemed to be University), Bhubaneswar, India, in 2018.

She is currently a Senior AI Scientist with Amygdala AI. It is a volunteer-driven global open-research community aiming to better tomorrow using AI tools and techniques. It also promotes scientific and engineering awareness with universal human values. Her research interests include machine learning, deep learning, and computer vision.

MANAS RANJAN PRADHAN (Member, IEEE) received the M.Tech. degree in computer science from Utkal University, India, and the Ph.D. degree in computer science from the University of Mysore, India.

He has vast experience in teaching, research, and academic administration in India and abroad. He is currently with Skyline University College, Sharjah, United Arab Emirates. As an Academic Leader, he was the Head of the Program with the University of Petroleum and Energy Studies (UPES), India, and the Dean (Faculty of IT and Science) of INTI International University, Malaysia. He has been with the IT industry for industry-academic collaboration, internships, placements, and workshops. He has executed the IBM Center of Education for Cloud Computing and Business Analytics, INTI International University, under Laureate International Universities, USA. He has presented and published many research papers in various conferences and journals. He has three Indian patents and three Australian patents to his credit. His research interests include business analytics, data mining, data warehouse, retail/e-commerce analytics, artificial intelligence, machine learning, and business process modeling. He has received the Mentor Award for the i-Talent Project Contest from the Confederation of Indian Industry (CII). He has played a vital role in organizing three international conferences, such as NGCT-2015 (UPES), ICQMOIT-2008 (ICFAI, India), and ICD-2019 (SUC, United Arab Emirates).



BISWARANJAN ACHARYA (Senior Member, IEEE) received the M.C.A. degree from IGNOU, New Delhi, India, in 2009, and the M.Tech. degree in computer science and engineering from the Biju Pattanaik University of Technology (BPUT), Rourkela, Odisha, India, in 2012. He is currently pursuing the Ph.D. degree in computer application with the Veer Surendra Sai University of Technology (VSSUT), Burla, Odisha.

He has a total of ten years of experience in both academia with some reputed universities, such as Ravenshaw University, and the software development field. He is an Assistant Professor with the Department of Computer Engineering-Artificial Intelligence and Big Data Analytics. He has published many research articles in internationally reputed journals and serves as a reviewer for many peer-reviewed journals. He has more than 50 patents on his credit. His research interests include multiprocessor scheduling along with different fields, such as data analytics, computer vision, machine learning, and the IoT. He is associated with various educational and research societies, such as IACSIT, CSI, IAENG, and ISC.

TANUPRIYA CHOUDHURY (Senior Member, IEEE) received the bachelor's degree in computer science engineering from the West Bengal University of Technology, Kolkata, India, the master's degree in computer science engineering from Dr. M.G.R. University, Chennai, India, and the Ph.D. degree, in 2016.

He has ten years of experience in teaching and research. He is currently an Associate Professor with the Department of Computer Science Engineering, University of Petroleum and Energy Studies (UPES), Dehradun. He has filed 14 patents to date and received 16 copyrights from MHRD for his software. His research interests include human computing, soft computing, cloud computing, and data mining. He has been associated with many conferences throughout India, as a TPC member and the session chair. He is a Lifetime Member of IETA, a member of IET (U.K.), and other renowned technical societies. Recently, he received the Global Outreach Education Award for Excellence in Best Young Researchers, in GOECA 2018.