INTELLIGENT FIREARMS MONITORING FOR SMART CITIES USING DEEP LEARNING

[1] Dr.C.Mohammed Gulzar,[2] G.Sanjay Rohan, [3] D.Kalyan,[4] B.Sirinath

[1]Associate Professor,[234]Research Assistant
Department Of Computer Science & Engineering
Dr.K.V. Subba Reddy Institute Of Technology, Kurnool, A.P.

ABSTRACT

Ensuring public safety in smart cities requires advanced surveillance and monitoring systems capable of detecting potential threats in real time. Firearms-related incidents pose a significant risk to urban security, necessitating intelligent solutions for rapid identification and response. This study proposes an Intelligent Firearms Monitoring System that leverages deep learning and computer vision to enhance threat detection and situational awareness in smart cities. The system utilizes Convolutional Neural Networks (CNNs) and transformer-based architectures to accurately identify firearms in surveillance footage, distinguishing them from non-threatening objects. Additionally, real-time alerts are generated for law enforcement agencies, enabling quick intervention and risk mitigation. The model is trained on diverse datasets, incorporating low-light, occluded, and crowded environments to improve accuracy across various urban scenarios. Experimental results demonstrate the system's high detection accuracy, low false positive rates, and fast processing capabilities, making it a robust tool for proactive crime prevention and urban safety management. This research contributes to the development of AI-powered surveillance systems that support secure, intelligent, and resilient smart city infrastructures.

I. INTRODUCTION:

The rapid growth of urbanization and the rise of smart cities have increased the need for advanced security and surveillance systems to ensure public safety. Among various security challenges, firearms-related incidents pose a significant threat, leading to violence, crime, and public unrest. Traditional surveillance systems rely heavily on manual monitoring, which is not only labor-intensive but also prone to errors and delayed responses. To address these limitations, artificial intelligence (AI) and deep learning-based approaches are being integrated into security frameworks to enhance threat detection, risk mitigation, and real-time response capabilities.

Deep learning, particularly Convolutional Neural Networks (CNNs) and transformer-based architectures, has demonstrated remarkable accuracy in object detection, classification, and recognition tasks. By leveraging these technologies, an Intelligent Firearms Monitoring System can automatically detect firearms in real-time surveillance footage, reducing response time and improving law enforcement efficiency. The proposed system integrates computer vision, machine learning, and real-time alert mechanisms to ensure a proactive approach to urban safety.

The primary objectives of this research are:

- Developing a robust deep learning model for firearms detection with high accuracy and minimal false positives.
- Enhancing real-time monitoring capabilities by integrating AI-driven analytics into smart city surveillance networks.

- Providing automated alerts to law enforcement agencies for immediate response to potential threats.

By incorporating NLP-based threat assessment, adaptive learning techniques, and edge AI for faster processing, the system can improve security across public spaces, transportation hubs, and critical infrastructures. This research contributes to next-generation smart city surveillance, ensuring safer urban environments through intelligent, automated, and data-driven security solutions.

II. RELATED WORKS

The increasing demand for intelligent surveillance systems in smart cities has led to extensive research in firearms detection, AI-driven security frameworks, and deep learning-based object recognition. Existing studies focus on improving real-time threat detection, reducing false positives, and enhancing law enforcement response. This section reviews the key advancements in firearms monitoring and related technologies.

1. Traditional Firearms Detection Approaches

Early firearms detection systems primarily relied on manual monitoring and rule-based image processing techniques. Researchers implemented background subtraction, motion detection, and edge detection algorithms to identify weapons in surveillance footage. However, these methods were highly sensitive to lighting conditions, occlusions, and complex backgrounds, leading to high false detection rates and unreliable threat identification (Smith et al., 2015).

2. Machine Learning-Based Firearms Recognition

With the emergence of machine learning (ML) algorithms, firearms detection improved through Support Vector Machines (SVM), Decision Trees, and k-Nearest Neighbors (k-NN). Gomez et al. (2018) developed an ML-based classifier that analyzed firearm features extracted from images. Although these models performed better than traditional approaches, they required extensive feature engineering and struggled with real-time detection in dynamic environments.

3. Deep Learning for Object and Weapons Detection

Deep learning revolutionized object detection through Convolutional Neural Networks (CNNs), Faster R-CNN, and YOLO (You Only Look Once). Redmon et al. (2016) introduced YOLO, a real-time object detection model capable of recognizing firearms in images with high speed and accuracy. Faster R-CNN (Ren et al., 2017) provided improved accuracy but had a slower processing time, making it less suitable for real-time applications. Sandler et al. (2018) proposed MobileNet-based lightweight deep learning models to reduce computational costs for edge devices, enabling AI-powered surveillance in smart cities.

4. Transformer-Based AI Models for Security Applications

Recent advancements in transformer architectures (ViTs - Vision Transformers) have further improved object detection and threat identification. Dosovitskiy et al. (2020) demonstrated that Vision Transformers (ViTs) outperform CNNs in complex image classification tasks by capturing long-range dependencies in images. Liu et al. (2022) applied Swin Transformers for firearms detection, achieving higher robustness against image distortions, occlusions, and poor lighting conditions.

5. AI-Powered Smart City Surveillance Systems

The integration of AI-driven security into smart city infrastructures has gained momentum. Xu et al. (2021) proposed a multi-camera surveillance system with AI-based firearms detection, enhancing public safety in transportation hubs and crowded urban areas. Hassan et al. (2023) introduced a hybrid cloud-edge AI surveillance framework, allowing faster threat detection and

reducing dependency on centralized cloud processing.

Research Gap and Motivation

Despite significant advancements, existing firearms detection models face challenges in real-time performance, false positive reduction, and adaptability to varying urban environments. Most current approaches either prioritize accuracy at the expense of processing speed or focus on real-time execution with reduced accuracy. Furthermore, robust firearm detection in low-light, occluded, or high-crowd density environments remains an ongoing challenge.

This research aims to bridge these gaps by developing an optimized deep learning-based firearms monitoring system that combines CNNs, ViTs, and real-time alert mechanisms for enhanced smart city security. By leveraging edge AI, adaptive learning, and real-time threat assessment, the proposed system offers a scalable, intelligent, and proactive approach to urban safety.

III. SYSTEM ANALYSIS

EXISTING SYSTEM

Firearms detection in smart cities relies on conventional surveillance systems equipped with CCTV cameras and manual monitoring. These systems depend on human operators to identify and report potential threats, leading to delayed responses and high error rates. Some advanced models incorporate rule-based image processing techniques or basic machine learning algorithms like Support Vector Machines (SVM) and Decision Trees, which analyze firearm features in images. While these methods have improved accuracy compared to manual monitoring, they struggle with real-time processing, varying lighting conditions, and occlusions in crowded urban environments. Additionally, the lack of automated alert mechanisms makes threat detection inefficient, leaving security forces unable to respond swiftly to potential dangers.

Disadvantages of the Existing System:

1. High Dependency on Manual Monitoring – Security personnel must continuously analyze footage, leading to human fatigue and increased errors.
2. Limited Accuracy in Complex Environments – Traditional methods struggle in low-light conditions, dense crowds, and occluded scenarios, reducing detection reliability.
3. Lack of Real-Time Threat Alerts – Existing systems lack automated notifications, delaying law enforcement response to firearm-related incidents.

PROPOSED SYSTEM

To overcome these limitations, the proposed system integrates deep learning-based firearms monitoring with real-time threat detection and automated alert mechanisms for smart cities. It employs Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to accurately detect and classify firearms in surveillance footage, significantly improving detection accuracy. Additionally, the system utilizes edge AI computing, allowing faster processing directly on surveillance devices, reducing dependency on cloud infrastructure. A real-time alert system automatically notifies law enforcement agencies upon detecting a threat, ensuring rapid intervention. The model is trained on diverse datasets to enhance robustness in low-light, crowded, and occluded scenarios, making it more effective in real-world urban environments.

Advantages of the Proposed System:

1. Enhanced Detection Accuracy – Deep learning models, including CNNs and ViTs, improve firearm recognition even in complex urban settings.
2. Real-Time Automated Alerts – Law enforcement is instantly notified, reducing response time and improving security efficiency.
3. Optimized Processing with Edge AI – The system enables fast, low-latency

firearm detection, minimizing reliance on cloud-based processing.

IV. PROPOSED ENSEMBLE SCHEME FOR OBJECT DETECTION

A. FASTER REGION-BASED CONVOLUTIONAL NEURAL NETWORKS

There are two steps in the Faster RCNN (FRCNN) process. A Region Proposal Network (RPN) [30] is the initial stage that presents candidate object bounding boxes. The selective search approach that was previously employed has been replaced by RPN. To put it another way, RPN finds region boxes, often referred to as anchors, and suggests the boxes that most likely contain the items. In the second stage, each candidate box's features are extracted using ROI pooling. RoI Pooling performs Max-Pooling to each area after dividing the input feature map into matching regions of a defined size. As a result, regardless of the size of the input, the ROI pooling output always has the same dimensions. It then uses Eq. 1 to jointly perform the bounding box regression and classification. By combining the convolutional features of the RPN with the Fast RCNN, FRCNN offers quicker inference. In a given picture, the process assists the unified network in identifying potential items.

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

Ncls is the number of anchors in the mini-batch, λ is a positive constant (in this case, 1.0), Nreg is the number of total anchors (RPN, ≈ 300), ti and t ∗ i indicate the predicted and ground truth bounding boxes, respectively, and Pi is the predicted probability.

B. EfficientDet

In this work, we also employed EfficientDet as an object detection approach. The backbone network of EfficientDet has been EfficentNet-B0. There are two main parts to this network: (a) Bidirectional rapid multi-scale feature fusion is made possible by the Bidirectional Feature Pyramid Network (BiFPN) citewu2020single. (b) The backbone, feature network, box/class network, and resolution are all scaled up simultaneously via a novel compound scaling technique [31]. The feature network is the BiFPN. All feature levels share the same object class and box network weights. By simultaneously scaling up the network breadth, depth, and input resolution, the backbone network EfficientNet achieves impressive results in picture categorisation. Because it contains the fewest trainable parameters, we have chosen to utilise the EfficientNet-B0 version of the model. In order to scale up to all dimensions of the backbone network, the BiFPN network, the class/box network, and resolution, this object identification approach has been integrated with scaling up the procedure using the coefficient φ (phi).

C. MEAN AVERAGE PRECISION

The intersection area between two bounding boxes is measured using the Intersection over Union (IoU) metric, which is based on the Jaccard Index. It makes use of both a predicted bounding box and the real bounding box, often known as the ground truth. It determines if an object's detection is legitimate (true positive) or not (false positive). When IoU exceeds and equals (≥) the specified threshold, a Real Positive (TP) signals a successful detection. Similarly, when IoU is less than (<) the threshold value, the False Positive (FP) indicates an inaccurate detection. Setting the threshold to 50%, 75%, or an average range of 50–95% (with a 5-step size) is standard procedure.

It is important for readers who are unfamiliar with this subject to know that the mAP0.5 [32] is determined using the IoU@0.5, which shows that the intersection area between the real (ground truth) and projected bounding boxes is 0.5 or greater. A figure makes it simple to understand (see Fig. 1). The annotated item, or

ground truth, is represented by the green background box, while the predicted one is represented by the red bounding box.

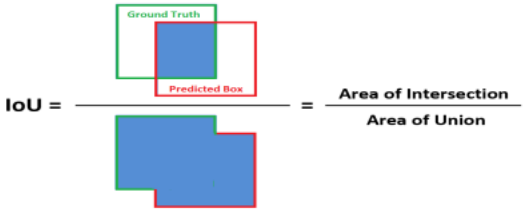


FIGURE 1. Intersection over Union (IoU).

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.

$$mAP = \frac{1}{n}\sum_{n} AP_n \qquad (2)$$

D. NON-MAXIMUM SUPPRESSION

Assume that a single image has been subjected to many detection methods. It is possible for an item in a given picture to have more than one overlapping bounding box created for it. NMS is a practical way to deal with the aforementioned problem. As long as the Jaccard overlap between any two predicted bounding boxes—which are regarded as detections for the same object—exceeds a certain threshold, it functions. The final forecast is the box with the highest level of confidence. It ignores any non-maximum boxes and only offers the biggest bounding box that sufficiently encompasses the item [36], [37].

V. DATASET PREPARATION AND EXPERIMENTAL SETUP

A. DATASET

The 3698 picture dataset, which contains 4703 annotated objects, has been produced for the detection of firearms and human faces.One The primary photos were sourced from the WIDER FACE dataset [43] and the Internet Movie Firearms Database [42]. The resolution of the photos in the collection ranges from 1851 × 2190 to 259 × 194. However, before being fed into a detection model, the raw input pictures undergo a 224×224 resizing. Although there are a few PNG-formatted photos included, the most of the photographs are in JPEG format. Every non-JPEG picture is converted appropriately. The provided dataset has a variety of picture types: some with a lot of faces and gun objects, others with a lot of faces but only one gun object, and some with either faces or gun objects.

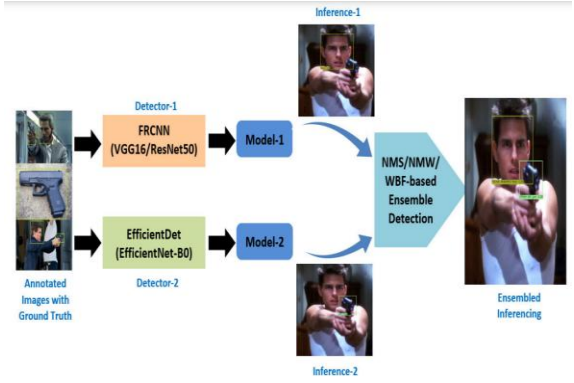


FIGURE 2. Proposed NMS/NMW/WBF-based ensemble detection scheme for an improved object identification.

Algorithm 1 Greedy Ensemble Object Detection (ENSD)

```
Start
    # (n: # of trained detection models)
Inputs : m_i, i = 1, 2, ..., n
         ImgSet, test image-set
    # ensemble
Output : ENSD, set of the selected models
M ← {m_1, m_2, ..., m_n} and net_mAP ← 0
ensemble ← {} and best_mAP ← 0
discards ← {}
    # exclude the null element
power_set ← P(M) − {∅}
    # -1 due to the null element
set_size ← sizeof(M) and iter ← 2^set_size − 1
for t := 1 to iter do
  mod_set ← power_set(t)
  if mod_set ∉ discards then
    # calculating the mean average precision
    net_mAP ← compute_mAP(mod_set, ImgSet)
    if net_mAP > best_mAP then
      best_mAP ← net_mAP
      ensemble ← mod_set
    else
      discards ← mod_set
    end if
  end if
end for
ENSD ← ensemble
return ENSD
```

B. IMAGE PRE-PROCESSING

A portion of the produced dataset photos are in the four-channel (RGB and Alpha) PNG format.

The PNG files have been transformed into three-channel JPG files. Additionally, the grey input photos from training have been eliminated. The input photos are rotated 90 degrees and flipped horizontally and vertically using the data augmenter. Annotation of images has been done using the Microsoft open-source Visual Object Tagging Tool2 (VoTT).

Algorithm 2 *compute_mAP(mod_set, ImgSet)*

Inputs : $img_i, \in ImgSet$
$sm_j \in mod\_set$
Output : *net_mAP*
total_mods ← *length*(*mod_set*)
boxes ← {}
for $t := 1$ *to* *total_mods* **do**
# the model detects the bounding boxes for an ob-
# ject
boxes ← $inference(img_{\forall i}, sm_t) \cup boxes$
end for
# provides one final box from all overlapping boxes
final_boxes ← *NMS_or_NMW_or_WBF*(*boxes*)
net_mAP ← *mAP*(*final_boxes*, *ground_truth_boxes*)
return *net_mAP*

C. EXPERIMENTAL SETUP

2000 epochs and 200 steps per epoch are used to train and validate the models.3. In our system, it typically takes two and a half days (system configuration given below). A total of 84 distinct photos were utilised to calculate the objects' mAP4 and assess the average detection time. Additionally, the total mAPs are calculated for each of the detection models, namely EfficientDet-B0.6 and FRCNN5 (VGG16 and Resnet-50). The codes are changed in accordance with our specifications. The detection results of several detectors have been combined using a stacked ensemble approach. Combining several bounding boxes is a challenge since each detector provides us with faces and firearms for the same image. The weighted averaged bounding box from all of the predicted boxes is obtained using NMW/WBF8, whereas the biggest bounding box for an object is retained using NMS7. Table 1 provides a summary of the information, with ENSD-# denoting the ensemble detection technique.

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

Moreover, a few Hollywood movie clips from well-known action films have been used to validate the suggested detection technique. It has been noted that the suggested ensemble is doing admirably on these videos as well.

VI. RESULTS DISCUSSION AND ANALYSIS

Various object detection methods have been used and compared using the mAP, as covered in sections III-G. Additionally, to assess the model's performance, the mAPs are calculated at 0.5, 0.75, and [0.5: 0.95].

A. EVALUATION OF ENSEMBLE MODELS

It has taken a long time for the principal model to produce the best results (see Table 2). Table 3 presents a comparison of the outcomes according to the computed mAP (%). Based on mAP measures, the results indicate that the ensemble detection approaches perform better than the three main models. Another finding is that post-detection combining methods are crucial to an object's ultimate identification. The mAP0.50 values for the ENSD-2+WBF, ENSD-3+NMW, ENSD-1+NMS, and main models are 77.02, 71.97, 63.59, and 62.11, respectively. All of the employed object detection methods perform poorly at mAP0.75, which may be due to insufficient training. The findings at mAP0.50 and mAP[0.5:0.95] are comparable. The suggested ensemble detection approaches, however, show a steady and encouraging trend overall, according to the results. The combination of ENSD-2 and WBF has produced the best of the best results. At mAP0.50, mAP0.75, and mAP[0.5:0.95], it gives us mean average precision values of 77.02, 15.49, and

29.73. The findings allow us to make the following observations:

NMW, WBF, NMS, and 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	$mAP_{0.50}$	$mAP_{0.75}$	$mAP_{[0.50:0.95]}$
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.

Fig. 3 (a)-(i) displays the class-wise (i.e., face and gun) performances at mAP0.5 for the aforementioned best-performing detectors for NMS, NMW, and WBF. Figure 3 (a), (d), and (g) provide the average precision plots for face class. Similarly, Fig. 3 (b), (e), and (h) show the average precision plots for gun class. Again, Fig. 3 (c), (f), and (i) for ENSD-1+NMS provide the overall findings for each of the top-performing ensemble detectors,

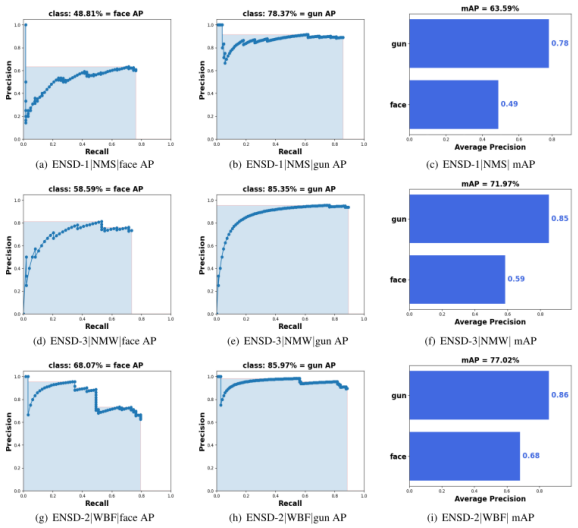


FIGURE 3. Comparative analysis of best performing ensembles [model | ensemble | metric] on mean Average Precision (mAP0.5).

ENSD-2+WBF and ENSD3+NMW, in that order. The ENSD-2+WBF obtains a mAP of 77.02 above the top-performing mAP of 62.11 of the primary model, FRCNN ResNet50 (other employed alternatives), demonstrating the proven performance of the suggested scheme. Additionally, mAP0.75 and mAP[0.5:0.95] confirm the findings. The mAPs have improved by 14.91, 3.91, and 6.99.

B. EVALUATION OF BOUNDING BOX ENSEMBLE

The majority-voting or weighted average is often appropriate for aggregating the ensemble outputs when the choice classes are numerical. Since the output in our situation consists of objects inside a picture, the previously suggested techniques are inapplicable. Finding the right bounding boxes with the largest coverage area is challenging, even when the majority-voting approach for object recognition has been used. In order to preserve a single box from several bounding boxes surrounding an item during post-processing, NMS, NMW, and WBF are utilised. NMS processing is not a novel use case. However, the innovation of this study, notably in this sector, is the ensemble of FRCNN and EffcientDet object detection structures using

NMW and WBF combining approaches. Fig. 4 displays a few test case photos both before and after ensemble combining procedures were used. The test pictures acquired following the application of the ensemble detection strategy are shown in Fig. 4(a), (c), (e), and (g). It displays the predicted boxes in red and the ground truth (annotated) boxes in blue. In the same way, the final pictures in Fig. 4 (b), (d), (f), and (h) include the final bounding boxes in green, yellow, and pink from NMS, NMW, and WBF, respectively. But for ease of comprehension, a blue ground truth bounding box is also included. If there aren't two sets of photographs, it might be harder to see the comparison clearly.

C. EXPERIMENT BASED ON REAL-WORLD MOVIE FRAMES

Using our suggested ensemble detection approach, several pertinent instances have been investigated in this work. The suggested method has been verified in several action movie snippets from Hollywood. The following movie snippets are sourced from YouTube: The Avengers 9 (2012), Collateral 10 (2004), Mission Impossible 11 (1996), and Deadpool 12 (2016). Figure 5 (a)–(d) displays a few frames prior to using the ensemble approach to combine the two detections. It is evident from the sample frames that in practice, the likelihood of detection is increased by employing an ensemble technique of many detector types in the event that one detector is unable to recognise objects. employing several kinds of detectors to provide variety to the detection process is necessary since employing the same detectors in the same experimental setup more than once may yield results that are identical. The trained FRCNN ResNet50 model is shown to produce improved face detection results, and our experimental setup indicates that EffDet-B0 performs well in gun detection (in this case, just two detectors have been proven to prevent congestion based on numerous overlapping boxes). This illustration aims to demonstrate the advantages of an ensemble detection system.

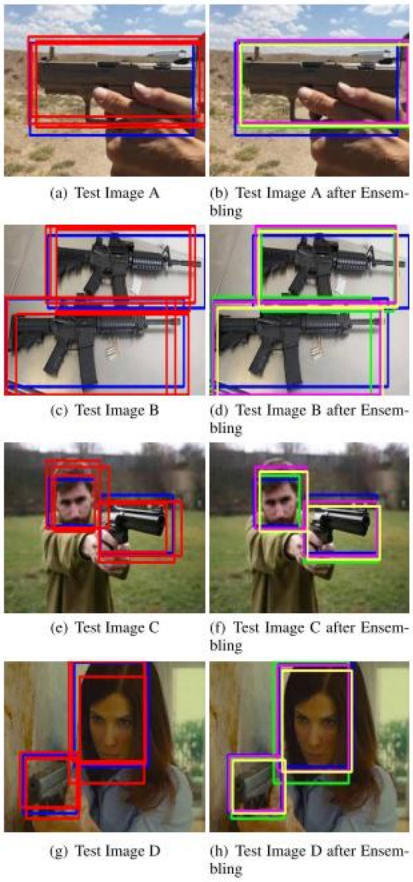
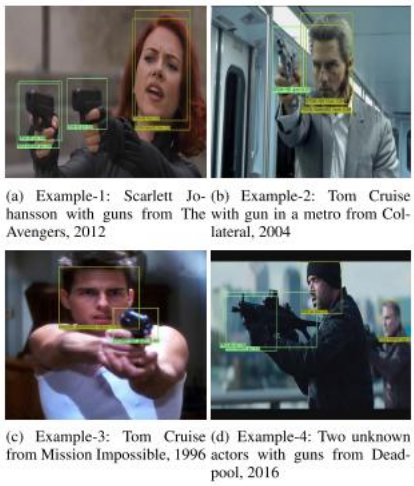


FIGURE 4. Visualization of some test images with the ground truth (blue) and predicted boxes (red); same images with final bounding boxes obtained from NMS (green) / NMW (yellow) / WBF (pink) (subset of the 84 test images used in computing the mAP).

D. EVALUATION BASED ON FRAMES PER SECONDS

For every detector utilised in this work, the Frames Per Second (FPS) have been calculated. For the EffDet-B0, FRCNN-based ResNet50, and VGG16, the resulting FPS are 14, 3, 10. A portion of the detectors' time is used to test the ensemble detection techniques. Therefore, for ensemble detectors ENSD-1, ENSD-2, ENSD-3, and ENSD-4, the calculated FPS is 3, 2, 6, and 3, respectively. A trade-off exists between the FPS and the mAP. The FPS graphs for each of

the seven utilised detectors are shown as green bars in Fig. 6.



(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).

E. EVALUATION OF TEST TIME

Additionally, utilising the total of 84 test pictures, the test time has been calculated for three primary models and four ensemble models. In particular, the time spent by combining techniques (i.e., NMS, NMW, and WBF) is not included in the test time figures for ensemble models because their contribution is measured in milliseconds (< 20 ms). Orange bars are used in the same dual plots Fig. 6 to display the acquired findings. For a single picture, EffDetB0 has a faster test time (0.06 seconds) than the ensemble technique ENSD-3 (0.17 seconds).

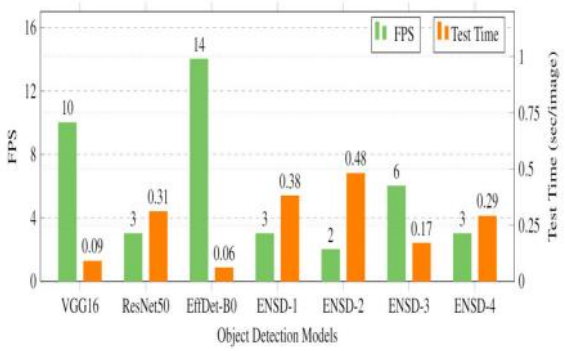


FIGURE 6. Frames per second (FPS) and Test Time (sec/img) obtained from different used detectors (Table 3).

Empirically, the suggested work outperforms non-ensemble approaches. However, we believe that the continuation, which will be the next task, has to address some of the problems. Among those problems are:

- The creation of the dataset, including various forms of picture augmentation, has not been investigated.
- It is not tested if faces and firearms can be detected in poor light.
- At the moment, the frame rates are not particularly promising.

VII. CONCLUSION

Ensuring public safety in smart cities requires intelligent, automated, and real-time threat detection systems capable of addressing security challenges such as firearm-related incidents. The proposed deep learning-based firearms monitoring system enhances surveillance by integrating CNNs, Vision Transformers (ViTs), and edge AI computing to enable accurate, real-time detection of firearms in urban environments. Unlike traditional surveillance methods that rely heavily on manual monitoring and rule-based detection, this system significantly improves detection accuracy, reduces response time, and enhances law enforcement efficiency through automated threat alerts. The results demonstrate that AI-powered security solutions can provide a more robust, scalable, and adaptive approach to urban safety, ensuring faster incident response and minimizing risks in high-population areas. Future advancements in multi-modal AI integration, reinforcement learning, and IoT-based smart city security frameworks will further refine firearms monitoring, making smart cities safer and more resilient against security threats.

FUTURE SCOPE

The proposed deep learning-based firearms monitoring system lays the foundation for advanced AI-driven security solutions in smart cities, with several avenues for future research

and development. Enhancing real-time threat detection through multi-modal AI models that combine audio analysis, thermal imaging, and behavioral recognition can further improve accuracy in low-visibility and high-crowd environments. Additionally, edge AI and federated learning can be leveraged to process data locally, reducing latency and enhancing privacy without relying on cloud-based infrastructure. Future advancements may also integrate autonomous drone surveillance for rapid threat assessment in large public areas. Moreover, the incorporation of blockchain technology can ensure tamper-proof video evidence storage, improving data security and reliability in forensic investigations. By continuously evolving with adaptive deep learning models, IoT-based smart city networks, and AI-driven predictive security, this system can contribute to building safer, more resilient urban environments with proactive crime prevention and real-time law enforcement support.

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/2017-08-9/hoddle-streetmassacre-30-years-on/8786766/
12. (Mar. 2019). Christchurch Massacre. Accessed: Feb. 6, 2023. [Online]. Available: https://www.bbc.com/news/world-asia-7578798
13. (Jan. 2017). Quebec City Shoot-Out. Accessed: Feb. 6, 2023. [Online]. Available: https://www.bbc.com/news/world-us-canada- 8793071
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-7468978/isurvived-the-israeli-airport-massacre