Advanced Baggage Security Screening Framework: Integration Of Yolov9 For Contraband Detection And Threat Mitigation

Rohan Reddy B
Department of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham Chennai-601103, India
rohanbadugula@gmail.com

SP. Chokkalingam
Department of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham Chennai-601103, India
chomas75@gmail.com

***Abstract*—In bustling urban areas, especially at public transportation hubs, the growing congestion highlights the urgent need to tackle the security concerns in contraband detection. This problem not only causes delays but also presents a significant risk to public safety. Thus, there is a pressing demand for swift, automated, and precise detection of prohibited items in X-ray scans. This study introduces a one-shot detection system utilizing the latest YOLOv9 architecture, which demonstrates exceptional performance in detecting prohibited objects within baggage. Through rigorous experimentation on the SIXray dataset, our model achieves a noteworthy mean Average Precision (mAP) score of 0.956, indicating its remarkable efficacy. Our findings underscore YOLOv9's potential to significantly enhance both the accuracy and speed of baggage inspection processes, thereby alleviating delays and congestion in transportation hubs. Furthermore, comparative analyses against YOLO-NAS and YOLOv8 reveal YOLOv9's superiority, surpassing YOLO-NAS by 3% and YOLOv8 by 14%. This research highlights the transformative potential of YOLOv9 in baggage inspection, marking a significant advancement in automated detection systems within transportation security.**

***Index Terms*—Contraband detection, X-Ray security inspection, YOLOv9,YOLO-NAS, YOLOv8, Prohibited objects Classification.**

I. Introduction

With the swift evolution of the contemporary world and extensive public transportation usage, X-ray baggage inspection has become a vital technical process in terminal operations, ensuring the safety and security of passengers. In today's scenario, human involvement is crucial in X-ray image inspection to prevent people from carrying harmful items onto public vehicles [1]. However, prolonged periods of concentrated work during inspections can result in a decline in performance over time, posing potential dangers to public safety [2]. In order to streamline the security process and alleviate the burden on security personnel, automatic detection of prohibited objects has emerged as a focal point [3]. This system is designed to precisely identify hazard objects and efficiently handle the influx of passengers. Traditionally, a variety of machine learning and deep learning algorithms, including Bag-of-Words (BOW), Support Vector Machines (SVM) [4], and Convolutional Neural Networks (CNNs) [5], were employed for accurate detection of contraband objects within transportation hubs. However, these approaches exhibited poor performance when faced with certain hazardous items during testing.

In the purview of this research, we harness the latest iteration of YOLO, designated as You Only Look Once – Version 9 (YOLOv9) [6], to augment the detection and identification of contraband and harmful objects within baggage. This cutting-edge technology is thoughtfully designed to confront the specific challenges posed by X-ray object detection, notably the enhancement of object recognition and adaptability to varying imaging conditions. Moreover, in order to provide a comprehensive evaluation of our approach, we meticulously conduct a comparative analysis of YOLOv9 against YOLO-NAS and YOLOv8 (version 8) [7], the predecessors in the YOLO family. This comparative study enables us to discern the enhancements and advancements achieved by YOLOv9 in the domain of X-ray baggage security screening. This research endeavors to make a significant contribution to the ongoing efforts aimed at augmenting the safety and security of public transportation industry, through the judicious utilization of state-of-the-art deep learning technologies.

II. Related Works

In the modern era, significant attention is placed on the meticulous implementation of X-ray technology for identifying prohibited items [8], thereby ensuring the safety of individuals across various public domains. Numerous studies have been conducted in baggage inspection, transitioning from traditional machine learning approaches to advanced object detection algorithms, aiming for enhanced accuracy in identifying potential threats.

Initially, traditional machine learning models on detecting prohibited objects in X-ray images relied on one-view dependency methods. Turcsany et al. (2013) [4] developed a Bag-of-Words (BoW) approach utilizing the Speeded-Up Robust Features detector and descriptor alongside a Support Vector Machine classifier to detect firearms in X-ray images. Riffo et al. (2015) [9] employed adapted implicit shape models and a visual vocabulary derived from a training dataset of representative X-ray images to automatically detect threat objects in single views of grayscale X-ray images acquired using a single energy acquisition system. Subsequently, advancements in multi-view detection techniques emerged to enhance object detection performance by addressing the limitations of one perspective imaging. Franzel et al. (2012) [10] brought forward a multi perspective detection method that incorporates single-view detections from multiple angles. This method involves analyzing object variations in X-ray images and adapting standard detection methods to dual-energy data.

Later, as a response to the constraints of traditional models, deep learning methods emerged as a promising solution. Akcay et al. (2016) [5] employed pre-trained CNNs, such as AlexNet and GoogLeNet, for image classification in X-ray baggage security screening, specifically targeting handgun detection. Their study revealed that GoogLeNet exhibited robust performance, even for similar classes, ultimately achieving superior mean Average Precision. Mery et al. (2016) [11]experimented with X-ray luggage classification on the GDXray dataset. They compared classic approaches, Bag-of-Words (BoWs), sparse representations, codebooks, and deep features. Additionally, they introduced two new methods by combining Bag-of-Words (BoW) with sparse K-nearest neighbors (KNN). Evaluation showed that techniques utilizing visual vocabularies and deep features achieved high recognition rates. Wu et al. (2024) [12] used the ESLA hybrid Self-Supervised Learning strategy, combining Contrastive Learning and Masked Image Model, for feature extraction. They also employed the Head-Tail Feature Pyramid (HTFP) to generate multi-level feature maps from the output of the last stage of plain Vision Transformers (ViT) for X-ray prohibited objects detection. Mohamed et al. (2023) [13] integrated Convolutional Neural Networks (CNNs) for extracting features from X-ray images and Gated Recurrent Units (GRUs) for analyzing sequences of X-ray scans, resulting in improved accuracy in threat detection.

In recent years, there has been rapid development in object detection algorithms especially for the X-ray contraband detection. These algorithms have emerged as crucial tools in elevating detection accuracy and efficiency in comparison with traditional methods. Bhowmik et al. (2019) [14] employed a Faster R-CNN and ResNet101 architecture to detect prohibited items in X-ray security imagery. Their comparison between real X-ray training imagery and synthetically composed imagery revealed that real X-ray imagery demonstrated superior performance. Wei et al. (2021) [15] employed YOLOv3 integrated with the Spatial Pyramid Pooling (SPP) model for feature extraction to detect prohibited items. Their approach exhibited superior performance compared to other one-stage object detection algorithms. Wang et al. (2022) [3] improved YOLOv5 for X-ray luggage image security detection by incorporating a transformer in the last convolution module for better feature extraction, introducing a global attention mechanism to handle complex backgrounds, and employing an adaptive spatial feature fusion algorithm for precise predictions.

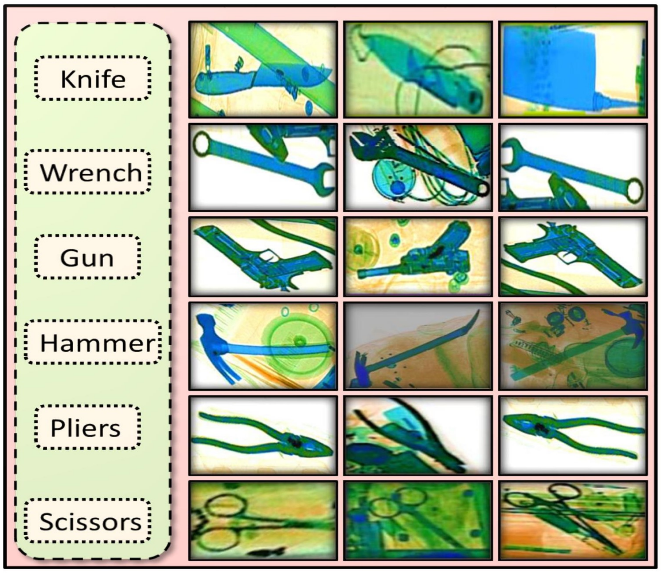


Fig. 1: Sample Images from Dataset Organized by Classes

Li et al. (2023) [1] introduced an enhanced version of the YOLOv7 algorithm for X-ray prohibited items detection. Their approach includes the integration of Bottleneck Transformers and Hydra Attention mechanisms to improve accuracy in crowded scenarios while optimizing computational efficiency. Furthermore, they introduced a convolution module within Hydra Attention to boost the representation of local details. Luo et al. (2023) [16] proposed an improved YOLOv8-based detection algorithm to address the shape and scale variability and overlap of prohibited items in X-ray images. They achieved this by incorporating ODConv into the backbone and DCNv3 into the Bottleneck of the C2f module, thereby improving feature extraction and adapting to object deformations.

III. Data Description

Dataset selection proves to be paramount in object detection tasks as it significantly influences accuracy of models. In the current research, we have employed the Security Inspection X-ray (SIXray) dataset [17] that consists of six most prevalent contraband items typically encountered in baggage screening scenarios. The entire dataset comprises 8,929 X-ray images containing the prohibited objects and its respective annotation files. Few samples of each class from the dataset are illustrated in the Figure 1. These X-ray images, sourced from an assortment of subway and railway stations, encapsulate contraband items spanning six distinct categories: knives, wrenches, guns, hammers, pliers, and scissors. Following their screening by inspection machines, these objects are visually distinguished and captured in JPEG format for subsequent analysis. The 8,929 images are distributed among the classes as shown in the Figure 3. The hammer class was excluded during the experiment due to the limited availability of samples. To facilitate robust evaluation, we adopted a conventional

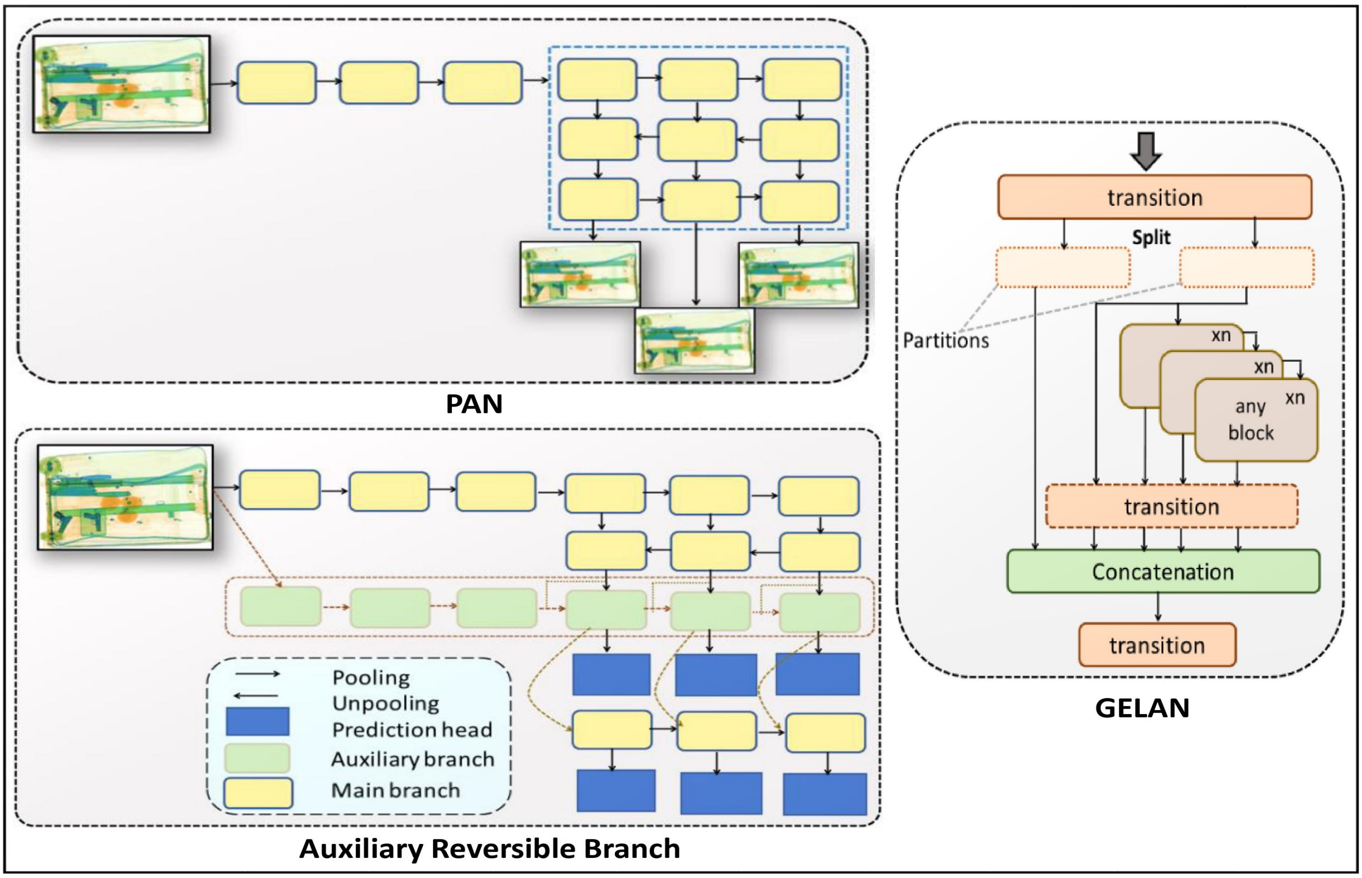


Fig. 2: Visual Depiction Illustrating the Structural Framework of YOLOv9

data partitioning strategy, allocating a 70-10-20 split ratio for training, validation, and test sets, respectively.

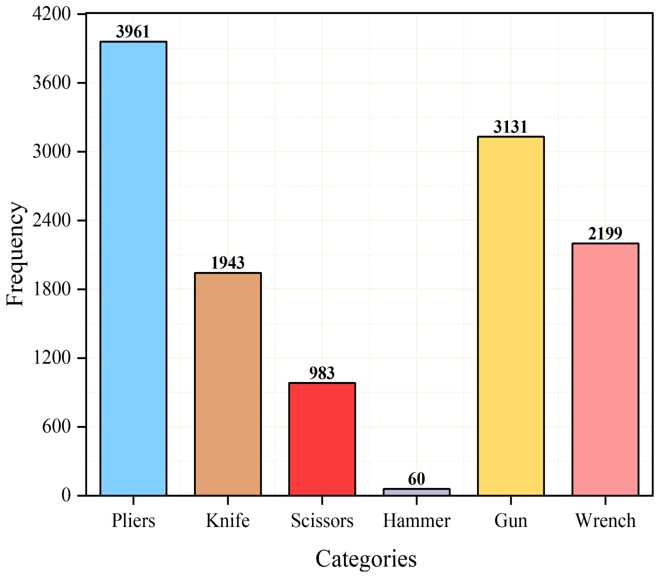


Fig. 3: Statistical Distribution of Class Instances Represented in the Dataset

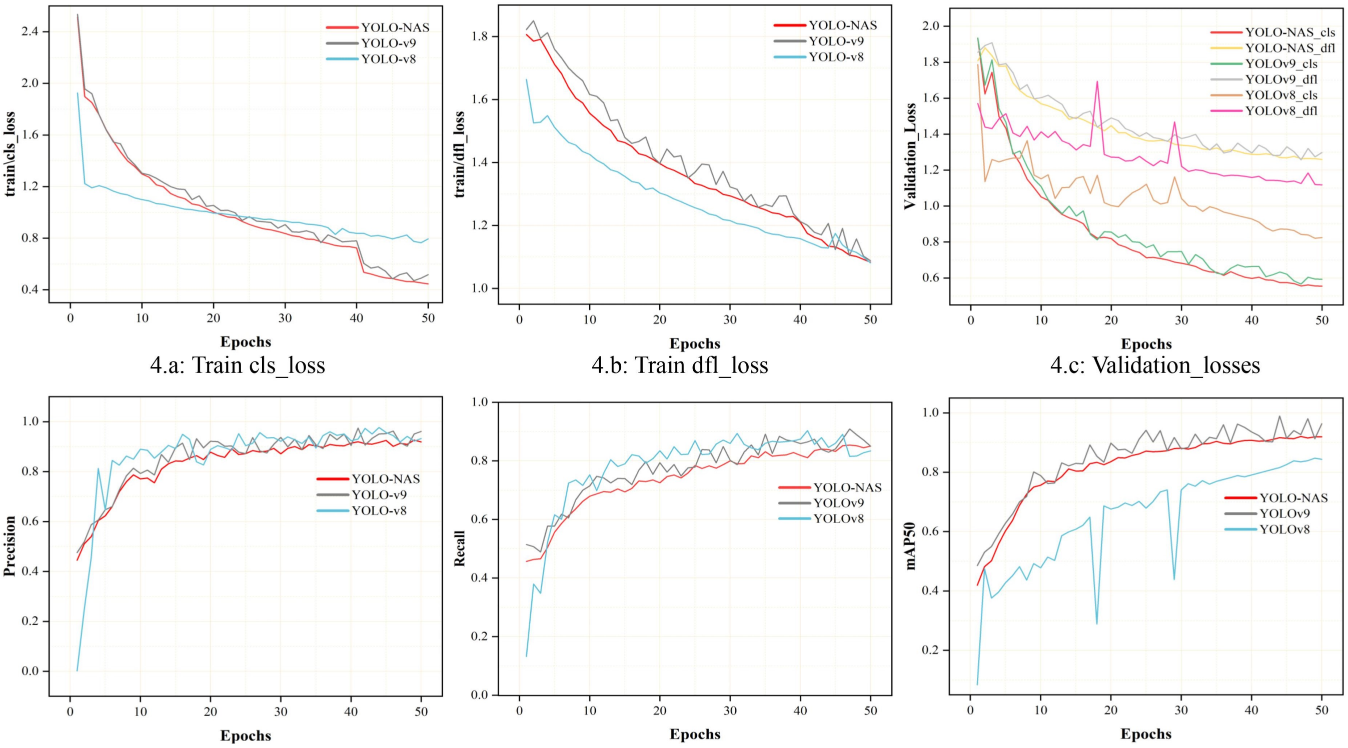
IV. Methodology

In this research, we employed YOLOv9, the most recent iteration within the YOLO series. This architecture builds upon the foundations laid by its predecessors, YOLOv8 and YOLOv7, while introducing innovative features. Previous versions of YOLO suffered from issues such as hindered gradient flow and information bottlenecks, resulting in prolonged training times and limited control over desired outputs. To address this, YOLOv9 incorporates novel modules such as Programmable Gradient Information (PGI) and the Generalized ELAN (GELAN) architecture.

The principle of information bottleneck [18] posits that as data propagates through successive layers of a neural network, there exists a risk of information loss. This phenomenon can impede the model's capacity to make precise predictions. Essentially, in the context of a simple neural network, each layer's operation on the input data may inadvertently result in a reduction of relevant information, thereby compromising the network's ability to accurately discern patterns and make reliable predictions.

$$Info\left(X,X\right) \geq Info\left(X,k_{\theta}\left(X\right)\right) \geq Info\left(X,l_{\phi}\left(k_{\theta}\left(X\right)\right)\right) \quad (1)$$

Here, "Info" represents the mutual information function between two parameters, while "k" and "l" denote transforma-



4.a: Train cls_loss

4.b: Train dfl_loss

4.c: Validation_losses

4.d: Precision

4.e: Recall

4.f: mAP50

Fig. 4: Comparing Training and Validation Metrics among YOLOv9, YOLO-NAS and YOLOv8 Models.

tion functions accompanied by their respective parameters in subsequent deeper layers. As the input data ”X” traverses through these layers governed by the specified functions, a discernible loss of information occurs. From the Equation 1 we can derive that the initial mutual information is always greater than that of the subsequent layer following transformation and so on for its further layers. This observation highlights the progressive mutual information reduction in the neural network's architecture as data progresses deeper into it. In addressing this, YOLO in its version nine has implemented reversible networks. This structure empowers the operational units within the network to revert the data back to its original form, thereby facilitating the retention of crucial information.

These reversible functions ensures the preservation of reliable gradients throughout the network's training process and serve as a effective measure against information loss.These networks have spurred the development of an auxiliary supervision framework known as Programmable Gradient Information (PGI). PGI comprises of three modules, the prime branch, auxiliary reversible branch, and multilevel auxiliary information, as depicted in Figure 2. The auxiliary reversible branch is specifically engineered to combat the network bottleneck inherent in deepening neural architectures. By generating gradients and updating network parameters during training, this branch plays a pivotal role in alleviating the information bottleneck's adverse effects. Within this framework, information mapping from data to targets is facilitated by the incorporation of reversible architecture. This integration serves to diminish the likelihood of spurious correlations and provides invaluable guidance to the loss function, thereby enhancing model performance. Notably, the auxiliary reversible branch, illustrated in Figure 2, serves as a critical mechanism for retaining information lost due to the progression of information bottleneck within the network. Furthermore, the network also generates a prediction head from the auxiliary branch. The gradient information derived from this auxiliary branch serves as a driving force for parameter learning, enabling the network to rectify outcomes for object detection tasks with heightened accuracy and efficacy.

The concept of Multilevel Auxiliary Information involves the utilization of distinct feature pyramids to predict objects of varying sizes. Each pyramid focuses on leveraging information pertinent to objects within its respective scale. This strategy aims to enhance object detection across a wide spectrum of sizes. The core to this concept is the integration of an intermediary network between the hierarchical layers of the auxiliary supervision's feature pyramid and the main branch. This network serves as a conduit for amalgamating the returned gradients from different prediction heads, as depicted in Figure 2. Subsequently, the integrated gradients encompassing all requisite outcomes are aggregated and conveyed to the main branch. This integration of multilevel auxiliary information offers a holistic approach to object detection, effectively leveraging hierarchical feature representations to capture diverse object scales.

At the core of YOLOv9 lies the Generalized Efficient Layer

Aggregation Network (GELAN), which represents a fusion of two pivotal components: the Cross Stage Partial Network (CSPNET) [19] and ELAN. The architecture of CSPNET involves splitting the input feature map into two distinct parts. One segment traverse through a dense block comprising multiple convolutional and pooling modules, while the other segment remains unaltered until the concatenation of these modules at the end. This split-and-merge approach facilitates an optimal gradient flow across the network, thereby enhancing its learning capacity. Adding upon this, ELAN establishes shortcuts to expedite the gradient flow from input layers to deeper layers within the network. This enables the model to effectively process multiple levels of input data concurrently, thus empowering object detection tasks to scrutinize both small and large objects simultaneously. Once individual detections are performed across varying granularities, the model consolidates these results, generating a higher resolution representation that encapsulates the pertinent objects of interest. Through the integration of CSPNET and ELAN, YOLOv9 extends its capabilities to facilitate robust information flow across the network. This synergistic combination equips the model with the ability to comprehend and internalize complex patterns.

V. Experiment Analysis

In this research, we have conducted a comprehensive experimental evaluation aimed at assessing the performance of YOLOv9 in comparison to its predecessors, YOLOv8 and YOLO-NAS, under controlled conditions. As the Dataset has almost 9,000 images, we have utilized a NVIDIA TESLA P100 GPU for the entire experimentation of three models across all phases including training, testing, and hyperparameter tuning. Considering the variability in model sizes within the YOLO family of nano, small, and large versions, we opted for the large variant for all three models in our analysis. Training was executed over 50 epochs, with YOLOv9 completing in approximately 5.2 hours, YOLO-NAS in 13.4 hours, and YOLOv8 in 4.5 hours. Notably, YOLO-NAS, being constructed from scratch without reliance on pre-established libraries, exhibited the lengthiest training duration. The mean Average Precision at 50 IoU threshold (mAP50) serves as the primary performance metric for tracking the model's advancement throughout the training process, offering a holistic assessment of its object detection capabilities. Figure 4 showcases a collection of six subplots, each delineating the evolution of training and validation losses alongside evaluation metrics such as Precision, Recall, and mAP50. From the plots it is evident that YOLOv9 proves to show a superior performance compared to the other models, in almost all the aspects within our experimental setup.

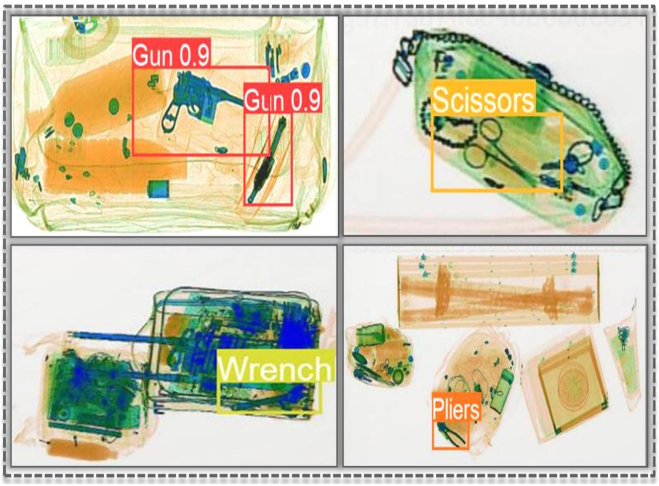


Fig. 5: Sample Predictions on few Test Images

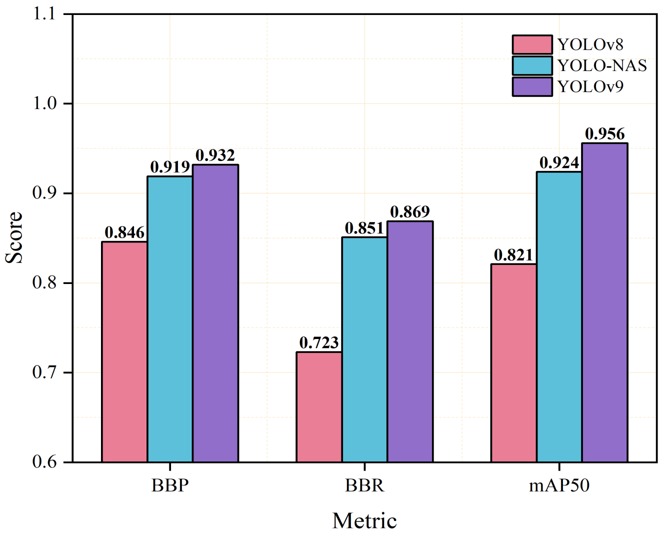


Fig. 6: Performance metric comprasion of YOLOv9, YOLO-NAS and YOLOv8 Models on test set

VI. Result and Discussion

During the testing phase, we conducted evaluations on the test dataset to assess the effectiveness of YOLOv9, alongside comparative analyses with YOLOv8 and YOLO-NAS. Sample detections from a subset of test images are depicted in Figure 5, providing visual insights into the detection outcomes. The Quantitative assessments of the models' detection capabilities were obtained through Bounding Box Precision (BP), Bounding Box Recall (BR), and Mean Average Precision (mAP). BP and BR were computed as follows,

$$BBP = \frac{\text{CP}}{\text{CP + IP}} \tag{2}$$

$$BBR = \frac{\text{CP}}{\text{CP + MB}} \tag{3}$$

Here, CP represents the count of bounding boxes predicted correctly, IP indicates the number of bounding boxes predicted incorrectly, and MB denotes the missed bounding boxes. The mAP serves as a comprehensive metric for assessing prediction accuracy across various object categories. It calculates the average precision, providing insight into a model's ability to detect objects. In particular, mAP50 emphasizes a critical facet of detection precision by evaluating predictions where the Intersection over Union (IoU) between predicted and ground-truth bounding boxes surpasses a predetermined threshold of 0.5.

Figure 6 provides a comparative illustration of the key performance metrics across the three models. YOLOv9 notably outperformed its counterparts, achieving a remarkable mAP50 score of 95.6%, followed by YOLO-NAS with 92.4%, and YOLOv8 with 82.1%. Additionally, YOLOv9 exhibited superior Bounding Box Precision and Bounding Box Recall values at 93.2% and 83.9%, respectively. Moreover, individual class metrics assume critical importance in contraband detection. These are meticulously detailed in Table I revealing nuanced performance insights. A primary examination indicates that the "gun" class attains the highest mAP50 score of 98.3% and "wrench" registers the lowest at 94%. Such findings underscore the effectiveness of the YOLOv9 model in contraband detection tasks, suggesting potential avenues for further refinement and deployment within security screening frameworks.

TABLE I: Performance Metrics of Detected Threat objects Classes

Class	**BBP**	**BBR**	**mAP50**
Gun	97.4%	96.2%	98.3%
Knife	92.6%	84.7%	96.2%
Pliers	94.2%	86.8%	97.4%
Scissors	93.3%	86.2%	95.1%
Wrench	88.1%	80.3%	94.0%

VII. Conclusion

In conclusion, the detection of contraband holds significant importance, particularly in densely populated areas such as transportation hubs. This research leverages the latest iteration of the YOLO family, YOLOv9, to assess its performance in comparison with its predecessors, YOLOv8 and YOLOv9. Through experimentation using the SIXray Dataset, our findings reveal a mAP50 score of 95.6%, surpassing the performance of YOLO-Nas and YOLOv8. YOLOv9 exhibits notable efficacy in both speed and accuracy. Moving forward, it is recommended that future research extends beyond simulation to real-world testing in operational centers. Additionally, the integration of the new 3D representation of baggage scanning objects in lidar format should be explored, offering potential advantages in scenarios where objects are stacked atop one another.

References

[1] S. Li and Y. Musha, "Detection of x-ray prohibited items based on improved yolov7," pp. 1874–1877, 2023.

[2] R. K. Putra and N. P. Utama, "Enhancing x-ray baggage inspection through similarity-based prohibited object identification," pp. 16–21, 2023.

[3] Z. Wang, H. Zhang, Z. Lin, X. Tan, and B. Zhou, "Prohibited items detection in baggage security based on improved yolov5," pp. 20–25, 2022.

[4] D. Turcsany, A. Mouton, and T. P. Breckon, "Improving feature-based object recognition for x-ray baggage security screening using primed visualwords," pp. 1140–1145, 2013.

[5] S. Akcay, M. Kundegorski, M. Devereux, and T. Breckon, "Transfer learning using convolutional neural networks for object classification within x-ray baggage security imagery," pp. 1057–1061, 09 2016.

[6] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "Yolov9: Learning what you want to learn using programmable gradient information," 2024.

[7] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas," *Machine Learning and Knowledge Extraction*, vol. 5, no. 4, p. 1680–1716, Nov. 2023.

[8] D. Mery, D. Saavedra, and M. Prasad, "X-ray baggage inspection with computer vision: A survey," *IEEE Access*, vol. 8, pp. 145 620–145 633, 2020.

[9] V. Riffo and D. Mery, "Automated detection of threat objects using adapted implicit shape model," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 46, no. 4, pp. 472–482, 2016.

[10] T. Franzel, U. Schmidt, and S. Roth, "Object detection in multi-view x-ray images," pp. 144–154, 2012.

[11] D. Mery, E. Svec, M. Arias, V. Riffo, J. Saavedra, and S. Banerjee, "Modern computer vision techniques for x-ray testing in baggage inspection," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. PP, pp. 1–11, 12 2016.

[12] J. Wu and X. Xu, "Eslaxdet: A new x-ray baggage security detection framework based on self-supervised vision transformers," *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107440, 2024.

[13] A. S. Mohamed, A. A. Sewisy, K. F. Hussain, and A. I. Taloba, "Strengthening the security of the kuwait international airport by detecting threats in x-ray images," *Appl. Math*, vol. 18, no. 1, pp. 23–32, 2024.

[14] N. Bhowmik, Q. Wang, Y. F. A. Gaus, M. Szarek, and T. P. Breckon, "The good, the bad and the ugly: Evaluating convolutional neural networks for prohibited item detection using real and synthetically composited x-ray imagery," *arXiv preprint arXiv:1909.11508*, 2019.

[15] Y. Wei, C. Dai, M. Chen, Z. Xu, Y. Liu, J. Fan, F. Ren, and Z. Liu, "Prohibited items detection in x-ray images in yolo network," pp. 1–6, 2021.

[16] Y. Luo and C. Liu, "Improved yolov8 detection algorithm for detecting contraband in x-ray security inspection image," pp. 1193–1197, 2023.

[17] C. Miao, L. Xie, F. Wan, C. Su, H. Liu, J. Jiao, and Q. Ye, "Sixray : A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images," 2019.

[18] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," 2015.

[19] C.-Y. Wang, H.-Y. M. Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen, and J.-W. Hsieh, "Cspnet: A new backbone that can enhance learning capability of cnn," 2019.