

Multi-View Detection of Prohibited Items in X-ray Images with Class Imbalance Handling

Soon-Hyuck Lee

*Luddy School of Informatics, Computing, and Engineering
Indiana University
Bloomington, United States
sl200@iu.edu*

Owen Randolph

*Luddy School of Informatics, Computing, and Engineering
Indiana University
Bloomington, United States
oyrandol@iu.edu*

Marcos Fernandez

*Luddy School of Informatics, Computing, and Engineering
Indiana University
Bloomington, United States
fernmarc@iu.edu*

Pratham Dedhiya

*Luddy School of Informatics, Computing, and Engineering
Indiana University
Bloomington, United States
pdedhiya@iu.edu*

Abstract—Airports and high-security checkpoints face escalating pressure to improve the accuracy, speed, and consistency of baggage screening. Current inspection systems rely heavily on human operators who must interpret complex transmission X-ray imagery under severe time constraints. This dependence introduces significant risks, including operator fatigue and a heightened susceptibility to human error when identifying small, occluded, or visually ambiguous threat items. To address these limitations, we propose a multi-stage deep learning framework for automated threat detection using the Dual-View X-ray (DvXray) dataset [14]. Our approach integrates a ResNet-18 multi-label classifier for global threat identification, achieving a classification mAP of 0.8122 on the DvXray dataset, and compares single-stage (YOLOv8-Nano) and two-stage (Faster R-CNN) architectures for precise localization. By implementing an automated multi-view late-fusion strategy to leverage orthogonal X-ray projections, we demonstrated a significant performance increase, elevating system recall from a baseline of 84.81% to 91.01%. Furthermore, we utilize Gradient-weighted Class Activation Mapping (Grad-CAM) to validate that the model focuses on high-density edges and critical geometric features. This study provides a scalable, explainable foundation for next-generation screening platforms designed to mitigate cognitive burden and enhance detection reliability in cluttered environments.

Index Terms—Deep learning, dual-view fusion, Grad-CAM, multi-label classification, object detection, security inspection, X-ray baggage screening, YOLO.

I. INTRODUCTION

The use of deep learning in X-ray security screening has increased significantly in recent years due to major advances in convolutional neural networks (CNNs). Traditional image-processing methods struggle with X-ray images because objects often overlap, vary in shape, and appear with low contrast. Deep neural networks can learn hierarchical visual patterns such as edges, textures, materials, and object shapes, making them well suited for analyzing complex baggage imagery containing multiple overlapping items. CNNs have become the standard method for image-based detection and classification because they learn features directly from raw

pixel data. The early layers capture simple patterns like edges and corners, while deeper layers learn more abstract structures such as shapes, semantic components, or entire objects [1]. This multi-level feature representation is essential for X-ray imaging, where important visual cues may be subtle, partially occluded, or distorted by overlapping items [1], [2].

Earlier CNN architectures such as VGG achieved strong performance but became difficult to train as they grew deeper [10]. Stacking additional layers often made optimization unstable and vulnerable to vanishing gradients [3], [10], preventing deeper networks from learning meaningful features. These limitations slowed progress in tasks requiring high-level abstraction, including the detection of small concealed objects in X-ray images.

Recent work in automated X-ray analysis has addressed challenges such as limited labeled data, strong object overlap, and variability in real baggage imagery. Kaminetzky and Mery [15] presented one of the most significant advances in simulation-based data augmentation for X-ray security screening. Their method uses 3D threat-object models and superimposes simulated X-ray projections onto real baggage images, incorporating diffusion-based distortions and semantic segmentation for realism. Using this synthetic dataset in combination with real SIXray images, they trained hundreds of YOLOv5 models and demonstrated substantial performance gains attributable solely to simulated data. Remarkably, training exclusively on 16,000 simulated grayscale wrench images produced a real-image mAP of 72.7%, and augmenting just 50 real handgun samples with 16,000 simulated examples increased detection accuracy from approximately 80% to over 90%. These results highlight the potential of simulation pipelines to significantly strengthen deep-learning-based X-ray detectors while reducing the need for large-scale manual annotation.

II. RELATED WORK

The SIXray benchmark itself (Miao et al.) remains one of the largest publicly available datasets for prohibited item detection [12] and is widely used for evaluating modern X-ray recognition systems. Containing over one million X-ray images with heavily overlapping objects, SIXray introduces realistic challenges such as clutter, occlusion, and scale variation. Models evaluated on SIXray must be capable of distinguishing extremely small and partially hidden objects in complex baggage scenes, making it a critical resource for training and benchmarking deep neural networks in security applications.

Complementary to synthetic data generation and large-scale benchmarking, prior work has also looked into the task of overcoming data scarcity through transfer learning. Akcay et al. [11] explore the use of pre-trained convolutional neural networks for X-ray object classification, noting that fully supervised end-to-end CNN training typically requires large, diverse datasets that are not readily available in the security domain. Their transfer learning framework adapts existing CNN models—originally trained on natural images—to the X-ray domain through secondary fine-tuning. Applied to handgun detection, their approach achieved a detection accuracy of 98.92% [11], outperforming earlier methods and demonstrating that domain-adapted transfer learning can effectively compensate for limited X-ray data.

Collectively, these works establish that progress in X-ray computer vision relies on three complementary strategies [10]–[12]: 1) generating realistic simulated X-ray imagery to expand training sets, 2) leveraging large-scale benchmarks such as SIXray to develop models robust to occlusion and clutter, and 3) applying transfer learning to mitigate domain-specific data scarcity. These innovations form the foundation of current state-of-the-art research in automated baggage inspection and are directly relevant to the development of robust, real-time X-ray threat detection systems.

III. DATA PREPARATION

To support diverse computer vision tasks—including object detection (YOLO, COCO) and multi-label classification—we implemented a comprehensive data processing pipeline for the DvXray dataset [14]. The pipeline, designed to automate the conversion of raw X-ray scans into structured formats, ensures data integrity and compatibility across different machine learning frameworks.

A. Data Source and Ingestion

The pipeline ingests the raw DvXray dataset, which consists of dual-view imagery and associated metadata. Each sample includes two orthogonal X-ray projections: the Optical Level (_OL) and the Side View (_SD). The dataset comprises both “Positive” samples, containing prohibited items, and “Negative” samples, representing clean or empty bags. Ground truth annotations are parsed from JSON files containing bounding boxes and class labels for each security scan.

B. Processing Workflow

The data transformation process executes sequentially to ensure robust dataset creation: Loading and Splitting: The system first loads raw positive and negative samples, creating a stratified split for training and validation to ensure reliable model evaluation.

Multi-Format Generation: The pipeline simultaneously exports data for three distinct architectures: YOLO: Generates a standard directory structure (images/ and labels/) with bounding boxes converted into normalized YOLO format. COCO: Consolidates annotations into a single instances_train.json file compliant with the COCO detection format. Classification: Extracts class labels from spatial annotations to generate 15-dimensional “multi-hot” vectors (saved as .npy files). These vectors indicate the global presence or absence of prohibited items in an image, facilitating multi-label learning.

C. Data Management

The pipeline explicitly manages dual-view correspondence by processing both OL and SD images for every sample. It also incorporates logic to handle edge cases, such as “difficult” samples or missing bounding boxes, ensuring the training set remains robust against annotation errors. The final output is organized into a structured data/processed/directory containing dedicated subfolders for YOLO, COCO, and Classification tasks.

IV. METHODOLOGY

To address the challenges of baggage screening, we propose a multi-faceted approach combining multi-label classification for global threat detection, object detection for precise localization, and explainable AI to validate model focus.

A. Multi-Label Classification (ResNet-18)

We developed a deep learning subsystem to perform multi-label image classification, serving as a high-recall filter to detect the presence of 15 distinct categories of prohibited items.

- Architecture: We utilized a ResNet-18 backbone pre-trained on ImageNet to leverage transfer learning [3]. To adapt the architecture for security screening, the final fully connected (FC) layer was replaced with a linear layer possessing 15 output neurons, corresponding to the specific threat classes in the DvXray dataset [14].
- Loss Function: We employed BCEWithLogitsLoss (Binary Cross-Entropy with Logits), which is well-suited for multi-label tasks where a single bag may contain multiple overlapping threats simultaneously.

B. Object Detection Architectures

To localize specific threats within the X-ray imagery, we implemented and compared two distinct detection frameworks:

- 1) Single-Stage Detection (YOLOv8)

We selected YOLOv8-Nano for its balance of speed and accuracy [6], aligning with the requirement for real-time screening.

- Configuration: The model was trained with an input resolution of 640 x 640 pixels. We employed Mosaic Augmentation during training, which combines four training images into one [13]. This forces the model to learn fractional object features, effectively countering the occlusion and overlap inherent in X-ray baggage scans.
- Dual-View Fusion: To exploit the multi-view nature of the dataset, we implemented a late-fusion logic [8]. Inferences are run independently on the Optical Level (OL) and Side View (SD) images, and the results are merged via a logical union. If an object is detected in either view, it is flagged, significantly improving recall for occluded items.

2) Two-Stage Detection (Faster R-CNN)

For comparison, we implemented a Faster R-CNN architecture [7], which typically offers higher localization precision at the cost of inference speed.

- Backbone: We employed a ResNet-50 backbone with a Feature Pyramid Network (FPN). This structure extracts multi-scale feature maps, capturing both low-level edges and high-level semantic information essential for recognizing objects of varying sizes.
- Region Proposal: The Region Proposal Network (RPN) scans feature maps to generate candidate object regions, which are then refined by a FastR-CNNPredictor modified for our 15-class problem.

C. Explainability (Grad-CAM)

To ensure the classification model relies on relevant visual features rather than background artifacts, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) [9]. This technique computes the gradients of a target class score with respect to the feature maps of the final convolutional layer. These gradients are pooled to generate a coarse localization map, highlighting the image regions that most influenced the model’s prediction.

V. EXPERIMENTAL RESULTS

The development and execution of the multi-stage detection framework were conducted using high-performance computing (HPC) resources at Indiana University. Model training and large-scale evaluations were performed on the BigRed 200 supercomputer and dedicated high-tier GPU servers at the Luddy School of Informatics, Computing, and Engineering.

A. Training Configuration

Classification (ResNet): Images were resized to 256 x 256 pixels and normalized (mean=0.5, std=0.5). The model was optimized using AdamW with a learning rate of 1e-4 for 20 epochs.

Detection (YOLO): The YOLOv8-Nano model was trained for 50 epochs. An attempt to extend training to 100 epochs resulted in overfitting, confirming that convergence was achieved early.

B. Classification Performance

The ResNet-18 classifier demonstrated strong potential as a primary filter. By Epoch 20, the model achieved a Mean Average Precision (mAP) of 0.8122 and a Micro F1 Score of 0.7442. The convergence behavior of the classifier over training epochs is illustrated in Fig. 1.

Per-Class Analysis: Performance was correlated with object density. Large, rigid items like Bat (AP 1.00), Hammer (AP 0.99), and Gun (AP 0.95) were detected with near-perfect reliability [1]. Conversely, thin or small items such as Razor Blade (AP 0.52) and Lighter (AP 0.62) proved challenging due to limited visual features in transmission imagery, as shown in Table I.

TABLE I
AVERAGE PRECISION (AP) SCORES FOR PROHIBITED ITEM CATEGORIES

Class Name	AP Score	Class Name	AP Score
Bat	1.000	Saw Blade	0.799
Hammer	0.991	Screwdriver	0.794
Gun	0.953	Knife	0.663
Pressure Vessel	0.927	Lighter	0.616
Battery	0.895	Razor Blade	0.517

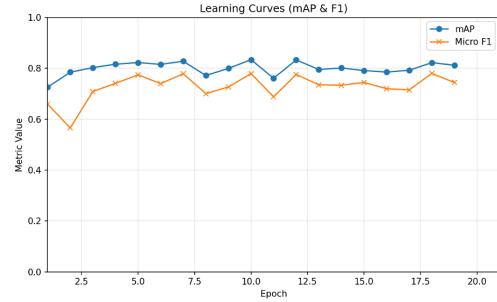


Fig. 1. Mean AP and F1 Score over training epochs.

C. Detection Performance

- 1) **YOLOv8 Results:** The model achieved a detection accuracy (mAP@0.50) of 84.81%. Implementing the multi-view fusion logic increased the system’s recall to 91.01%, demonstrating that leveraging orthogonal views effectively mitigates occlusion. The class-level confusion patterns are shown in Fig. 2, where thin metallic objects such as Razor Blade and Scissors emerge as the most frequently misclassified categories. Qualitative detection results produced by the YOLOv8-Nano model on the validation set are visualized in Fig. 3.
- 2) **Faster R-CNN Results:** While Faster R-CNN achieved high Intersection over Union (IoU) scores—indicating

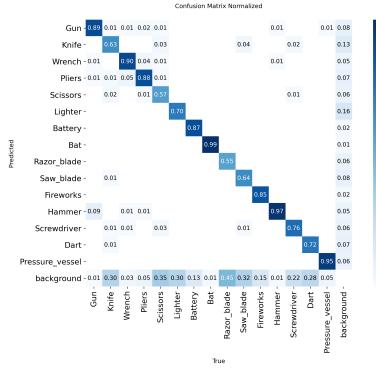


Fig. 2. Confusion matrix, showing that Razor Blade and Scissors are the most frequently misclassified classes.

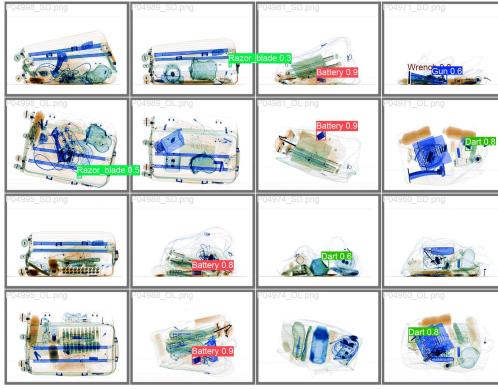


Fig. 3. A batch of validation predictions produced by the YOLOv8-Nano model.

precise bounding boxes—it struggled with correct classification (e.g., misclassifying a Battery as a Bat). This suggests the model learned spatial localization well but requires more data to resolve class ambiguity compared to the YOLO approach. An illustrative example of this behavior is shown in Fig. 4. Quantitative detection and localization metrics for the Faster R-CNN model are reported in Table II, highlighting the contrast between strong localization accuracy and weak class discrimination.

TABLE II
FASTER R-CNN PERFORMANCE: MAP AND IOU METRICS

Metric	Value	Metric	Value
mAP@0.50 (VOC)	0.0287	Val Mean IoU	0.7597
mAP@0.75	0.0247	Val Median IoU	0.8843
mAP@0.50:0.95	0.0216	Val IoU ≥ 0.50	0.8556
Train Mean IoU	0.7553	Val IoU ≥ 0.30	0.8926
Train Median IoU	0.8918		
Train IoU ≥ 0.50	0.8516		
Train IoU ≥ 0.30	0.8668		

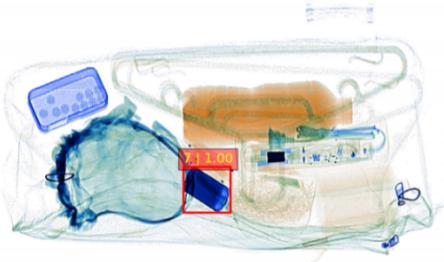


Fig. 4. An illustration of the model correctly detecting the object's location while incorrectly classifying it as a bat rather than a battery as Faster R-CNN.

D. Qualitative Analysis

Grad-CAM visualizations for the ResNet classifier are presented in Fig. 5, demonstrating that the model consistently attends to high-density edges and object cores rather than background regions. This observation confirms that the achieved mAP performance is driven by meaningful feature learning rather than dataset bias.

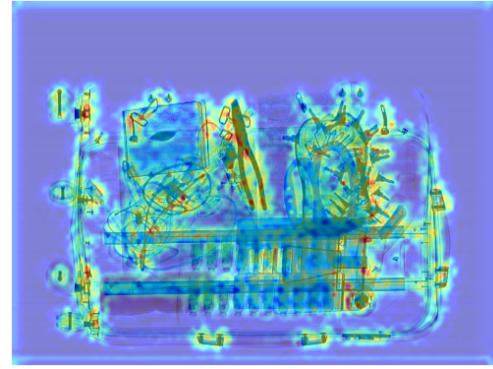


Fig. 5. Grad-CAM visualization highlighting salient regions used by the Custom ResNet classifier for threat prediction.

VI. DISCUSSION

The experimental results demonstrate the complementary strengths of single-stage (YOLOv8) and two-stage (Faster R-CNN) architectures, as well as the utility of multi-label classification for automated security screening.

Our analysis shows that YOLOv8-Nano, despite its lightweight architecture, achieved superior detection accuracy (84.81% mAP) compared to Faster R-CNN. This success is largely attributed to the use of Mosaic Augmentation [13], which artificially increases the complexity of training samples by mashing four images together, forcing the model to learn localized features even when objects are partially occluded. Furthermore, the implementation of Multi-View Late Fusion proved to be the most significant performance booster, raising the system's recall to approximately 91.01%. This confirms the operational hypothesis that utilizing orthogonal views (OL and SD) allows the system to resolve ambiguities [8] that occur when an item is "hidden" in a single 2D projection.

In contrast, Faster R-CNN exhibited high localization precision (high IoU) but struggled with class recognition, often

misclassifying objects with similar geometric profiles—such as mistaking a *Battery* for a *Bat*. This suggests that while the two-stage architecture’s Region Proposal Network (RPN) is structurally sound for finding “where” an object is, it requires a larger or more balanced dataset to accurately determine “what” the object is compared to the YOLO framework.

Finally, the Grad-CAM visualizations provided necessary validation for our ResNet-18 classification model [3]. By highlighting that the model focuses on the high-density edges and distinct silhouettes of prohibited items rather than background noise, we confirm that the classification subsystem can effectively serve as a primary “threat/no-threat” filter before more computationally expensive detection models are triggered.

VII. LIMITATIONS AND FUTURE WORK

Despite the high recall achieved through multi-view fusion, several limitations remain. The dataset size, while functional for a final project, led to early overfitting—specifically in YOLOv8 training beyond 50 epochs and in Faster R-CNN’s difficulty with fine-grained class separation. Furthermore, X-ray transmission imaging inherently suffers from feature fragmentation; small, thin metallic items like Razor Blades (AP 0.517) and Lighters (AP 0.616) remain significantly more difficult to detect than large, rigid objects like Guns or Bats.

To address these limitations, future research should explore:

- Enhanced Data Augmentation: Incorporating intensity shifts, synthetic noise, and random flips to simulate the varied orientations of baggage items in a real-world conveyor system. This approach identifies a clear path for future optimization through advanced data augmentation and architectural scaling.
- Learning Rate Scheduling: Applying cosine annealing or “ReduceLROnPlateau” strategies could refine model convergence in the final epochs.
- Alternative Backbones: Transitioning to heavier backbones like ResNet-101 or efficiency-focused models like MobileNet-FPN to further explore the accuracy-speed tradeoff.
- Operational Deployment: Exporting models to ONNX or TensorRT formats would enable real-time inference on GPU servers or web-based security monitoring applications.

VIII. CONCLUSION

This project successfully developed an end-to-end deep learning pipeline for the Dual-View X-ray (DvXray) baggage detection task. By integrating ResNet-18 for global classification, YOLOv8 and Faster R-CNN for localized object detection, and Grad-CAM for explainability, we created a robust framework for automated threat identification.

Our findings underscore the critical importance of multi-view reasoning in security contexts; the transition from single-view inference to a fused dual-view approach increased detection recall from 84% to over 91%. While localization remains precise across models, the disparity in per-class

performance—ranging from perfect retrieval for large tools to moderate success for small metallic items—identifies a clear path for future optimization through advanced data augmentation and architectural scaling. Ultimately, this multi-faceted approach provides a viable foundation for reducing human operator workload and increasing the consistency of airport security screenings.

REFERENCES

- [1] D. Mery, Computer Vision for X-Ray Testing: A Practical Guide. Springer, 2015.
- [2] D. Turcsany, A. Mouton, and T. P. Breckon, “Improving feature-based object recognition for X-ray baggage security screening,” AVSS, 2013.
- [3] K. He et al., “Deep residual learning for image recognition,” CVPR, 2016.
- [4] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for CNNs,” ICML, 2019.
- [5] J. Redmon et al., “You only look once: Unified, real-time object detection,” CVPR, 2016.
- [6] G. Jocher, “YOLOv8: Ultralytics,” 2023.
- [7] S. Ren et al., “Faster R-CNN: Towards real-time object detection with region proposal networks,” NeurIPS, 2015..
- [8] D. Mery and E. Svec, “Multi-view X-ray imaging for cargo and baggage inspection,” IEEE TIP, 2019.
- [9] R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” ICCV, 2017.
- [10] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” ICLR, 2015.
- [11] S. Akçay et al., “Transfer learning using CNNs for X-ray baggage imagery,” ICIP, 2018.
- [12] C. Miao et al., “SIXRay: A large-scale security inspection X-ray benchmark for prohibited item discovery in overlapping images,” CVPR, 2019.
- [13] A. Bochkovskiy et al., “YOLOv4: Optimal speed and accuracy of object detection,” arXiv, 2020.
- [14] B. Ma et al., “Toward Dual-View X-Ray Baggage Inspection: A Large-Scale Benchmark and Adaptive Hierarchical Cross Refinement for Prohibited Item Discovery,” IEEE TIFS, vol. 19, pp. 3866–3878, 2024.
- [15] A. Kaminetzky and D. Mery, “Improving Automated Baggage Inspection Using Simulated X-ray Images of 3D Models,” in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (CIARP 2023)*, vol. 13867, pp. 102–115, Springer, 2023.