DEVELOPMENT OF ROBUST DEEP LEARNING FRAMEWORKS FOR PROHIBITED ITEM DETECTION IN X-RAY SECURITY IMAGERY

A Thesis Submitted
In Partial Fulfillment of the Requirements
for the Degree of

M.TECH BY RESEARCH
in
INFORMATION TECHNOLOGY
with
specialization in
DATA SCIENCE AND ENGINEERING

by

Ranu Singh
(24/RIT/02)

Under the Supervision of
Prof. Dinesh Kumar Vishwakarma
&
Dr. Bindu Verma



Department of Information Technology
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India
May, 2026

ACKNOWLEDGEMENT

I express my sincere gratitude and respect to my supervisors, **Prof. Dinesh Kumar Vishwakarma** and **Dr. Bindu Verma**, Department of Information Technology, Delhi Technological University, for their continuous guidance, valuable suggestions, motivation, and constant support throughout my M.Tech research. Their knowledge, thoughtful discussions, and research-oriented approach provided inspiration at every stage of this thesis. I am deeply grateful for their support, encouragement, and trust, which helped me successfully complete this work.

I sincerely thank **Dr. Bindu Verma** for her constant guidance, patience, and encouragement throughout the course of this research. Her technical advice, feedback, and guidance were invaluable, and her constant assistance supported my academic and professional growth.

I also express my deep gratitude to **Prof. Dinesh Kumar Vishwakarma** for his outstanding guidance, visionary suggestions, and continuous encouragement during my M.Tech journey. His assistance and encouragement helped me identify the research direction and improve the quality of this work.

I extend special thanks to **Dr. Kavinder Singh** for his continuous support and guidance throughout the academic course. His encouragement and backend support provided motivation during difficult phases of my research and academic journey.

I am deeply thankful to my friends **Harsh Sinha** and **Ms. Bhavana Verma**, who were constant sources of support and motivation throughout this journey. Their cooperation, discussions, and companionship made this research journey much smoother and more memorable.

I sincerely thank all faculty members and staff of the Department of Information Technology, Delhi Technological University, for providing a motivating academic environment and necessary facilities for carrying out this research work.

Finally, I express my deepest gratitude to my parents and family members for their unconditional love, blessings, sacrifices, and continuous encouragement throughout my academic journey. Their faith has always been my greatest source of strength and inspiration.

Above all, I thank Almighty God for the strength, patience, and determination required to successfully complete this thesis work.

Ranu Singh



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Bawana Road-Delhi-42

DECLARATION

I, RANU SINGH, hereby certifies that the work being presented in this thesis entitled "DEVELOPMENT OF ROBUST DEEP LEARNING FRAMEWORKS FOR PROHIBITED ITEM DETECTION IN X-RAY SECURITY IMAGERY" submitted in partial fulfillment of the requirements for the award of the degree of Master of Technology in Department of Information Technology, Delhi Technological University, is an authentic record of the candidate's own work carried out during the period from August 2024 to May 2026 under the supervision of Prof. Dinesh Kumar Vishwakarma & Dr. Bindu Verma.

The matter presented in this thesis has not been submitted by the candidate for the award of any other degree of this or any other Institute.

Candidate's Signature

Signature of Supervisor(s)

Signature of External Examiner



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Bawana Road-Delhi-42

CERTIFICATE BY THE SUPERVISOR'S

Certified that Mr. **Ranu Singh** (**24/RIT/02**) has carried out the research work presented in this thesis entitled **"Development of Robust Deep Learning Frameworks for Prohibited Item Detection in X-Ray Security Imagery"** submitted for the award of **M.Tech by Research** in **Data Science and Engineering**, Department of Information Technology, Delhi Technological University.

This thesis embodies the original work of the candidate and fulfills the requirements of the award. The work has been carried out under the guidance and supervision of the undersigned supervisors.

Prof. Dinesh Kumar Vishwakarma
Supervisor
Department of Information Technology,
Delhi Technological University, New
Delhi-110042 India

Dr. Bindu Verma
Co-Supervisor
Department of Information Technology,
Delhi Technological University, New
Delhi-110042 India

ABSTRACT

The detection of prohibited items in X-ray baggage imagery plays a vital role in ensuring public safety at transportation hubs, airports, railway stations, metro stations, and other high-security environments. X-ray image analysis provides critical visual information about concealed and overlapping objects inside baggage, enabling security personnel to identify potential threats more effectively. However, manual inspection of X-ray scans is time-consuming, labor-intensive, and prone to human error, especially in challenging scenarios involving cluttered backgrounds, severe occlusion, object overlap, and complex object orientations. Therefore, AI-based prohibited object detection systems have gained significant attention in recent years because of their potential to improve screening efficiency, detection accuracy, and operational reliability.

This thesis investigates key challenges associated with prohibited item detection in X-ray baggage imagery. Although deep learning-based object detection methods have achieved substantial progress over the past decade, challenges related to occlusion handling, feature degradation, multi-scale object representation, and robust detection in cluttered X-ray imagery remain insufficiently addressed. To overcome these limitations, this thesis proposes deep learning frameworks that effectively capture discriminative spatial and frequency-domain features from complex X-ray scans and accurately detect prohibited items under challenging conditions.

Specifically, two complementary detection frameworks are proposed. First, an Edge-Aware Deformable Feature Pyramid Network (EDFPN) is introduced to improve feature representation by enhancing object boundary information and spatial adaptability. The framework incorporates edge-aware feature enhancement operators and deformable convolution mechanisms to better capture irregular object structures and improve localization performance in cluttered environments. An empirical analysis is also conducted to evaluate the effectiveness of deformable feature learning and edge-guided representations for prohibited item detection.

Second, a Frequency-Spatial Guided Deformable Fusion Network is proposed for robust prohibited object detection in highly occluded and overlapping X-ray imagery. The proposed framework integrates frequency-domain feature extraction using Ga-

bor filtering with spatially adaptive deformable learning to capture both textural and structural information effectively. Furthermore, channel modulation and multi-path feature interaction mechanisms are incorporated to strengthen feature fusion and improve discriminative learning across multiple scales. These components collectively enhance the network's capability to detect concealed prohibited objects under challenging imaging conditions.

Both proposed approaches are designed, implemented, and extensively evaluated on benchmark X-ray security datasets, including PIDray, OPIXray, and HiXray. Experimental results demonstrate significant improvements in detection accuracy, robustness against occlusion and clutter, and real-time inference performance compared with existing state-of-the-art methods. The proposed systems provide efficient, reliable, and intelligent solutions for automated security screening applications, contributing toward safer and more effective public security infrastructure.

LIST OF PUBLICATIONS

Ranu Singh, Bindu Verma, and Dinesh Kumar Vishwakarma. "EDFPN: Edge-Aware Deformable Feature Pyramid Network for Occluded Object Detection in X-ray Imagery." In *Proceedings of International Conference on Communication and Computational Technologies (ICCCT)*, Springer, Lecture Notes in Networks and Systems (LNNS), 2026. (Scopus Indexed) (Accepted)

Ranu Singh, Bindu Verma, and Dinesh Kumar Vishwakarma. "Frequency–Spatial Guided Deformable Fusion for Prohibited Item Detection in X-ray Imagery." *Journal of Real-Time Image Processing*, Springer, 2026. (SCIE Indexed, Q2, IF: 3.0) DOI: 10.1007/s11554-026-01895-z. (Published).

Contents

Acknowledgement **i**

Candidate's Declaration **iii**

Certificate by the Supervisor **iv**

Abstract **v**

List of Publications **vii**

Table of Contents **viii**

List of Figures **xi**

List of Tables **xii**

List of Abbreviations **xiii**

1 INTRODUCTION **1**

1.1 Background of X-ray Security Screening Systems 2

1.2 Challenges in Prohibited Item Detection using X-ray Images 4

1.2.1 Object Overlapping 4

1.2.2 Occlusion 4

1.2.3 Low Contrast and Weak Texture 5

1.2.4 Cluttered Background 5

1.2.5 Geometric Deformation and Orientation Variation 5

1.2.6 Small Object Detection 5

1.2.7 Computational Complexity 5

1.3 Problem Statement 6

1.4 Research Objectives 6

1.5 Scope of the Thesis 7

1.6 Motivation of the Research 7

1.7 Contributions of the Thesis 8
1.8 Organization of the Thesis 8

2 Literature Survey **10**

3 Development of Edge-Aware Deformable Feature Pyramid Network **14**
3.1 Introduction 14
3.2 Proposed Methodology 16
3.2.1 Enhanced-Deformable Feature Pyramid Network 16
3.2.2 Edge-Aware Feature Operator(EFO) 17
3.2.3 Deformable Convolution Network 18
3.2.4 Neck 19
3.2.5 Head: Generalized Focal Loss 19
3.3 Experimental Setup 20
3.3.1 Dataset 20
3.3.2 Implementation Details 20
3.3.3 Evaluation criteria 20
3.4 Experimental Results and Discussion 21
3.4.1 Quantitative Results 21
3.4.2 Ablation Study 22
3.4.3 Computational Complexity Analysis 23
3.5 Conclusion 23

4 Development of Frequency–Spatial Guided Deformable Fusion Network **24**
4.1 Introduction 24
4.2 Proposed Methodology 25
4.2.1 Frequency-Spatial Guided Deformable Convolution Block (FSG-DCN) 26
4.2.2 Balanced Channel Bottleneck (BCB) 28
4.2.3 Tri-Path Frequency-Spatial Modulation (TriFSM) 29
4.2.4 Depthwise-Deformable Feature Pyramid Network (DDFPN) . 31
4.3 Experimental Setup 33
4.3.1 Datasets 33
4.3.2 Implementation Details 34
4.4 Experimental Results 34
4.4.1 Quantitative Analysis 34
4.4.2 Qualitative Analysis 37
4.4.3 Ablation Study 38

4.4.4 Run-time Analysis 40
4.5 Conclusion 40

5 Conclusion and Future Work **42**
5.1 Conclusion 42
5.2 Future Scope 44

Bibliography **46**

List of Figures

1.1 Examples of different occlusion levels in the PIDray dataset: (a) Easy - prohibited items are clearly visible with low overlapping, (b) Hard - objects are partially overlapped with other items, (c) Hidden - targets are heavily concealed within cluttered baggage, making detection challenging. 4

3.1 Architecture flowchart of the proposed EDFPN framework. The input image is processed through the backbone, edge-aware feature operator (EFO), deformable convolution modules, feature pyramid network (FPN), and detection head. 16
3.2 Overview of the EDFPN architecture. 17
3.3 Epoch-wise loss curves showing the decrease in total loss, classification loss, and bounding-box loss. 22

4.1 Overall architecture for detecting prohibited items. 26
4.2 Architecture of the proposed Frequency-Spatial Guided Deformable Convolutional Network (FSGDCN). 27
4.3 Structure of the TriFSM module. 31
4.4 Visual results of prohibited item detection on PIDRAY dataset [10]. Row 1: Ground truth boxes; Row 2: Prediction using ATSS; Row 3: Prediction using YOLOX; and Row 4: Prediction using proposed method. 38

List of Tables

3.1 Comparison with SOTA detectors on OPIXray [8] using VOC metrics. FK, SK, Sc, UK, and MK are stands for folding knife, straight knife, scissor, utility knife, and multitool knife, respectively. 21
3.2 Ablation results on the OPIXray dataset showing model component contributions. 23

4.1 Benchmark datasets of prohibited item detection. 33
4.2 Training hyperparameters used for proposed method. 34
4.3 Comparison with SOTA detectors on OPIXray [8]. FK, SK, Sc, UK, and MK denote folding knife, straight knife, scissor, utility knife, and multitool knife, respectively. A dash indicates unavailable data. 35
4.4 Evaluation of the proposed framework on the HiXray [11] dataset using VOC metrics. 36
4.5 Evaluation of the proposed framework on the PIDray [10] dataset using COCO metrics. 37
4.6 Impact of each module of proposed method on the OPIXray dataset. . 39
4.7 Impact of fixed vs. learnable Gabor filters on frequency response extraction. 39
4.8 Run-time performance comparison of different methods with input resolution of 640×640. 41
4.9 Deployment-oriented inference performance using PyTorch FP32 on a single GPU. 41

LIST OF ABBREVIATIONS

Abbreviation	Full Form
AI	Artificial Intelligence
AO-DETR	Anti-Overlapping Detection Transformer
AP	Average Precision
ATSS	Adaptive Training Sample Selection
BCB	Balanced Channel Bottleneck
BN	Batch Normalization
CA	Channel Attention
CNN	Convolutional Neural Network
COCO	Common Objects in Context
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
cuDNN	CUDA Deep Neural Network Library
DCN	Deformable Convolutional Network
DDFPN	Deformable Depthwise Separable Feature Pyramid Network
DETR	Detection Transformer
DFL	Distribution Focal Loss
DINO	DETR with Improved DeNoising Anchor Boxes
DML	Dual-Mode Learning
DOAM	De-occlusion Attention Module
EDFPN	Edge-Aware Deformable Feature Pyramid Network
EFO	Edge-Aware Feature Operator
FCOS	Fully Convolutional One-Stage Object Detection
FDTNet	Frequency Dual-Stream Transformer Network
FK	Folding Knife
FN	False Negative
FP	False Positive

FP16	16-bit Floating Point
FP32	32-bit Floating Point
FPN	Feature Pyramid Network
FPS	Frames Per Second
FSG-DCN	Frequency–Spatial Guided Deformable Convolution Network
GADet	Geometry-Aware Detector
GB	Gigabyte
GFL	Generalized Focal Loss
GFLOP	Giga Floating-Point Operation
GFLOPs	Giga Floating-Point Operations
GIoU	Generalized Intersection over Union
GPU	Graphics Processing Unit
HiXray	High-Quality X-ray Dataset
HSV	Hue, Saturation, and Value
IF	Impact Factor
IoU	Intersection over Union
L1	Least Absolute Deviations Loss
LIM	Latent Interaction Module
LLS	Label-Level Smoothing
LNNS	Lecture Notes in Networks and Systems
LR	Learning Rate
MCIA-FPN	Material-Aware Cross-Channel Interaction Attention Feature Pyramid Network
MK	Multitool Knife
MKL	Math Kernel Library
MM	Multimedia
MSANet	Multi-Scale Attention Network
NMS	Non-Maximum Suppression
OPIXray	Occluded Prohibited Items X-ray Dataset
PIDray	Prohibited Item Detection X-ray Dataset
Q2	Second Quartile
QFL	Quality Focal Loss
R-CNN	Region-Based Convolutional Neural Network
RCNN	Region-Based Convolutional Neural Network
RGB	Red, Green, and Blue

RT-DETR	Real-Time Detection Transformer
SCIE	Science Citation Index Expanded
SK	Straight Knife
SOTA	State of the Art
SSD	Single Shot MultiBox Detector
TP	True Positive
TriFSM	Tri-path Frequency–Spatial Modulation
UK	Utility Knife
VOC	Visual Object Classes
XAI	Explainable Artificial Intelligence
YOLO	You Only Look Once
YOLOX	You Only Look Once X

CHAPTER 1

INTRODUCTION

In recent years, ensuring public safety has become one of the most critical concerns for governments and security agencies across the world. Rapid urbanization, increasing passenger movement, and rising security threats have significantly increased the need for intelligent surveillance and inspection systems. Security screening systems are extensively deployed at airports, railway stations, metro stations, border checkpoints, government buildings, shopping malls, and public gathering areas for detecting dangerous and prohibited items such as firearms, knives, explosives, sharp tools, and other hazardous materials. Among various security inspection technologies, X-ray baggage screening systems are widely adopted because they provide non-invasive inspection of luggage and parcels without opening the baggage physically. These systems generate X-ray images that reveal the internal contents of baggage based on the material absorption characteristics of objects. Security operators analyze these scans manually to identify suspicious or prohibited items hidden inside the luggage.

However, manual inspection of X-ray images is a highly challenging and time-consuming task. In real-world scenarios, baggage often contains densely packed objects with severe overlap and clutter. Dangerous items may appear partially hidden behind other objects or may be visually deformed because of arbitrary orientations and occlusions. Moreover, X-ray images generally lack rich texture and color information that are normally available in natural RGB images. These characteristics make the identification of prohibited objects extremely difficult even for experienced security personnel. Human operators are also affected by fatigue, stress, workload, and lack of concentration during long screening hours, which may lead to missed detections and false alarms. Therefore, intelligent automated prohibited item detection systems are becoming increasingly important to assist human operators and improve the efficiency and reliability of security screening systems.

Artificial intelligence and deep learning have made a significant progress in com-

puter vision. In the field of natural image datasets, deep learning object detection algorithms such as Faster R-CNN, YOLO, SSD, RetinaNet and DETR have demonstrated impressive performance. These models learn hierarchical feature representations automatically and perform well in complex visual tasks. Despite these developments, object detection in X-ray imagery remains challenging, since many conventional object detection assumptions do not directly apply. The characteristics of X-ray images are different from those of natural images. Heavy occlusion, object overlap, low contrast, cluttered backgrounds, transparency and geometric distortion significantly affect the performance of standard detectors. The detection of small items like blades, wires, cutters and small arms is particularly challenging because the boundaries and structure of the object can become obscured by surrounding objects.

To overcome these drawbacks, researchers have suggested some special frameworks for detection of prohibited items in X-ray imaging. The approaches involve contour-based enhancement of features, deformable convolution networks, multi-scale feature fusion, frequency-aware learning strategies, Attention mechanisms, and transformer-based architectures. These approaches have enhanced the ability to detect objects, but there are still numerous issues that need to be addressed, especially when dealing with severe occlusion and overlapping objects.

In this thesis, two advanced deep learning frameworks are developed for robust prohibited item detection in X-ray imagery. The first framework introduces an Edge-Aware Deformable Feature Pyramid Network (EDFPN), which improves boundary-sensitive representation learning and geometric adaptability using edge-aware feature extraction and deformable convolutions. The second framework proposes a Frequency–Spatial Guided Deformable Fusion Network that combines frequency-domain analysis and spatial feature modeling for improved multi-scale feature representation and occlusion-aware detection. The proposed frameworks aim to improve the reliability, robustness, and accuracy of intelligent X-ray screening systems under real-world complex security conditions.

1.1 Background of X-ray Security Screening Systems

X-ray imaging systems are one of the most commonly used technologies for baggage inspection and security surveillance. These systems utilize electromagnetic radiation to penetrate baggage and generate internal structural representations of objects. Different materials absorb X-ray radiation differently depending upon their density and composition. Based on the absorption patterns, X-ray systems generate grayscale or pseudo-colored images representing the contents inside the luggage. Security screen-

ing systems play an essential role in detecting prohibited objects before they pose a threat to public safety. Such systems are widely used in:

- Airports
- Railway stations
- Metro stations
- Government organizations
- Military checkpoints
- Commercial complexes
- Border security
- Public event venues

Traditional X-ray baggage inspection systems mainly depend on human operators for detecting suspicious objects. However, manual inspection suffers from several practical limitations:

- Large passenger volume
- Continuous monitoring workload
- Fatigue during long inspection hours
- Human errors and missed detections
- Difficulty in analyzing cluttered baggage

As a result, automated detection systems based on computer vision and artificial intelligence are increasingly being adopted for assisting security personnel.

Deep learning has significantly improved automated image analysis and object detection capabilities. Convolutional Neural Networks (CNNs) can automatically learn discriminative feature representations from large-scale image datasets. Modern detection frameworks such as Faster R-CNN, YOLO, RetinaNet, and DETR have achieved state-of-the-art performance in natural image analysis.

However, X-ray prohibited item detection remains considerably more challenging than natural image detection because of several unique visual properties associated with X-ray imagery.

1.2 Challenges in Prohibited Item Detection using X-ray Images

Although deep learning methods have achieved impressive performance in natural image analysis, several major challenges still exist in X-ray prohibited item detection systems.

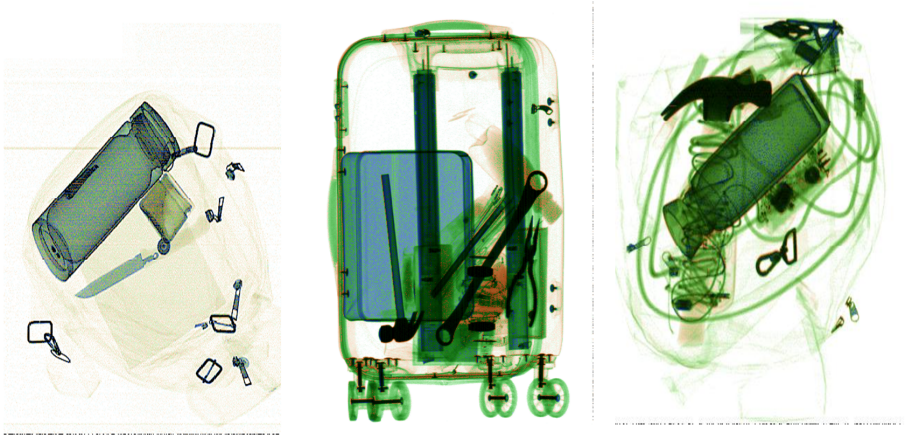


Figure 1.1: Examples of different occlusion levels in the PIDray dataset: (a) Easy - prohibited items are clearly visible with low overlapping, (b) Hard - objects are partially overlapped with other items, (c) Hidden - targets are heavily concealed within cluttered baggage, making detection challenging.

1.2.1 Object Overlapping

One of the most critical challenges in X-ray baggage analysis is object overlap. Multiple objects are usually packed closely inside baggage, causing different items to overlap in the X-ray image. Dangerous objects may become partially or fully hidden behind other benign items. Such overlap significantly reduces object visibility and makes feature extraction difficult for conventional detectors.

1.2.2 Occlusion

Occlusion occurs when prohibited items are partially concealed by surrounding objects. In practical baggage scenarios, weapons and hazardous items may appear with only a small visible region. Heavy occlusion leads to:

- Poor localization
- Boundary ambiguity

- Misclassification
- Missed detection

Conventional CNN-based detectors generally fail to capture discriminative features under severe occlusion conditions.

1.2.3 Low Contrast and Weak Texture

Unlike RGB images, X-ray scans lack rich texture and color information. Many prohibited items appear with weak structural contrast against surrounding objects. Low contrast conditions reduce the effectiveness of standard convolutional operations because object boundaries become difficult to distinguish.

1.2.4 Cluttered Background

X-ray baggage images often contain large numbers of unrelated objects arranged randomly. Complex cluttered backgrounds introduce noise and irrelevant features, reducing the robustness of feature learning. This challenge becomes more severe when multiple metallic and organic materials coexist inside the baggage.

1.2.5 Geometric Deformation and Orientation Variation

Objects inside baggage may appear in arbitrary orientations, scales, and deformed shapes. Conventional convolution kernels use fixed sampling locations and therefore fail to adapt effectively to irregular object geometries. This reduces detection performance for small and distorted prohibited items.

1.2.6 Small Object Detection

Several prohibited items such as blades, wires, cutters, and small firearms occupy very small regions in X-ray images. Detecting such small objects is difficult because important structural details are often lost during feature downsampling.

1.2.7 Computational Complexity

Modern transformer-based and attention-driven architectures often achieve high detection accuracy but introduce significant computational overhead. Real-time security systems require models that are both accurate and computationally efficient. Therefore, balancing detection performance and inference speed remains an important research problem.

1.3 Problem Statement

The main objective of this research is to develop robust deep learning algorithms for detecting prohibited items in X-ray images under challenging conditions, such as severe object overlap, occlusion, cluttered backgrounds, low contrast and geometric distortion. Current object detection systems have several limitations:

- Limited ability to capture fine structural boundaries
- Weak feature representation in overlapping regions
- Poor geometric adaptability.
- Inefficient multi-scale feature fusion
- High computational complexity
- Limited frequency-domain representation learning

In this respect, efficient multi-scale fusion and feature enhancement frameworks that can enhance the detection robustness in complex X-ray security environments are needed. This thesis aims to address these limitations by incorporating:

- Edge-aware feature learning
- Frequency-domain representation
- Deformable convolution
- multi-scale Feature Fusion
- Channel modulation
- Spatial-frequency interaction mechanisms

1.4 Research Objectives

The major objectives of this thesis are summarized below:

1. To investigate edge-aware and deformable feature learning mechanisms for enhancing overlap handling, boundary localization, and geometric adaptability of X-ray prohibited item detection systems.
2. To develop an effective frequency–spatial guided deep learning system for improving texture representation, structural feature extraction and multi-scale detection performance in complex baggage screening scenarios.

1.5 Scope of the Thesis

This thesis proposes novel deep learning-based object detection architectures that can be used to detect prohibited objects in X-ray security imagery. This study includes the following aspects:

- Object detection under occlusion
- Analysis of overlapping objects
- Learning multi-scale features
- Frequency-spatial representation
- Deformable feature extraction
- Efficient feature pyramid networks

The developed models are tested on publicly available benchmark datasets for X-rays and are intended for use in intelligent baggage screening systems. While the proposed approaches have improved detection accuracy significantly, this thesis is limited to 2D X-ray image analysis and does not address multi-view or 3D baggage reconstruction.

1.6 Motivation of the Research

This research is motivated by the increasing demand for intelligent and reliable automated security screening systems. Current baggage inspection systems continue to rely on human interpretation, and remain susceptible to human error in complex screening environments. Current deep learning models struggle to effectively handle complex X-ray imaging scenarios, such as those with substantial clutter and overlapping objects, since they primarily focus on spatial information, while neglecting key characteristics in the frequency domain and other structural aspects. Moreover, traditional feature pyramid structures adopt fixed convolution operations that lack geometric adaptability. These limitations motivate the development of more sophisticated feature enhancement systems that are able to represent both spatial and frequency-aware features. The proposed research aims to contribute to:

- Safer public security systems
- Intelligent automated baggage inspection

- Improved prohibited item detection accuracy
- Reduced false alarms
- Real-time security applications

1.7 Contributions of the Thesis

The major contributions of this thesis focus on developing advanced deep learning frameworks for robust prohibited item detection in X-ray security imagery under challenging conditions such as severe overlap, occlusion, and cluttered backgrounds.

An **Edge-Aware Deformable Feature Pyramid Network (EDFPN)** is proposed to improve contour-sensitive feature representation and localization accuracy in heavily occluded X-ray images. The framework integrates edge-aware feature enhancement, deformable convolution, and multi-scale feature fusion for robust prohibited item detection. An Edge-Aware Feature Operator (EFO) is introduced to enhance boundary-sensitive information using gradient-based edge extraction. The proposed module improves contour localization, thin object detection, and overlap handling capability.

Further, a **Frequency–Spatial Guided Deformable Fusion Framework** is developed to jointly exploit spatial-domain and frequency-domain representations for enhanced structural feature extraction. The framework incorporates learnable Gabor filtering, frequency-aware channel encoding, tri-path modulation, and deformable depthwise convolution for adaptive feature learning. In this framework, a Balanced Channel Bottleneck (BCB) mechanism is proposed for adaptive channel recalibration and suppression of redundant feature responses, thereby improving discriminative representation learning.

Extensive experiments conducted on benchmark datasets, including PIDray, OPIXray, and HiXray, demonstrate that the proposed frameworks achieve superior prohibited item detection performance compared with several existing state-of-the-art methods.

1.8 Organization of the Thesis

The thesis is organized into six chapters as follows:

Chapter 1: Introduction

This chapter presents the background, motivation, research challenges, objectives, contributions, and organization of the thesis.

Chapter 2: Literature Survey

This chapter reviews existing research related to prohibited item detection, deep learning-based object detection, deformable convolution, frequency-domain learning, and multi-scale feature fusion.

Chapter 3: Development of Edge-Aware Deformable Feature Pyramid Network

This chapter presents the proposed EDFPN framework including edge-aware feature extraction, deformable convolution refinement, and feature pyramid fusion.

Chapter 4: Development of Frequency–Spatial Guided Deformable Fusion Framework

This chapter presents the proposed frequency–spatial guided network including FSG-DCN, balanced channel bottleneck, tri-path modulation, and deformable depthwise feature pyramids.

Chapter 5: Conclusion and Future Scope

This chapter summarizes the research contributions and presents possible future research directions.

CHAPTER 2

LITERATURE SURVEY

Automatic prohibited item detection in X-ray imagery has become an important research area due to the increasing deployment of intelligent security screening systems in airports, railway stations, metro stations, border checkpoints, and public surveillance infrastructures. X-ray baggage inspection systems are extensively used to identify dangerous and prohibited objects such as firearms, knives, explosives, and sharp metallic tools concealed inside luggage. However, automatic detection in X-ray imagery remains highly challenging because X-ray scans generally contain severe object overlap, cluttered backgrounds, low texture contrast, transparency effects, and partial occlusion. These characteristics significantly reduce the effectiveness of conventional object detection frameworks and motivate the development of specialized deep learning architectures for prohibited item detection.

In the initial research on prohibited object detection, there was a focus on manually designed feature extraction techniques and using standard machine learning classifiers. Turcsany et al. [1] adopted handcrafted image descriptors and statistical feature representation to improve the object recognition process for X-ray baggage screening. Likewise, Mery et al. [2] introduced the use of sparse representation for analysing images to inspect the luggage. For this, they borrowed the concept of adaptive sparse coding to enhance the ability to recognize objects in X-ray images acquired under sub-optimal imaging conditions. These manual techniques were found to be effective in clean baggage scenes with minimal clutter and moderate typical occlusion and overlap, but were not yet well developed in high occluded and over-lappable ones.

Deep Learning greatly revolutionized Object Detection. Convolutional Neural Network (CNN) -based detectors performed greatly well on visual recognition tasks due to their capability of learning the hierarchical representation of features from large-scale datasets. Redmon et al. [24] proposed a unified regression problem to solve object detection in the real-time scenario, named YOLO. Liu et al. [33] introduced the Single

Shot MultiBox Detector (SSD) that detects objects of varying scales using the multi-scale feature map. Faster R-CNN (Ren's et al. [12]) has been proposed by Ren et al., which uses Region Proposal Networks (RPNs) to generate and localize the regions of interest efficient. These general purpose detectors were later modified by several studies for the purpose of X-ray prohibited item detection. However, Galvez et al. [4], Bhowmik et al. [5] and Chaturvedi et al. [6] fine-tuned CNN-based baggage screening detectors and achieved better results than handcrafted ones. However, even with these improvements, when the images are heavily overlapped and occluded, deep-learning models still operate poorly on X-ray images since there is a lack of texture and color information in X-ray images compared to natural RGB images.

For this reason, different methods for occlusion-aware feature enhancement were investigated. Akcay et. al. [3] explored transfer learning methods for fine-tuning pre-trained CNN models to the X-ray domain, where data is scarce. They found that transferring the knowledge derived from one good detector to another one could greatly boost the detection performance of the restricted items when labeled data was scarce. Pre-trained models based on RGB images, however, are not enough to capture the structural details in X-ray images. To handle cluttered baggages, Wei et al. [8] introduced the De-Occlusion Attention Module (DOAM) to combine the edge enhancement and material-aware contextual cues for prohibited items detection. The attention mechanism extends to Southeast regions that carry useful information and/or background interference to generate a higher level of detection performance in severe overlap. Shao et al. [29] proposed a framework of foreground–background separation with dual decoders and attention mechanisms to separate the prohibited objects in the foreground and suppress background irrelevant clutter to enhance the detection accuracy of prohibited objects. They proposed a framework to enhance feature disentanglement and alleviate the interference from overlapped background objects. Likewise, ForkNet [34] is a foreground and background feature disentanglement framework which uses dual backbones to learn foreground and background representations separately. This architecture significantly reduced the clutter interference and improved the extraction of features from overlap sensitive regions. To boost the performance of X-ray object detection, Zhang et al. [30] proposed an improved Mask R-CNN with low-level feature aggregation and Soft-NMS [32, 31]. Under overlapping conditions, the integration of Soft-NMS reduced the suppression errors due to CNM by introducing a smooth transition between winner and loser subsystems.The integration of Soft-NMS reduced the suppression errors due to CNM, by introducing a smooth transition between the winner and loser subsystems, under the overlapping condition.

Despite these advancements, severe overlap and partial occlusion remain major

challenges in X-ray prohibited item detection because hazardous objects frequently appear distorted, partially hidden, or overlapped with other baggage contents. To improve geometric adaptability under such conditions, deformable convolution-based frameworks were introduced. Dai et al. [25] proposed Deformable Convolutional Networks (DCN), which introduced learnable spatial offsets for adaptive feature sampling. Unlike standard convolution operations that use fixed receptive fields, deformable convolution dynamically adjusts sampling locations according to object geometry, thereby improving localization under geometric deformation and irregular object structures. Wang et al. [37] further improved this concept through DCNv2, which enables flexible off-grid sampling and enhances spatial adaptability to pose variations and irregular boundaries. Deformable convolution significantly improved feature alignment and overlap handling in dense object detection tasks. Its success also motivated transformer-based deformable attention architectures. Zhu et al. [16] proposed Deformable DETR, which extends transformer-based object detection using sparse deformable attention sampling. By focusing on informative multi-scale locations, Deformable DETR achieved improved convergence rates and small-object detection accuracy compared with conventional DETR models.

Recent research works further focused on improving structural representation and contextual modeling for prohibited item detection. Ma et al. [17] proposed a feature enhancement strategy integrating learnable Gabor filters, spatial attention mechanisms, and global context-aware multi-scale aggregation for improving robustness under severe overlap conditions. Their approach demonstrated that frequency-aware feature extraction can significantly improve contour and texture representation in X-ray imagery. Similarly, Zhu et al. [22] proposed FDTNet, a dual-stream transformer framework incorporating frequency-aware representations for capturing high-frequency edge and contour information. The proposed frequency-domain learning strategy improved localization of thin and partially occluded prohibited objects. Li et al. [23] proposed AO-DETR, an overlap-aware transformer framework designed specifically for heavily overlapping prohibited item detection.

In addition, the framework introduced occlusion-sensitive attention mechanisms and spatial separation strategies to improve localization accuracy in densely overlapping settings. Yang et al. [40] proposed DML-Net, a dual-mode learning framework that jointly learns shared and dataset-specific feature representations across multiple datasets. This approach improved generalization capability in diverse X-ray imaging environments. Li et al. [38] proposed GADet, which uses geometric constraints and IoU-aware label assignment mechanisms to address severely overlapped prohibited objects. Their method achieved promising localization improvements for heavily

overlapping prohibited objects. Chen et al. [39] used mix-paste augmentation strategies to reduce the influence of noisy annotation data by simulating overlap through mixed object patches. Wei et al. [41] introduced multi-task teacher-student knowledge transfer strategies to improve hard-sample learning and prohibited item representation capability. More recently, Wu et al. [43] introduced MSANet, which adopts triple-attention mechanisms with multi-scale contextual fusion. This framework improves feature representation for occluded and scale-varied prohibited items, enabling better detection of partially visible prohibited objects.

These techniques have helped to improve accuracy in detecting prohibited items while addressing some issues remain unsolved. Fine structural boundaries, contour-sensitive information and frequency-domain representations are all necessary for detecting thin, partially occluded, and highly overlapping prohibited objects and still remain problematic with many existing approaches. In addition, transformer models tend to have excessive computation overhead and need massive amounts of training data to be practical to deploy.

The restrictions underscore the requirement for an efficient and powerful feature representation framework that can represent jointly edge sensitive structures, frequency, adaptive geometric alignment and multi-scale contextual information. From the above observations, the following contributions are made to tackle the prohibited item detection problem in difficult X-ray security screening scenes.Based on the above observations, the following contributions are presented in this thesis to address the problem of prohibited item detection in problem scenes during X-ray security screening.

CHAPTER 3

DEVELOPMENT OF EDGE-AWARE DEFORMABLE FEATURE PYRAMID NETWORK

3.1 Introduction

Nowadays, detection and recognition of objects is a very important application in security systems, particularly in the detection of dangerous object, including weapons, explosives and sharp objects. These systems are utilized in airports, government organizations, transportation centers, as well as in huge numbers of people involvement. Even though the object detection process has been greatly developed, it is still difficult to recognize these objects under the challenging condition of occlusion. Sometimes, a target object is partially or entirely obstructed by other objects, making it hard for typical detectors to identify. In a real situation, harmful products can be protected by clothes, electronics or other items in the middle of tightly packed or crowded bags and luggage. This could contribute to an undetected event and grave security issues. When it comes to x-ray images, it makes this challenge even more difficult. Unlike natural RGB images, X-ray images are poor in texture, may have cluttered background, overlapping items and have similar categories of items. As the receptive field and sampling pattern are fixed, it is hard for the standard detectors (Ren et al. [12], Redmon et al. [24] and Lin et al. [28]) to separate the contour of objects or to identify overlapping objects in X-ray image. To overcome these challenges, researchers have studied different methods that rely only on X-rays; however, these methods have been seen as a major challenge in the literature.In literature, researchers have studied and analyzed different methods that involve only X-ray, and are referred to as X-ray specific approaches. These techniques however, have their limitations in extreme occlusion instances and in data-limited cases.

Early approaches on X-ray object recognition mainly depended on handcrafted features and sparse representation techniques, however, these approaches were not robust enough for the complex scenarios of the real world. With the advent of deep learning, CNN-based detectors like Faster R-CNN, YOLO and RetinaNet increased detection accuracy to a large extent by learning hierarchical feature representations. More recent studies have explored occlusion-aware and attention-driven strategies to enhance prohibited item detection. For instance, Hassan et al. [7] introduced contour proposal learning, while Wei et al. [8] proposed a de-occlusion attention module to improve feature recovery under heavy occlusion. Zhao et al. [9] developed a label-aware overlap handling mechanism to mitigate object interference. Additionally, Wang et al. [21] proposed the material-aware cross-channel interaction network (MCIA-FPN) to improve feature fusion for occluded prohibited item detection. Zhu et al. [22] introduced a frequency-aware transformer architecture, namely FDTNet, for improved contextual modeling in X-ray imagery. Jia et al. [34] explored ForkNet-based disentanglement frameworks to separate overlapping objects in security scans. Li et al. [23] proposed Anti-Overlapping DETR (AO-DETR) to address overlapping object detection in complex X-ray scenes. Despite their promising performance gains, many of these approaches suffer from high computational cost, increased model complexity, or limited boundary sensitivity under severe occlusion.

To improve detection under such circumstances, this study presents **EDFPN (Edge-Aware Deformable Feature Pyramid Network)**, a compact yet effective feature enhancement module. By using an edge-aware feature operator (EFO) and deformable convolutions, EDFPN improves spatial flexibility and reinforces boundary information. When combined, these elements produce richer multi-scale features that improve detection performance in highly congested X-ray settings.

The primary contributions of this work are outlined below:

- This thesis proposes **EDFPN (Edge-Aware Deformable Feature Pyramid Network)**, a novel feature pyramid enhancement framework designed to improve prohibited item detection in highly occluded X-ray imagery.

- This thesis introduces an **Edge-Aware Feature Operator (EFO)** that explicitly enhances boundary and contour information using gradient-based edge extraction, improving localization of overlapping and low-contrast objects without adding learnable parameters.

- To solve the above problem, this thesis integrates **deformable convolution** into the feature pyramid, in order to achieve the effect of spatially adaptive recep-

tive fields, which will increase the geometric alignment and robustness of the network under severe occlusion.

- Extensive experiments are conducted on the dataset of the OPIXray benchmark and prove that EDFPN is able to gain up to 83.2% of the mAP, performing better in comparison with several state-of-the-art detectors that operate in the conditions of occlusion.

3.2 Proposed Methodology

This section describes the methodology of EDFPN, which consists of three key components: an edge-aware feature operator for enhancing boundary information, deformable convolutions for spatially adaptive refinement, and an FPN-style neck for multi-scale feature fusion. It also outlines the loss functions used in the detection head, namely Quality Focal Loss, Distribution Focal Loss, and GIoU Loss. The overall architecture of the proposed EDFPN framework is illustrated in Fig. 3.1.

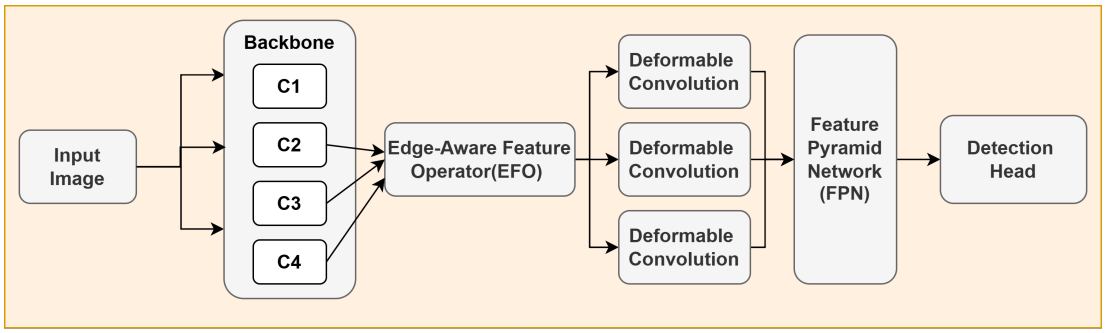


Figure 3.1: Architecture flowchart of the proposed EDFPN framework. The input image is processed through the backbone, edge-aware feature operator (EFO), deformable convolution modules, feature pyramid network (FPN), and detection head.

3.2.1 Enhanced-Deformable Feature Pyramid Network

In this work, the EDFPN module is introduced as a neck network to enhance multi-scale feature representation. As shown in Fig. 3.2, EDFPN comprises three components: an Edge-Aware Feature Operator for extracting boundary information, deformable convolutions for improving geometric adaptability, and a top-down FPN-style fusion for generating semantically consistent multi-scale features. The refined features are merged through lateral and output convolutions, while an additional max-pooling layer extends the pyramid for deeper detection. The following sections briefly explain each component of EDFPN.

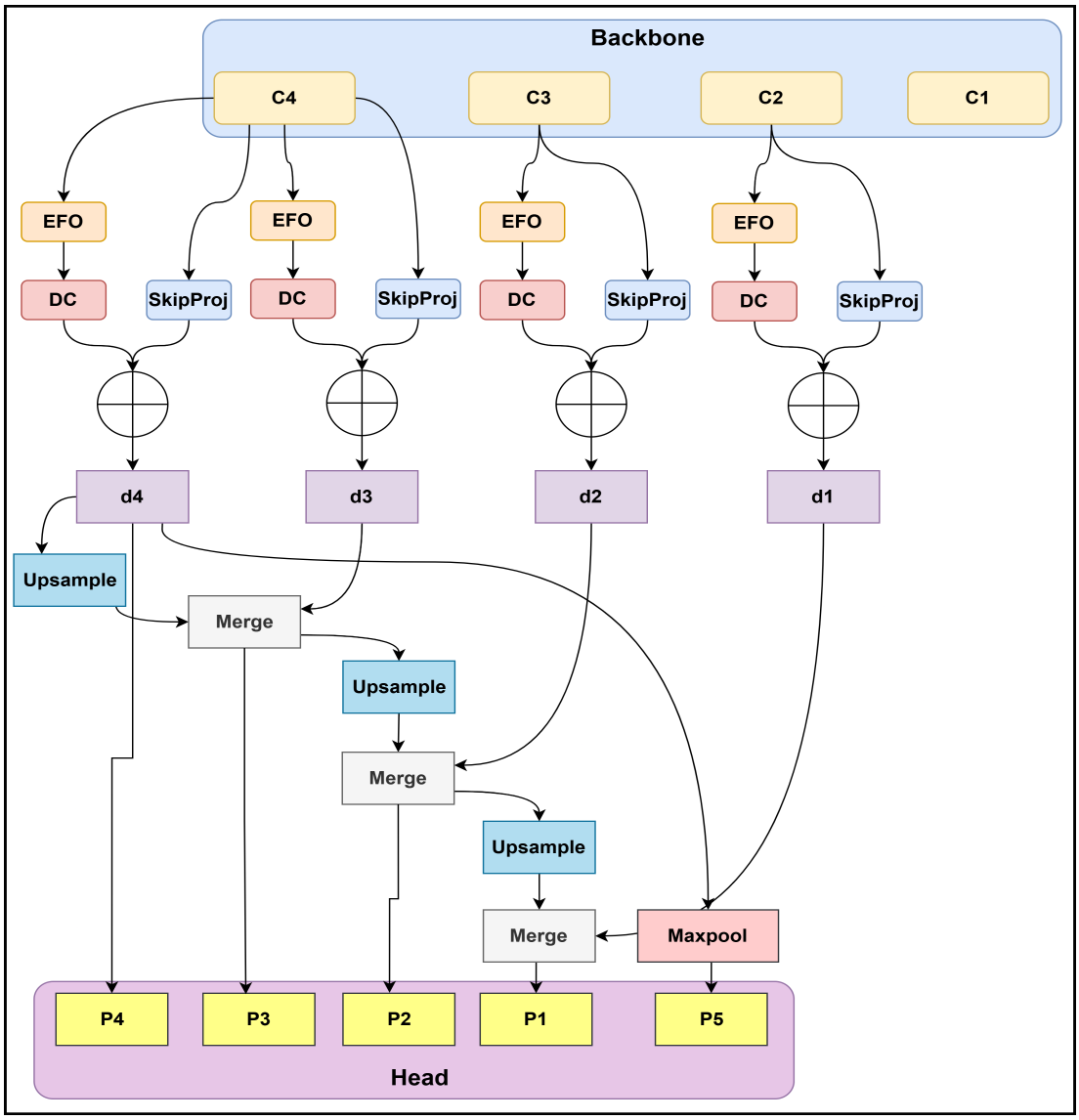


Figure 3.2: Overview of the EDFPN architecture.

3.2.2 Edge-Aware Feature Operator(EFO)

This module enhances spatial features by explicitly extracting edge information, which is crucial for identifying the boundaries of prohibited items in X-ray images. As boundary cues are usually blurred in semantically interesting objects by overlapping objects and low-contrast materials, a gradient-based operator is developed to deliver on-demand boundary information.

Given an input feature map, $F \in \mathbb{R}^{C\times H\times W}$ in the form of a matrix, a fixed 1×1 convolution then makes a grayscale projection:

$$F_{\text{gray}} = \frac{1}{C} \sum_{c=1}^{C} F_c. \tag{3.1}$$

The spatial gradients are then calculated via predetermined Sobel filters: $G_x = S_x * F_{\text{gray}}, \qquad G_y = S_y * F_{\text{gray}}$, and the edge magnitude derived as:

$$F_{\text{edge}} = \sqrt{G_x^2 + G_y^2 + \epsilon} \tag{3.2}$$

The resultant edge map is then added to the original feature map by channel concatenation, $F_{\text{out}} = \text{Concat}(F, F_{\text{edge}})$. Since this operation introduces no additional learnable parameters, it offers a simple way to reinforce boundary information. As a result, the network gains clearer structural cues, which helps in locating thin, overlapping, or partially hidden objects commonly seen in X-ray imagery.

3.2.3 Deformable Convolution Network

To enhance spatial adaptability, EDFPN applies deformable After edge enhancement, EDFPN applies deformable convolutions to the Sobel-augmented features. In contrast to standard convolutions that rely on fixed sampling locations, deformable convolutions learn spatial offsets, allowing the network to adjust its sampling pattern and better handle irregular shapes, occlusion, and other variations commonly seen in X-ray images.

Let the combined feature map be represented by $F' = \text{Concat}(F, F_{\text{edge}})$. For every spatial position p, the deformable convolution gathers data using learned offsets. Δp_k:

$$F_d(p) = \sum_{k=1}^{K} w_k \, F'(p + p_k + \Delta p_k) \tag{3.3}$$

where w_k are the convolution weights and p_k are the standard kernel offsets. This technique helps the network focus on physically important regions, such as metal boundaries or overlapping object edges.

To preserve the integrity of the refined features with the backbone's semantic structure, each level C_l is added back to the deformable output following a 1×1 projection: $\hat{D}_l = D_l + \text{Proj}(C_l)$ ensures that the updated features preserve high-level context while enhancing spatial detail.

3.2.4 Neck

After deformable refinement, EDFPN incorporates multi-scale features using an FPN-style structure. Let $\{F_1, F_2, \ldots, F_L\}$ represent the feature maps from different backbone levels. First, each feature is projected to a single channel dimension using a 1×1 convolution: $\hat{F}_l = \text{Conv}_{1\times1}(F_l), \quad l = 1, \ldots, L$ A top-down process is then used to upsample and integrate higher-level features with lower-level features: $G_l = \hat{F}_l + \text{Up}(\hat{F}_{l+1}), \quad l = L-1, \ldots, 1$, where $\text{Up}(\cdot)$ denotes $2\times$ upsampling. This method helps detect thin or small objects by adding stronger semantic information to the bottom layers. Each merged feature is then subjected to a 3×3 convolution. $P_l = \text{Conv}_{3\times3}(G_l), \quad l = 1, \ldots, L$

To extend the hierarchy, a max-pooling layer generates an additional deeper feature, $P_{L+1} = \text{MaxPool}(P_L)$,. The multi-scale set $\{P_1, P_2, \ldots, P_{L+1}\}$ is the result of this. This structure enhances the detector's effectiveness across a variety of item sizes and in busy or obstructed environments by fusing high-level context with spatial information.

3.2.5 Head: Generalized Focal Loss

The Generalized Focal Loss (GFL) head, which simultaneously learns localization and classification, is used by the model. GFL uses Quality Focal Loss (QFL) to integrate the localization quality into the classification step rather than handling these tasks independently. The IoU score $q \in [0, 1]$ serves as a continuous monitoring signal in this case:

$$\mathcal{L}_{\text{QFL}} = -|p_t - q|^{\beta} \log(p_t). \tag{3.4}$$

Distribution Focal Loss (DFL) and the GIoU loss are combined by GFL for regression. Every coordinate is represented by DFL as a probability distribution across K bins:

$$\mathcal{L}_{\text{DFL}} = \sum_{i=1}^{K} \text{SoftLabel}(x, i) \log(p_i), \tag{3.5}$$

while GIoU provides a more informative penalty when predicted and ground-truth boxes do not overlap:

$$\mathcal{L}_{\text{GIoU}} = 1 - \left(\text{IoU} - \frac{|C \setminus (A \cup B)|}{|C|} \right). \tag{3.6}$$

Together, these losses help the model generate better-calibrated confidence scores and more accurate bounding boxes, which is particularly useful for detecting small,

crowded, or overlapping objects in X-ray imagery.

3.3 Experimental Setup

3.3.1 Dataset

One of the most popular datasets for evaluating detection accuracy of the prohibited object recognition in dual-energy X-ray images is the OPIXray dataset [8]. The data collection is intended to simulate real scenarios when objects can be hidden in bags filled with a lot of other items, heavily occluded, or be obscured by other objects. A total of 7,109 images were available for training and 1,776 images for testing, focusing on five classes of bladed objects - utility knife, multi-tool knife, scissors, folding knife and straight knife.

3.3.2 Implementation Details

All experiments were conducted on an Ubuntu system using Python 3.8.20 and a single NVIDIA RTX A5000 GPU with CUDA 11.8 and cuDNN 9.1. The models were trained using MMEngine 0.10.7 and implemented in PyTorch 2.4.1 using TorchVision 0.19.1. Data preparation was handled by OpenCV 4.10.0, while CPU activities were accelerated by using Intel MKL with OpenMP. For reproducibility, training was conducted in a non-distributed setting with fixed random seeds and `cudnn_benchmark` deactivated. This arrangement provides a dependable balance between accuracy and efficiency.

3.3.3 Evaluation criteria

The spatial agreement between predicted and ground-truth boxes is ignored by traditional metrics such as classification accuracy, making them insufficient for object detection. As a result, *mean Average Precision (mAP)* is frequently used to assess the accuracy of a forecast using Intersection over Union (IoU).

True positives (TP), false positives (FP), and false negatives (FN) are the three categories of detections based on IoU. The definition of precision and recall is: Precision $= \frac{TP}{TP+FP}$, Recall $= \frac{TP}{TP+FN}$. The area under the Precision–Recall curve is used to calculate AP for each class: $AP = \int_0^1 P(R)\, dR$. The mean Average Precision is:

$$mAP = \frac{1}{C} \sum_{i=1}^{C} AP_i. \tag{3.7}$$

Thus, mAP provides a more comprehensive evaluation than accuracy, as it reflects both classification reliability and localization quality across all object categories.

Table 3.1: Comparison with SOTA detectors on OPIXray [8] using VOC metrics. FK, SK, Sc, UK, and MK are stands for folding knife, straight knife, scissor, utility knife, and multitool knife, respectively.

Models	**Ref.**	**Parameter (M)**	**G-FLOPs**	**mAP**	**FK**	**SK**	**Sc**	**UK**	**MK**
FasterR-CNN (TPAMI'17)	[12]	41.19	20.36	70.9	75.1	45.8	88.4	67.7	77.3
RetinaNet (ICCV'17)	[28]	38	239	70.9	77.1	36.1	94.6	64.6	82.2
CascadeR-CNN (CVPR'18)	[15]	68.97	22.37	72.8	75.7	50.0	89.4	70.0	79.0
GFLv1 (NeurIPS'20)	[19]	50.70	28.51	75.6	80.0	53.6	89.3	71.7	83.4
SparseR-CNN (CVPR'21)	[26]	106	278	81.8	87.7	64.3	89.3	79.9	87.9
DINO(ICLR'23)	[20]	58.38	26.89	78.2	83.2	58.8	89.4	72.7	86.7
DOAM(ACM MM'20)	[9]	90.79	-	82.41	86.7	68.5	90.2	78.8	87.6
MCIA-FPN (TVC'23)	[21]	-	-	82.5	89.0	74.4	89.9	86.1	89.7
FDTNet (EAAI'24)	[22]	66.17	-	82.0	87.9	60.2	96.1	78.9	87.1
AO-DETR(R) (TNNLS'25)	[23]	58.38	26.89	79.2	83.8	60.5	90.1	74.7	87.1
AO-DETR(S) (TNNLS'25)	[23]	229.0	156	80.8	84.8	63.0	90.1	77.7	88.4
Proposed Model	-	**45.91**	**44.25**	**83.2**	**88.5**	**66.9**	**90.3**	**82.6**	**87.7**

3.4 Experimental Results and Discussion

3.4.1 Quantitative Results

The proposed EDFPN achieves 83.2% mAP on the OPIXray dataset, outperforming multiple state-of-the-art detectors including Faster R-CNN, RetinaNet, Sparse R-CNN, DINO, and AO-DETR, as shown in Table 3.1. The improvement demonstrates the effectiveness of combining edge-aware boundary enhancement with deformable feature alignment, which strengthens detection of partially occluded and overlapping prohibited items.

EDFPN demonstrates significant improvement on the detection of folding knives (88.5%), utility knives (82.6%), and scissors (90.3%), which shows better boundary localization and resistance to noise in baggage with clutter. EDFPN is efficient as it provides similar accuracy to AO-DETR(S) and has lower computational overhead.

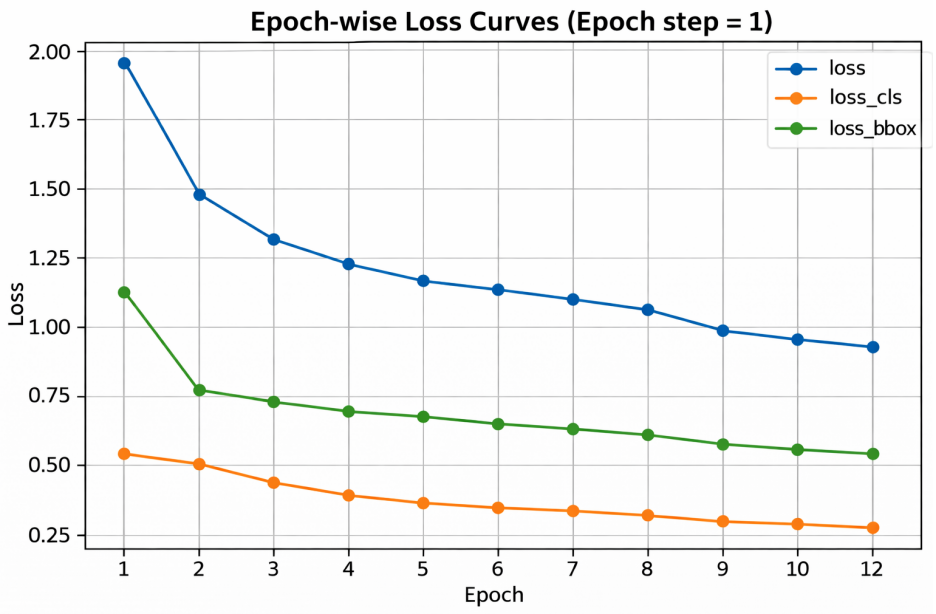


Figure 3.3: Epoch-wise loss curves showing the decrease in total loss, classification loss, and bounding-box loss.

This can be seen in Figure 3.3 where all loss elements have been decreasing steadily, which shows consistent optimization and increasingly improved feature weighting. Overall, the results confirm that EDFPN significantly improves multi-scale representation quality and enhances the robustness of automated X-ray security inspection systems.

3.4.2 Ablation Study

An ablation study was performed on the OPIXray dataset to assess the contribution of each component, as indicated in Table 3.2. 79.9% mAP was attained by the baseline model. Performance was greatly enhanced to 81.0% by adding deformable convolution, demonstrating the advantage of variable spatial sampling. Accuracy goes to 80.4% when the EFO module was included, suggesting that explicit edge-based augmentation yields comparable benefits. The best result of 83.2% was obtained by the entire EDFPN architecture, demonstrating that the greatest significant improvement is obtained when deformable feature extraction and optimal fusion are combined. Overall, each part makes a significant contribution; the biggest increases in accuracy are

provided by deformable convolution and the whole EDFPN.

Table 3.2: Ablation results on the OPIXray dataset showing model component contributions.

Model Variant	**mAP (%)**
baseline	79.9
+ Deformable Convolution	81.0
+ EFO	80.4
+ EDFPN(full model)	**83.2**

3.4.3 Computational Complexity Analysis

Although additional feature enhancement operations are introduced, the proposed EDFPN framework remains computationally efficient. The edge-aware module introduces negligible computational overhead because it utilizes fixed Sobel filters without trainable parameters The deformable convolution operations improve feature adaptability while maintaining practical inference speed suitable for real-time security screening systems.

3.5 Conclusion

This chapter presented the proposed Edge-Aware Deformable Feature Pyramid Network (EDFPN) for prohibited item detection in heavily occluded X-ray imagery.The proposed framework integrates edge-aware feature enhancement, deformable convolution-based geometric refinement, and multi-scale feature pyramid fusion for improving structural representation and localization accuracy. Experimental results on the OPIXray dataset demonstrate that the proposed EDFPN framework achieves superior performance compared to several existing state-of-the-art methods.

The next chapter presents the proposed Frequency–Spatial Guided Deformable Fusion framework for further improving prohibited item detection using frequency-domain representation learning and spatial-frequency interaction mechanisms.

CHAPTER 4

DEVELOPMENT OF FREQUENCY–SPATIAL GUIDED DEFORMABLE FUSION NETWORK

4.1 Introduction

Automated prohibited item detection using X-ray security imagery has become an important research problem in intelligent surveillance systems. Despite significant progress in deep learning-based object detection, identifying prohibited items in X-ray baggage scans remains highly challenging due to severe object overlap, low contrast, geometric distortion, and cluttered backgrounds.

Traditional convolution-based detectors mainly focus on spatial representations while often ignoring important frequency-domain characteristics. However, X-ray imagery contains significant structural and texture information that can be effectively represented in the frequency domain. Fine-grained patterns such as metallic edges, contours, and directional textures are often more distinguishable in frequency-space representations than in purely spatial feature maps.

Furthermore, conventional convolution operations rely on fixed receptive fields, limiting their ability to adapt to irregular object geometries and varying object orientations commonly present in baggage scans. Although transformer-based methods provide strong global contextual modeling, they introduce high computational overhead and are difficult to deploy in real-time security systems.

To address these limitations, this chapter presents a Frequency–Spatial Guided Deformable Fusion Network for prohibited item detection in X-ray imagery. The proposed framework jointly exploits frequency-domain and spatial-domain representations while integrating deformable geometric adaptation for improved multi-scale feature learning.

The proposed framework introduces three key components:

- Frequency–Spatial Guided Deformable Convolution Block (FSG-DCN)
- Balanced Channel Bottleneck (BCB)
- Tri-path Frequency–Spatial Modulation (TriFSM)

These components are integrated within a deformable depthwise separable feature pyramid network to improve structural representation, channel selectivity, and overlap-aware detection capability.

The proposed framework is evaluated on benchmark X-ray datasets including PIDray, OPIXray, and HiXray and demonstrates improved detection accuracy and real-time performance.

4.2 Proposed Methodology

This section presents the proposed Frequency-Spatial Guided Deformable Fusion Network, which enhances multi-scale features for prohibited item detection in X-ray images. The proposed method overcomes three challenges in X-ray baggage analysis: (i) poor edge contrast between metallic and organic materials, (ii) shape deformations and occlusions due to object overlap, and (iii) the inability of conventional feature pyramid networks to capture fine frequency-based cues. In general, the design of the model adopts CSPNet network structure along with three important modules, namely, Frequency-Spatial Guided Deformable Convolution Block (FSG-DCN), Balanced Channel Bottleneck (BCB), and Tri-Path Frequency-Spatial Modulation (TriFSM). The integration of these modules happens before deformable depthwise-separable feature pyramid to enhance performance.

The CSPNet backbone generates hierarchical features (C_3, C_4, and C_5) based on the input X-ray image. Each feature map undergoes further refinement through the proposed TriFSM module, which performs operations on the feature maps using the parallel FSG-DCN and BCB modules. FSG-DCN increases the importance of structure and texture-based cues in the images using fixed and learnable Gabor filters along with depthwise deformable convolutions, whereas BCB focuses on channel-based information. The two results are combined using a residual connection, which allows the model to perform multiplicative modulation and identify salient patterns in the frequency-spatial domain, particularly those containing overlapping or occluded objects. These multi-scale features are then processed using a revised FPN architecture,

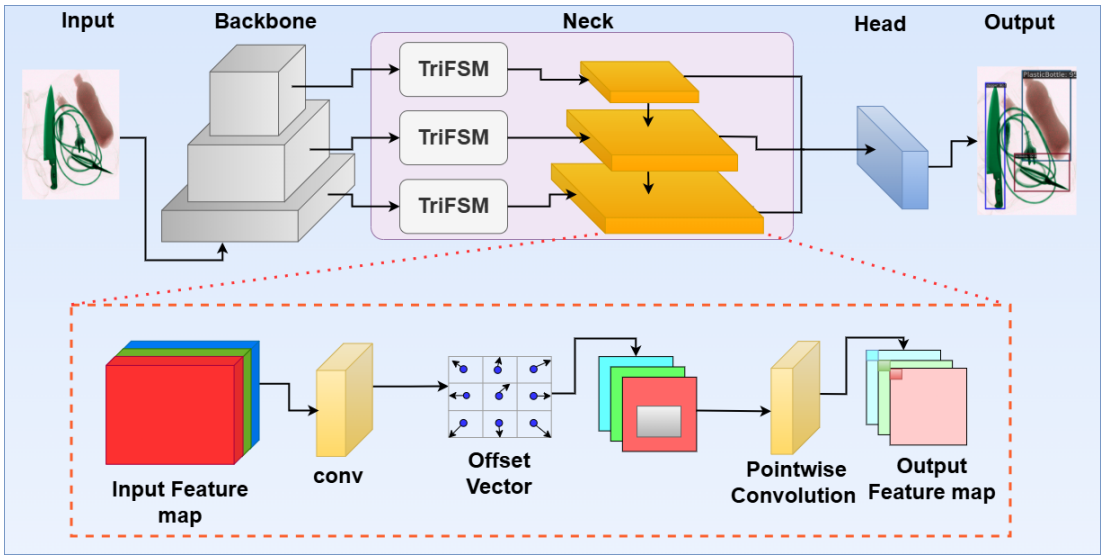


Figure 4.1: Overall architecture for detecting prohibited items.

wherein normal convolutions are replaced with depthwise separable deformable convolutions. The final set of features is then sent to the detection head for classification and localization tasks, as shown in Fig. 4.1.

4.2.1 Frequency-Spatial Guided Deformable Convolution Block (FSG-DCN)

FSG-DCN enhances feature representations by jointly modeling frequency-domain structure and spatial adaptivity, both of which are essential for robust prohibited-item detection. Since X-ray scans often contain overlapping objects, irregular shapes, and material-dependent absorption patterns, a standard convolution is insufficient to capture these complex variations. The FSG-DCN therefore integrates two complementary components: a learnable frequency-adaptive path and a fixed-prior frequency path, followed by deformable convolution to improve spatial flexibility, as given in Figure 4.2.

Let $F \in \mathbb{R}^{C\times H\times W}$ denote the input feature. The first transformation extracts frequency-aware features using a learnable Gabor kernel. A two-dimensional Gabor kernel with trainable parameters is expressed as:

$$G_{\theta,\sigma,\lambda,\psi}(x,y) = \exp\left(-\frac{x_\theta^2 + y_\theta^2}{2\sigma^2}\right)\cos\left(\frac{2\pi}{\lambda}x_\theta + \psi\right), \tag{4.1}$$

where $x_\theta = x\cos\theta + y\sin\theta$,and $y_\theta = -x\sin\theta + y\cos\theta$. The parameters $(\theta, \sigma, \lambda, \psi)$ are learned during training, allowing the kernel to adapt to the frequency characteristics of metallic edges and material textures in X-ray images. Applying the generated kernel

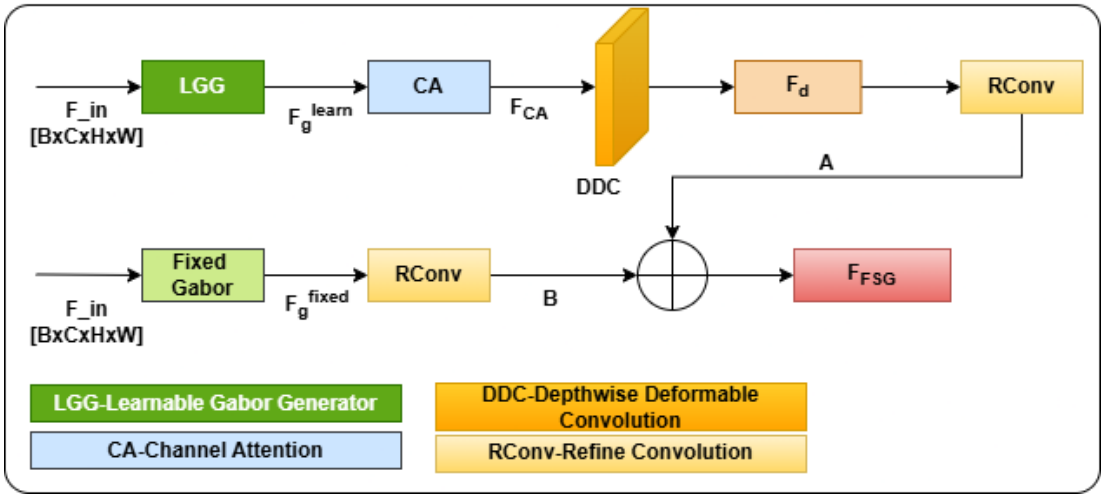


Figure 4.2: Architecture of the proposed Frequency-Spatial Guided Deformable Convolutional Network (FSGDCN).

extracts the learnable Gabor feature:

$$F_g^{\text{learn}} = G_{\theta,\sigma,\lambda,\psi} * F \tag{4.2}$$

This operation highlights directional textures and frequency components that are often indicative of prohibited metallic objects. To refine these features, a channel-attention mechanism further evaluates the importance of each frequency channel. The channel-attention operator refinement is expressed as:

$$F_{\text{ca}} = \text{CA}(F_g^{\text{learn}}) \tag{4.3}$$

It selectively enhances informative frequency channels. The refined frequency representation is processed by a depthwise deformable convolution. For a spatial location p, the deformable convolution is expressed as:

$$F_d(p) = \sum_{k=1}^{K} w_k \, F_{\text{ca}}(p + p_k + \Delta p_k) \tag{4.4}$$

where w_k denotes depthwise convolution weights, p_k is a predefined offset, and Δp_k is a learned offset. Unlike standard deformable convolution that learns offsets from spatial activations alone, the proposed method applies deformation after frequency enhancement. Consequently, the learned offsets are implicitly conditioned on orientation-sensitive and high-frequency structural cues, which improves alignment with thin metallic edges and partially occluded boundaries in overlapped X-ray images. A standard

convolution consolidates this representation:

$$A = \text{Conv}(F_d). \tag{4.5}$$

A parallel branch incorporates fixed Gabor filters, which provide stable and orientation-selective frequency priors. A predefined filter set G_{fixed} estimates:

$$F_g^{\text{fixed}} = G_{\text{fixed}} * F, \tag{4.6}$$

It is followed by a projection convolution:

$$B = \text{Conv}(F_g^{\text{fixed}}). \tag{4.7}$$

Finally, the outputs of the learnable and fixed frequency paths are concatenated as:

$$F_{\text{FSG}} = \text{Concat}(A, B). \tag{4.8}$$

In the proposed work, deformable convolution is applied after the frequency enhancement stage, allowing the learned sampling offsets to be implicitly conditioned on orientation-sensitive high-frequency cues. This design encourages the deformable kernels to adapt their receptive fields according to edge-dominant structures that frequently appear in X-ray imagery, such as thin metallic objects and partially occluded boundaries. Consequently, the deformable sampling process becomes guided by frequency-aware representations rather than operating solely on spatial feature maps.

4.2.2 Balanced Channel Bottleneck (BCB)

BCB is a Channel-analysis module that generates a stable,interpretable channel–weight distribution. As far as conventional attention mechanisms are concerned, they typically re-calibrate features by multiplying directly with weights that have been learned from the data. On the other hand, the proposed model uses a BCB to generate a softmax-normalized channel descriptor which is then used inside the TriFSM module. This formulation leaves out only the trivial variations in the feature representation, and yet it is informative.

A bottleneck pointwise convolution architecture is adopted at the BCB in order to effectively model inter-channel dependencies in an efficient fashion. To get a compact latent representation, the input feature F is first convolved with 1×1 filter whose

channel dimensionality is halved, which is followed by a dense layer:

$$F_r = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}_{1\times1}(F; C/4))) \,. \tag{4.9}$$

This reduction allows for capturing meaningful cross channel interactions in the model with limited computational burden, especially when dealing with identifying low contrast and fine-grained structures in X-ray images. This reduced feature is then reprojected back to original dimensionality by again pointwise convolving:

$$F_e = \mathrm{BN}(\mathrm{Conv}_{1\times1}(F_r; C)) \,. \tag{4.10}$$

This operation gives back the channel capacity, but does not lose the important relationship information that is contained in the compressed representation. In BCB, step-by-step softmax operation is performed along channel dimension:

$$M_{c,h,w} = \frac{\exp(F_{e,c,h,w})}{\sum_{j=1}^{C} \exp(F_{e,j,h,w})}, \qquad c = 1, \ldots, C. \tag{4.11}$$

In this form, there is a competitive interaction between the channels naturally, resulting in a distribution of the type of probability. This creates a weight vector M that defines the spatial variation of the relevance of the channel, enabling that network to give different weights to different channels at each pixel. This can be helpful when analysing X-ray images because the way in which items are prohibited and visible in them could differ in different spatial locations.

The BCB provides a robust mechanism for computing spatially distributed channel-importance maps without directly altering the original feature representation. The balanced channel distribution produced by the softmax operation enhances interpretability, prevents channel dominance, and enables the TriFSM module to perform more informed frequency–spatial fusion. This design proves advantageous for X-ray security applications, where small, overlapping, and low-contrast object signatures necessitate fine-grained spatial and channel-level discrimination.

4.2.3 Tri-Path Frequency-Spatial Modulation (TriFSM)

The Tri-Path Frequency-Spatial Modulation (TriFSM) block serves as the fusion unit by combining three complementary sources of information: frequency response, spatial adaptation, and spatially distributed channel importance. As shown in Figure 4.3, this design is particularly important for X-ray prohibited item detection because objects often appear overlapped, rotated, or partially transparent.

Given a backbone feature map F, the TriFSM module generates an enhanced output through the coordination of three parallel transformations. The first path propagates the backbone feature through the FSG-DCN:

$$F_{\text{FSG}} = \text{FSG-DCN}(F) \tag{4.12}$$

F_{FSG} incorporates frequency responses derived from both learnable and fixed Gabor filters, with spatially flexible receptive fields. Consequently, F_{FSG} encodes boundary cues, fine edges, metallic textures, and other discriminative components that are characteristic of X-ray scanned contraband.

In the second path, the backbone feature is processed by the Balanced Channel Bottleneck, which produces a spatially distributed channel-weighting map:

$$M = \text{BCB}(F). \tag{4.13}$$

Unlike conventional channel-attention mechanisms that assign global weights, the spatial softmax-driven formulation ensures that channel importance is evaluated at each pixel. This enables the network to highlight discriminative channels differently across complex spatial regions, which is crucial for identifying partially occluded or thin metallic structures. The weighting map is then used to modulate the FSG-DCN output as:

$$F_B = F_{\text{FSG}} \odot M. \tag{4.14}$$

where $\cdot$ represents element-wise multiplication. This operation helps the module to adaptively emphasize relevant frequency–spatial patterns.

The third path refines the structural alignment between the input feature and the modulated response. The backbone features are processed as:

$$F_p = \text{Conv}_{1\times1}(F), \tag{4.15}$$

It projects the feature into a refined channel space that is better aligned with the frequency-spatial representation. The projected feature is then multiplied with the modulated representation from the first two paths:

$$F_{\text{mod}} = F_p \odot F_B. \tag{4.16}$$

To ensure that the enhanced representation preserves the semantic information of

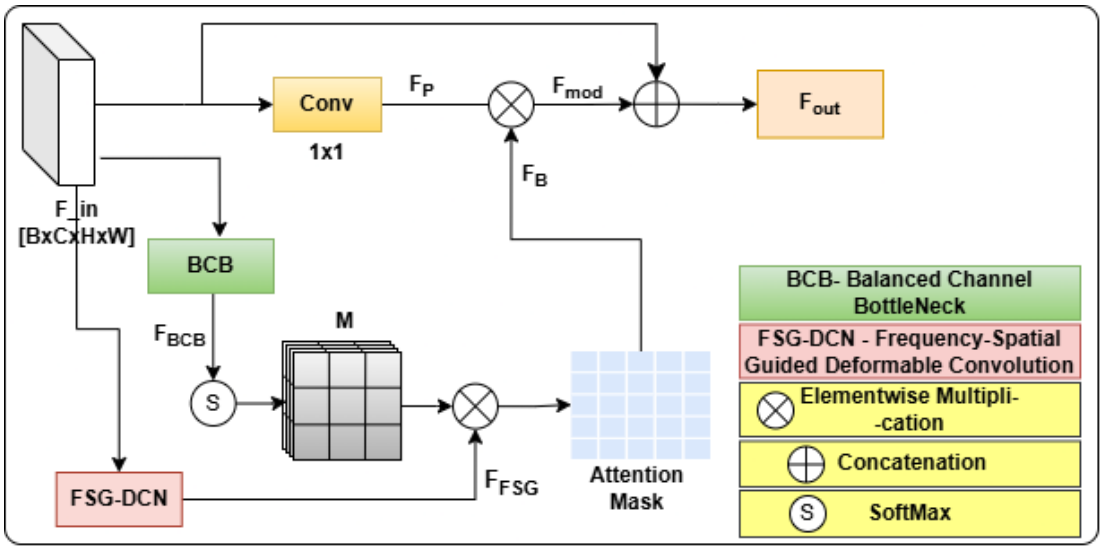


Figure 4.3: Structure of the TriFSM module.

the backbone feature, a residual connection is employed:

$$F_{\text{out}} = F + F_{\text{mod}}. \tag{4.17}$$

This residual formulation stabilizes training and maintains the integrity of high-level semantic cues while injecting enriched frequency-spatial information.

TriFSM does not replace feature pyramid fusion; instead, it refines backbone features before pyramid construction. In particular, the module operates on frequency-enhanced representations, allowing spatial attention and channel interactions to be conditioned on orientation-sensitive high-frequency cues. The multiplicative interaction differs from the additive aggregation typically used in FPNs: it selectively reinforces feature responses that are jointly supported by frequency-enhanced structures and spatially distributed channel relevance, while suppressing inconsistent activations arising from overlapping materials and clutter. The residual connection ensures semantic stability and maintains compatibility with subsequent DDFPN aggregation.

4.2.4 Depthwise-Deformable Feature Pyramid Network (DDFPN)

The DDFPN works as a multi-scale fusion and pyramid processing mechanism within the proposed framework. Traditional feature pyramids predominantly rely on fixed-grid convolutions and, consequently, struggle to adapt to severe geometric distortions, irregular orientations, and heavy occlusion. To address this limitation, DDFPN integrates deformable sampling with computationally efficient depthwise separable convolution, enabling the network to model both geometric variability and fine structural

cues with minimal overhead.

Let $\hat{F}_3$, $\hat{F}_4$, and $\hat{F}_5$ be refined TriFSM feature maps. The goal of the DDFPN is to produce a set of enhanced multi-scale features (P_3, P_4, P_5) that are a deformable and spatially aligned version of $\hat{F}_i$ enriched with semantic information from higher levels. Given a feature map $\hat{F}_i$, deformable depthwise convolution is applied to increase geometric flexibility by learning spatial offsets for each sampling location. Formally, for a spatial position p, the operation is expressed as

$$F_i^{\text{dd}}(p) = \sum_{k=1}^{K} w_k \, \hat{F}_i\big(p + p_k + \Delta p_k\big), \tag{4.18}$$

where w_k denotes the depthwise convolution weight, p_k represents the predefined kernel offset, and Δp_k is the learnable deformation offset. A subsequent 1×1 pointwise convolution captures cross-channel interactions in a computationally efficient manner. This formulation enables the convolutional kernel to adapt to distorted object boundaries and irregular spatial arrangements, which is essential for robust X-ray feature extraction.

The DDFPN follows a top-down pathway that progressively propagates high-level semantic information to lower-resolution features. Starting with the highest-level feature:

$$P_5 = \text{DDConv}(\hat{F}_5). \tag{4.19}$$

For the remaining levels, the fusion is carried out by upsampling the previous pyramid output and aligning it with the corresponding feature using deformable depthwise convolution:

$$U_i = \text{Upsample}(P_{i+1}), \tag{4.20}$$

$$\tilde{F}_i = \text{DDConv}(\hat{F}_i), \tag{4.21}$$

$$P_i = \tilde{F}_i + U_i, \tag{4.22}$$

for $i \in \{3, 4\}$. This process ensures that each lower-level feature receives semantically rich information from higher levels while simultaneously benefiting from localized spatial adaptation enabled by deformable sampling. The additive fusion also preserves the structural details inherent in each resolution level.

The DDFPN provides several advantages for X-ray images. First, the deformable offsets enable the feature pyramid to adapt to irregular and overlapped shapes. Second, the depthwise separable design significantly reduces computational cost while retaining representational capacity. Third, top-down fusion enhances the consistency of

Table 4.1: Benchmark datasets of prohibited item detection.

Dataset	Train/Test	Classes	Class Names
OPIXray[8]	7,019 / 1,776	5	"Folding Knife, Straight Knife, Scissor, Utility Knife, Multi-tool Knife"
HiXray[11]	36,295 / 9,069	8	"Portable Charger 1, Portable Charger 2, Water, Laptop, Mobile Phone, Tablet, Cosmetic, Nonmetallic Lighter"
PIDray[10]	29,457 / 18,220	12	"Gun, Knife, Wrench, Pliers, Scissors, Hammer, Handcuffs, Baton, Sprayer, Power Bank, Lighter, Bullet"

semantic cues across scales. Consequently, the DDFPN forms an essential component of the proposed architecture by delivering spatially flexible and multi-scale enriched features to the detection head.

4.3 Experimental Setup

4.3.1 Datasets

The proposed method was evaluated on three benchmark X-ray datasets [10], [8], [11] commonly used for prohibited-item detection. These datasets simulate real-world scenarios in public security settings such as airports and subways, where items are often occluded, overlapped, or concealed within baggage. A consolidated summary of dataset statistics is provided in Table 4.1.

The PIDray[10] dataset is one of the largest and most diverse datasets for prohibited item detection in real-world X-ray images. It contains 47,677 annotated images of 12 categories, including guns, knives, hammers, and electronic devices such as batteries. The dataset includes various forms of occlusion and cluttered backgrounds and provides clear instance-level annotations for each object.

The OPIXray[8] dataset is designed specifically for object detection under occlusion, especially for knives, which are commonly carried and easily hidden. It consists of five different kinds of bladed items: utility knives, multi-tool knives, scissors, folding knives, and straight knives. To improve the robustness of detection algorithms, the dataset offers 1776 testing images and 7109 training photos with different levels of background and occlusion complexity.

Lastly, the HIXray dataset [11] introduces a hierarchical structure for item detec-

tion. It contains annotations not only at the object level but also at the part level, enabling more fine-grained detection and classification. HIXray supports multi-level labeling for complex composite items and includes a wide range of daily-use and potentially dangerous objects.

4.3.2 Implementation Details

Table 4.2: Training hyperparameters used for proposed method.

Parameter	**Value**
Input resolution	640×640
Batch size	16
Optimizer	Stochastic Gradient Descent
Base learning rate	0.01
LR schedule	Quadratic warm-up (5 epochs) + Cosine annealing (up to 45 epochs) + Constant (final 20 epochs)
Total epochs	70
Loss functions	Cross-Entropy (cls/obj), IoU loss (bbox), L1 loss (refinement)
Data augmentation	Mosaic, MixUp, RandomAffine, HSV adjustment, RandomFlip

4.4 Experimental Results

4.4.1 Quantitative Analysis

Comparison on OPIXray

Table 4.3 shows the results on the OPIXray dataset using the VOC metric. OPIXray dataset is particularly challenging as it focuses on five knife-related categories under various degrees of occlusion and clutter.

From Table 4.3, it can be observed that the proposed method achieves an mAP of 86.8%, which surpasses both classical detectors and transformer-based approaches. Compared with YOLOX-s (83.6% mAP), the proposed method obtains an absolute gain of 3.2 percentage points. The performance margin of proposed method compared to AO-DETR(S) [23] (80.8% mAP) and DINO [20] (78.2% mAP) is large. YOLOv8 [46] achieves 84.1% mAP, while the proposed method further improves it to 86.8% with significantly fewer parameters and FLOPs. Specifically, YOLOv8 requires 11.2M parameters and 28.6 GFLOPs, whereas the proposed network uses only 10.0M

Table 4.3: Comparison with SOTA detectors on OPIXray [8]. FK, SK, Sc, UK, and MK denote folding knife, straight knife, scissor, utility knife, and multitool knife, respectively. A dash indicates unavailable data.

Models	Ref.	Parameter (M)	GFLOPs	mAP	FK	SK	Sc	UK	MK
Faster R-CNN	[12]	41.19	20.36	70.9	75.1	45.8	88.4	67.7	77.3
RetinaNet	[28]	38	239	70.9	77.1	36.1	94.6	64.6	82.2
Mask R-CNN	[27]	41.42	20.36	74.7	77.9	51.4	89.5	69.4	85.5
Cascade R-CNN	[15]	68.97	22.37	72.8	75.7	50.0	89.4	70.0	79.0
ATSS	[18]	51.14	27.82	67.5	72.8	38.0	88.6	58.0	80.2
GFL	[19]	50.70	28.51	75.6	80.0	53.6	89.3	71.7	83.4
Sparse R-CNN	[26]	106	278	81.8	87.7	64.3	89.3	79.9	87.9
Deformable DETR	[16]	52.14	13.47	63.4	70.1	29.0	86.0	55.7	76.4
DINO	[20]	58.38	26.89	78.2	83.2	58.8	89.4	72.7	86.7
DOAM	[9]	90.79	-	82.4	86.7	68.5	90.2	78.8	87.6
MCIA-FPN	[21]	-	-	82.5	89.0	74.4	89.9	86.1	89.7
FDTNet	[22]	66.17	-	82.0	87.9	60.2	96.1	78.9	87.1
AO-DETR(R)	[23]	58.38	26.89	79.2	83.8	60.5	90.1	74.7	87.1
AO-DETR(S)	[23]	229.0	156	80.8	84.8	63.0	90.1	77.7	88.4
RT-DETR(S)	[42]	42.81	17.07	61.8	61.1	26.0	88.6	56.4	76.8
GADet-S	[38]	8.94	3.33	69.6	72.6	43.6	86.6	67.5	77.5
GADet-L	[38]	54.16	19.4	77.7	81.8	54.0	89.8	75.7	85.2
GADet-X	[38]	99.01	35.2	78.1	83.1	56.3	89.8	75.7	85.5
Mix-Paste + LLS	[39]	-	-	83.7	-	-	-	-	-
DML-Net	[40]	36.20	-	52.7	38.1	42.7	54.7	38.9	35.9
YOLOv5-s	[44]	7.2	16.5	83.6	88.4	69.9	90.4	80.1	89.1
YOLOv6-s	[45]	18.5	45.3	84.9	88.8	76.6	90.4	80.8	88.2
YOLOv7-tiny	[36]	6.2	13.8	77.9	88.1	51.1	89.6	73.2	87.2
YOLOv8-s	[46]	11.2	28.6	84.1	87.3	73.5	90.3	80.7	88.5
YOLOX-s	[13]	9.0	26.8	83.6	86.3	73.5	89.7	80.1	88.6
Proposed Model	-	**10.0**	**21.1**	**86.8**	**89.3**	**77.4**	**90.5**	**87.2**	**89.6**

parameters and 21.1 GFLOPs. This shows that the proposed method outperforms existing methods without incurring a high computational cost, making the proposed method more suitable for real-time systems.

The proposed method consistently achieves the highest or near-highest AP across all categories. For folding knife (FK), the proposed method obtains 89.3% AP, which is slightly higher than MCIA-FPN [21] (89.0%) and DOAM [9] (86.7%). For straight knife (SK) category, the proposed method achieves 77.4%, which is comparably higher than DOAM (68.5%) and MCIA-FPN (74.4%). Similarly, for utility knife (UK) and multitool knife (MK), the proposed model achieves 87.2% and 89.6% AP, respectively, outperforming DOAM (78.8% and 87.6%) and FDTNet [22] (78.9% and 87.1%). These improvements clearly indicate that the proposed method is particularly effec-

tive in distinguishing blade-like structures under severe occlusion and clutter.

Table 4.4: Evaluation of the proposed framework on the HiXray [11] dataset using VOC metrics.

Model	**mAP**
Faster RCNN [12]	72.6
Cascade RCNN [15]	72.4
FCOS [14]	75.7
GFL [19]	70.3
DOAM [9]	76.2
LIM [11]	77.3
AO-DETR [23]	75.8
MSANet [43]	71.5
YOLOX [13]	77.4
Ours	**78.5**

Comparison on HiXray

The results on the HiXray dataset are given in Table 4.4. This dataset contains eight categories, many of which correspond to benign daily objects (e.g., laptops, tablets, mobile phones, cosmetic items) that are still relevant in security screening contexts. The proposed method achieves mAP of 78.5%, outperforming Faster R-CNN (72.6%), Cascade R-CNN (72.4%), and FCOS (75.7%). Compared to the X-ray specific DOAM [9] (76.2%) and the LIM framework [11] (77.3%), the proposed method yields improvements of 2.3% and 1.2% percentage points, respectively. YOLOX achieves 77.4%, and the proposed method further improves it by 1.1 percentage points. Although the absolute gain on HiXray is relatively modest compared to OPIXray, it is important to note that HiXray includes a mixture of metallic and non-metallic objects with less extreme occlusion patterns. In such a scenario, many modern detectors already perform reasonably well, and further improvements are non-trivial. The consistent gain indicates that the proposed method not only handles highly occluded knife-like objects but also generalize to a broader set of everyday items, confirming the robustness and versatility of the proposed design.

Comparison on PIDray and occlusion-level

Table 4.5 presents the results on PIDray dataset using COCO-metric, in three difficulty levels (Easy, Hard, Hidden) and overall performance. This dataset is closer to real-world baggage inspection in terms of scene complexity and occlusion, and is an important benchmark for evaluating practical deployment feasibility.

Table 4.5: Evaluation of the proposed framework on the PIDray [10] dataset using COCO metrics.

Model	Easy	Hard	Hidden	Overall
FCOS [14]	61.8	51.7	37.5	50.3
RetinaNet [28]	61.8	52.2	40.6	51.5
Faster RCNN [12]	63.3	57.2	42.1	54.2
Libra RCNN [35]	64.7	58.8	42.9	55.5
Mask RCNN[27]	64.7	59.0	43.8	55.8
SSD [33]	68.1	58.9	45.7	57.6
Cascade RCNN [15]	69.3	62.8	48.0	60.0
Cascade Mask RCNN [15]	70.9	64.0	48.0	61.0
Deformable DETR [16]	72.1	65.7	46.2	61.3
SDANet [10]	71.2	64.2	49.5	61.6
YOLOX-s [13]	70.2	63.3	50.5	61.3
Ours	**72.9**	**69.6**	**52.6**	**65.0**

Faster R-CNN, RetinaNet, and FCOS achieve overall mAP scores in the range of 50-55%. Cascade R-CNN and CascadeMask R-CNN further improve the overall mAP to 60.0% and 61.0%, respectively. SDANet [10], which is specifically designed for the PIDray dataset, achieved an overall mAP of 61.6%. YOLOX-s achieves 61.3%, with 70.2%, 63.3%, and 50.5% mAP in the Easy, Hard, and Hidden subsets, respectively. In contrast, the proposed method achieves an overall mAP of 65.0%, representing an absolute improvement of 3.7% over YOLOX-s and 3.4% over SDANet. Analysing the difficulty-wise performance, the results show that the proposed method improves the mAP by 2.7%, and 6.3% on easy, hard and hidden subsets, respectively. On the most challenging Hidden subset, which corresponds to heavily occluded items, the proposed method improves mAP from 50.5% to 52.6%.

The relatively larger improvement on the Hard subset is particularly noteworthy. It indicates that the proposed method is effective in handling overlapping objects and complex baggage layouts. Even in the Hidden subset, where only small portions of prohibited items may be visible, the model benefits from the enhanced edge and texture representations produced by FSG-DCN and the spatially adaptive modulation of TriFSM. Overall, the results on PIDray clearly demonstrate superiority of the proposed method.

4.4.2 Qualitative Analysis

Figure 4.4 shows the visual comparison of the proposed method with existing object detection methods. All results used for visual comparison are inferred from the mod-



Figure 4.4: Visual results of prohibited item detection on PIDRAY dataset [10]. Row 1: Ground truth boxes; Row 2: Prediction using ATSS; Row 3: Prediction using YOLOX; and Row 4: Prediction using proposed method.

els trained on PIDray dataset. It can be observed from Figure 4.4 that the proposed method detects majority of objects with few false positives. However, the results of ATSS and YOLOX shows various false negatives which is not suitable for security inspections. In first column, only the proposed method detects the overlapped object whereas ATSS and YOLOX failed in detecting overlapped objects. Similarly, the results of the proposed method show the detection of majority of object where ATSS and YOLOX failed for few objects from column 2-4. In last column of Figure 4.4, it can be observed that YOLOX and ATSS failed in detecting the object which is wrapped with wire to intentionally hide the object. The proposed method successfully detected the intentionally hidden object more effectively than existing methods. It shows the superiority of the proposed method as compared to the existing methods.

4.4.3 Ablation Study

To evaluate the contribution of each proposed component, the study performed various ablation experiments on the OPIXray dataset. The contribution of each component using different network configurations. All experiments were performed under identical

training conditions to ensure consistency and fairness in comparison.

Table 4.6: Impact of each module of proposed method on the OPIXray dataset.

Configuration	**mAP@0.5**
Baseline [13]	83.6
with DDFPN	85.6
with BCB + DDFPN	85.9
with FSG-DCN + DDFPN	86.1
with TriFSM (without DDFPN)	85.5
Ours	**86.8**

Impact of Individual Components

Table 4.6 summarizes the effect of each module. The baseline model [13] achieves an mAP@0.5 of 83.6% on the OPIXray dataset. Incorporating DDFPN improves performance by +2.0%, demonstrating the benefit of deformable multi-scale aggregation for handling irregular object boundaries and overlapping structures commonly observed in X-ray images. Adding BCB with DDFPN further improves performance to 85.9%, indicating that softmax weighting enables more stable and discriminative feature utilization by preventing disproportionate activation across channels. When the FSG-DCN module is integrated with DDFPN, performance increases to 86.1%, underscoring the importance of joint frequency–spatial modeling. The fixed and learnable Gabor representations, together with deformable convolution, enhance the model's ability to capture orientation-specific structures and fine-grained metallic textures. The TriFSM module, when evaluated independently without DDFPN, achieves 85.5%, confirming its capacity to modulate frequency-enriched and spatially adaptive features effectively. However, its benefit is maximized when combined with the deformable feature pyramid. The proposed method with all modules achieves the highest mAP@0.5 of 86.8%, validating the contribution of each proposed component and the overall design of the framework.

Table 4.7: Impact of fixed vs. learnable Gabor filters on frequency response extraction.

Gabor Configuration	**mAP %**
Only Fixed Gabor Filters	86.1
Only Learnable Gabor Generator	86.2
Fixed + Learnable (Proposed FSG-DCN)	**86.8**

Effect of Learnable and Fixed Gabor features

The effect of Gabor-based frequency extraction in the FSG-DCN module is evaluated in Table 4.7. The use of fixed Gabor filters only in FSG-DCN achieves an mAP of 86.1%, which shows that it effectively highlights contours and directional gradients characteristic of prohibited items. The use of only learnable Gabor filters achieves 86.2%, which shows that data-driven frequency adaptation provides additional flexibility for intra-class shape variations and imaging artifacts. The combination of fixed and learnable Gabor filters, as adopted in the proposed FSG-DCN, achieves the best performance of 86.8%. The proposed FSG-DCN produces richer frequency-spatial representations, leading to superior performance.

4.4.4 Run-time Analysis

To evaluate computational efficiency, runtime analysis was conducted on an NVIDIA RTX A5000 GPU with an input resolution of 640×640. Table 4.8 shows that Faster R-CNN and RetinaNet require significantly higher computational overhead. Deformable DETR provides better spatial modeling but remains computationally intensive with 173 GFLOPs. In contrast, the YOLOX-S model provides substantially higher efficiency, achieving 219 FPS. The proposed model maintains real-time performance despite introducing additional frequency–spatial processing and deformable multi-scale fusion. Specifically, the proposed framework achieves 220 FPS with 10 M parameters and 21.1 GFLOPs. This balance between runtime and detection performance demonstrates that the integration of FSG-DCN, BCB, and TriFSM does not impose prohibitive computational cost. The proposed model achieves state-of-the-art detection performance while retaining the computational characteristics required for real-time X-ray inspection systems. As shown in Table 4.9, the proposed model achieves 4.5 ms latency per image ($\approx$ 220.4 FPS) with batch size 1 using PyTorch FP32 on a single GPU, demonstrating real-time inference capability. The throughput is computed as (Throughput $=$ FPS $\times$ Batch Size), reaching up to 313.6 images/s for batch size 16 while maintaining moderate GPU memory usage. These results indicate that the proposed method is suitable for practical deployment, and further acceleration can be achieved using FP16 or TensorRT optimization.

4.5 Conclusion

This study designed a frequency–spatial guided deformable fusion network to detect prohibited items in the X-ray image. The proposed model includes three specialized

Table 4.8: Run-time performance comparison of different methods with input resolution of 640×640.

Model	Params (M)	GFLOPs	FPS
Faster R-CNN	42	180	27
RetinaNet	34.0	151.5	28
Deformable DETR	40	173	19
YOLOv5-S	7.2	16.5	86
YOLOX-S	9.0	26.8	219
Ours	**10.0**	**21.1**	**220.4**

Table 4.9: Deployment-oriented inference performance using PyTorch FP32 on a single GPU.

Batch	FPS	Time/Batch (ms)	Latency (ms/img)	Throughput (img/s)	Mem (GB)
1	220.4	4.5	4.5	220.4	7.56
8	33.8	29.6	3.7	270.4	6.29
16	19.6	51.6	3.2	313.6	7.87

modules, namely FSG-DCN, BCB, and TriFSM along with the deformable depthwise feature pyramid, which contributes to multi-scale discrimination. By integrating both fixed and learned Gabor filters and adaptive spatial convolution, the network can efficiently capture fine structural details, metallic texture, and boundaries of objects under the condition of overlapping and occlusion. Experiments on three public datasets reveal that the proposed framework can achieve state-of-the-art results without compromising inference efficiency.

As future research directions, global contextual reasoning and multi-view X-ray information will be incorporated into the proposed model.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

In recent years, increasing security threats and public safety concerns have created a strong need for intelligent automated baggage screening systems. X-ray security imaging systems are widely deployed at airports, railway stations, metro stations, border checkpoints, and government facilities to identify dangerous and prohibited objects concealed inside baggage. However, manual inspection of X-ray images by security personnel is difficult and time-consuming because X-ray scans often contain severe object overlap, cluttered backgrounds, low-contrast structures, and partial occlusion. These challenges can lead to missed detections and false alarms, thereby reducing the reliability and efficiency of security screening systems.

While deep learning based object detection systems have demonstrated impressive success in analyzing natural images, their ability to analyze X-ray imagery has still been constrained by the nature of an X-ray scan. In general, traditional CNN-based approaches tend to only process spatial characteristics, and struggle to extract fine boundaries, geometry variations and texture-related features from X-ray images. Furthermore, the use of sampling locations fixed in the two domains limit the adaptability of the traditional feature extraction mechanisms in the presence of severe overlap and irregular arrangement of the objects.

The thesis presented two advanced deep learning approaches for detecting prohibited items in an X-ray image to overcome these limitations. The first framework introduced an Edge-Aware Deformable Feature Pyramid Network (EDFPN) to enhance the features of the image using the edges and also to deform the feature convolutional network to capture image geometry changes under strong occlusions, for the detection of prohibited items. The second model put forward a Frequency–Spatial Guided Deformable Fusion Network, which combines the frequency domain representation

learning with spatial feature modeling, to extract structural features and achieve multi-scale feature fusion.

The model is proposed in EDFPN, is based mainly on the improvement of the contour-sensitive feature representation and adaptive geometric modeling. New Edge-Aware Feature Operator (EFO) has been added to boost boundary sensitive information by applying gradient based edge extraction. The edge-aware mechanism assisted in achieving a more accurate portrayal of partially obscured prohibited items and contour localization. Additionally, deformable convolution was integrated into feature refinement end to enhance adaptive spatial sampling and the geometric alignment. Conventional convolutions sample fixed positions, but deformable convolutions optimally sample positions based on object geometry to enhance localization of objects with high overlap and with non-uniform orientations.

The effectiveness of the proposed EDFPN framework was evaluated experimentally, and significant enhancement in the performance of prohibition item detection was shown on the OPIXray dataset. The framework was able to successfully identify partially visible and heavily overlapping forbidden objects with efficient computational complexity on the basis that it is realistically implementable. Overall, the proposed edge-aware enhancement and deformable convolution significantly enhanced the localization precision and overlap resolution performance over existing CNN based detectors.

The second major contribution of this thesis was the development of the Frequency–Spatial Guided Deformable Fusion Network. Existing object detection frameworks generally rely only on spatial-domain feature extraction and ignore important frequency-domain characteristics such as directional textures, edge structures, and contour-sensitive information. However, X-ray imagery contains significant frequency-domain information that can improve structural representation learning under low-contrast and cluttered conditions.

To exploit this information, the proposed framework introduced a Frequency–Spatial Guided Deformable Convolution (FSG-DCN) module that integrates learnable Gabor filters with deformable convolution. The proposed frequency-aware representation learning mechanism effectively captures directional texture and contour patterns useful for prohibited item detection. Additionally, a Balanced Channel Bottleneck (BCB) module was proposed to adaptively emphasize informative feature channels and suppress redundant feature responses. A Tri-path Frequency–Spatial Modulation (TriFSM) mechanism was further introduced to improve interaction between frequency-enhanced features and spatial representations.

It also introduces deformable depthwise separable feature pyramid fusion for effi-

cient multi-scale feature aggregation into the proposed frequency–spatial framework. This approach proved to be more effective in detecting small-sized prohibited objects with low computational complexity, it is suitable for real-time systems.

Numerous experiments were performed on standard database setups such as PIDray, OPIXray and HiXray. The experimental results showed that the proposed frameworks were always superior to other representative state-of-the-art object detection techniques when dealing with the challenging overlap and occlusion conditions. The proposed methods provided higher localization accuracy, more localized contour representation, better overlap handling and better detection of small prohibited items.

The experimental evaluation also showed that the representation learning with the edge awareness module greatly improved the boundary-sensitive localization in the experimental settings, and the 2-D SGP leveraged the frequency domain feature extraction to obtain good structural and texture representation. Deformable convolution was used to enhance the adaptability of geometric modeling, and robust feature extraction was achieved in the case of irregular arrangement of the object. The multi-scale fusion of features found to be effective in terms of detection performance at different scales of objects.

There was also an important finding in the experimental outcomes – the proposed schemes remain practically computationally efficient despite integrating powerful feature enhancement operations. New lightweight edge extraction, depthwise separable convolution and efficient deformable feature refinement allowed for inference in real-time, enabling practical security screening systems.

In conclusion, the research efforts in this thesis have achieved a number of goals which provide some improvements to the existing prohibited item detection systems. Edge-aware learning combined with frequency-domain representation, deformable convolution, and adaptive multi-scale fusion enabled the provision of strong and efficient solutions for intelligent X-ray prohibited item detection. The suggested frameworks help realize dependable automatic baggage check systems that may support man-in-the-loop security situations.

5.2 Future Scope

The proposed frameworks proved to have successfully enhanced the ability to detect banned items, however, there are still several research areas to be explored for further improvement.

Future systems can now be designed to incorporate multi-view X-ray imaging. In the current study, considered are two dimensional images from X-ray photography,

in which the object is often completely covered by another object, thus geometric information is lost. Multi-view or dual-view X-ray analysis can give other geometric clues, and enhance the three-dimensional understanding of contents of the bags.

Another direction that can be taken for better item detection in a scene with a lot of clutter is hybrid CNN-transformer architectures. A transformer model can be used to get a good global contextual model, and a convolutional network can be employed to extract local spatial features well. By combining both methods, further improvement of the detection performance under complex screening conditions may be achieved.

Self-supervised or unsupervised representation learning can also be explored in future research so as to free up on the requirement of large annotated X-ray baggage datasets. Manual annotation of X-ray images is costly and labor-intensive as it needs the expertise. Self-supervised Learning Strategies (SSL) can leverage on unlabeled data for better feature representation and better generalizability to diverse security settings.

Prohibited item detection systems can additionally be enhanced by implementing explainable AI (XAI) techniques to make the systems more interpretable and trustworthy. Explainable frameworks can also assist security in comprehending detection choices and boost their trust in automatic screening systems.

Domain adaptation methods can be explored further to make it more robust to change the imaging condition and X-ray scanner. Different scanners, types of bags and acquisition environments can cause domain shifts that can limit the generalizability of the model. Deploying in a cross domain scenario can be aided with the use of domain adaptation frameworks for enhancing cross domain robustness and scalability.

Another key direction for lightweight modeling is also being pursued. The proposed frameworks provide sufficient computational efficiency while pruning, quantization and knowledge distillation methods can be used to facilitate deployment on embedded devices and edge devices with constrained computational resources.

Moreover, hyperspectral imaging, millimeter-wave imaging, and thermal imaging could also be fused with X-ray analysis to enhance the detection of banned items in very difficult scenarios.

Finally, three dimensional baggage reconstruction and volumetric object reasoning can be investigated to enhance the perception of the structure of baggage scenes with high density. These could improve the accuracy of locating prohibited items and the resolution for overlapping items for next-generation intelligent security item screening systems.

BIBLIOGRAPHY

[1] D. Turcsany, A. Mouton, and T. Breckon, "Improving Feature-based Object Recognition for X-ray Baggage Security Screening using Primed Visual Words," in *Proc. IEEE Int. Conf. Industrial Technology*, 2013, pp. 1140–1145.

[2] D. Mery, E. Svec, and M. Arias, "Object recognition in baggage inspection using adaptive sparse representations of X-ray images," in *Pacific-Rim Symposium on Image and Video Technology*, LNCS vol. 9431, Springer, Cham, 2016, pp. 709–720.

[3] S. Akcay, M. E. Kundegorski, C. G. Willcocks, and T. P. Breckon, "Using Deep Convolutional Neural Network Architectures for Object Classification and Detection Within X-Ray Baggage Security Imagery," *IEEE Transactions on Information Forensics and Security*, vol. 13, pp. 2203–2215, 2018.

[4] R. L. Galvez, E. P. Dadios, A. A. Bandala, and R. R. P. Vicerra, "YOLO-Based Threat Object Detection in X-Ray Images," in *Proc. IEEE Int. Conf. Humanoid, NanoTech., IT, Communication and Control, Environment, and Management*, Laoag, Philippines, 2019, pp. 1–5.

[5] N. Bhowmik, Y. F. A. Gaus, and T. P. Breckon, "On the Impact of Using X-ray Energy Response Imagery for Object Detection via Convolutional Neural Networks," in *Proc. IEEE Int. Conf. Image Processing*, Anchorage, AK, USA, 2021, pp. 1224–1228.

[6] K. Chaturvedi, A. Braytee, D. K. Vishwakarma, M. Saqib, D. Mery, and M. Prasad, "Automated Threat Objects Detection with Synthetic Data for Real-Time X-ray Baggage Inspection," in *Proc. Int. Joint Conf. Neural Networks*, Shenzhen, China, 2021, pp. 1–8.

[7] T. Hassan, M. Bettayeb, S. Akcay, S. Khan, M. Bennamoun, and N. Werghi, "Detecting Prohibited Items in X-Ray Images: A Contour Proposal Learning

Approach," in *Proc. IEEE Int. Conf. Image Processing*, Abu Dhabi, UAE, 2020, pp. 2016–2020.

[8] Y. Wei, R. Tao, Z. Wu, Y. Ma, L. Zhang, and X. Liu, "Occluded Prohibited Items Detection: An X-ray Security Inspection Benchmark and De-occlusion Attention Module," in *Proc. 28th ACM Int. Conf. Multimedia*, New York, NY, USA, 2020, pp. 138–146.

[9] C. Zhao, L. Zhu, S. Dou, W. Deng, and L. Wang, "Detecting Overlapped Objects in X-Ray Security Imagery by a Label-Aware Mechanism," *IEEE Trans. on Information Forensics and Security*, vol. 17, pp. 998–1009, 2022.

[10] B. Wang, L. Zhang, L. Wen, X. Liu, and Y. Wu, "Towards Real-World Prohibited Item Detection: A Large-Scale X-ray Benchmark," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021, pp. 5412–5421.

[11] R. Tao, Y. Wei, X. Jiang, H. Li, H. Qin, J. Wang, Y. Ma, L. Zhang, and X. Liu, "Towards Real-world X-ray Security Inspection: A High-quality Benchmark and Lateral Inhibition Module for Prohibited Items Detection," in *Proc. IEEE Int. Conf. Computer Vision*, 2021.

[12] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[13] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO Series in 2021," *ArXiv*, vol. abs/2107.08430, 2021.

[14] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully Convolutional One-Stage Object Detection," in *Proc. IEEE Int. Conf. Computer Vision*, 2019, pp. 9627–9636.

[15] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into High Quality Object Detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6154–6162.

[16] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

[17] C. Ma, L. Zhuo, J. Li, Y. Zhang, and J. Zhang, "Occluded prohibited object detection in X-ray images with global context-aware multi-scale feature aggregation," *Neurocomputing*, vol. 519, pp. 1–16, Jan. 2023.

[18] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 9759–9768.

[19] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yan, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, Dec. 2020, pp. 21002–21012.

[20] H. Zhang, F. Chen, C. Zhang, X. Huang, L. Wang, Y. Yu, and H. Ouyang, "DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection," arXiv preprint arXiv:2203.03605, 2022.

[21] M. Wang, H. Du, W. Mei, S. Wang, and D. Yuan, "Material-aware cross-channel interaction attention (MCIA) for occluded prohibited item detection," *Visual Computer*, vol. 39, no. 7, pp. 2865–2877, Jul. 2023.

[22] Z. Zhu, Y. Zhu, H. Wang, N. Wang, J. Ye, and X. Ling, "FDTNet: Enhancing frequency-aware representation for prohibited object detection from X-ray images via dual-stream transformers," *Engineering Applications of Artificial intelligence*, vol. 133, Jul. 2024, Art. no. 108076.

[23] M. Li, T. Jia, H. Wang, B. Ma, H. Lu, S. Lin, D. Cai, and D. Chen, "AO-DETR: Anti-Overlapping DETR for X-Ray Prohibited Items Detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 7, pp. 12076–12090, Jul. 2025.

[24] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788 (2016). https://doi.org/10.1109/CVPR.2016.91

[25] Dai, Jifeng, et al. "Deformable convolutional networks." Proceedings of the IEEE international conference on computer vision. 2017.

[26] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, and P. Luo, "Sparse R-CNN: End-to-end object detection with learnable proposals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, Virtual, USA, 2021, pp. 14449–14458.

[27] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2961–2969.

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2999–3007.

[29] F. Shao, J. Liu, P. Wu, Z. Yang, and Z. Wu, "Exploiting foreground and background separation for prohibited item detection in overlapping X-ray images," *Pattern Recognition*, vol. 122, p. 108261, 2022.

[30] J. Zhang, X. Song, J. Feng, J. Fei, and S. Jiang, "X-ray image recognition based on improved Mask R-CNN algorithm," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–14, 2021.

[31] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proceedings of International Conference on Pattern Recognition*, 2006, pp. 850–855.

[32] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS: Improving object detection with one line of code," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Los Alamitos, CA, USA, 2017, pp. 5561–5569.

[33] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," in B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, vol. 9905. Springer, Cham, 2016, pp. 21–37.

[34] T. Jia, B. Ma, H. Wang, M. Li, S. Lin, and D. Chen, "ForkNet: Overlapping Image Disentanglement for Accurate Prohibited Item Detection," *IEEE Trans. on Instrumentation and Measurement*, vol. 73, pp. 1–12, 2024.

[35] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards Balanced Learning for Object Detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 821–830.

[36] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv*, 2022.

[37] R. Wang, R. Shivanna, D. Z. Cheng, S. Jain, D. Lin, L. Hong, and E. H. Chi, "DCN-V2: Improved Deep & Cross Network and Practical Lessons for Web-Scale Learning to Rank Systems," arXiv preprint arXiv:2008.13535, 2020.

[38] M. Li, B. Ma, H. Wang, D. Chen and T. Jia, "GADet: A Geometry-Aware X-Ray Prohibited Items Detector," in IEEE Sensors Journal, vol. 24, no. 2, pp. 1665-1678, 15 Jan.15, 2024.

[39] R. Chen, Y. Yan, J. -H. Xue, Y. Lu and H. Wang, "Augmentation Matters: A Mix-Paste Method for X-Ray Prohibited Item Detection Under Noisy Annotations," in IEEE Transactions on Information Forensics and Security, vol. 20, pp. 234-248, 2025.

[40] F. Yang, R. Jiang, Y. Yan, J. -H. Xue, B. Wang and H. Wang, "Dual-Mode Learning for Multi-Dataset X-Ray Security Image Detection," in IEEE Transactions on Information Forensics and Security, vol. 19, pp. 3510-3524, 2024.

[41] Wei, Y. and Wang, H., 2024. Cooperative distillation with X-ray images classifiers for prohibited item detection. Engineering Applications of Artificial intelligence, 127, p.107276.

[42] Y. Zhao et al., "DETRs Beat YOLOs on Real-time Object Detection," 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024, pp. 16965-16974.

[43] Zhengping Wu, Peng Zhu, and Bangjun Lei. 2024. MSANet: Multi scale prohibited item detection with tripe attention. In Proceedings of International Conference on Frontiers of Artificial intelligence and Machine Learning. Association for Computing Machinery, New York, NY, USA, 27–31.

[44] Khanam, Rahima, and Muhammad Hussain. "What is YOLOv5: A deep look into the internal features of the popular object detector." arXiv preprint arXiv:2407.20892 (2024).

[45] Li, Chuyi, et al. "YOLOv6: A single-stage object detection framework for industrial applications." arXiv preprint arXiv:2209.02976 (2022).

[46] Yaseen, Muhammad. "What is YOLOv8: An In-Depth Exploration of the Internal Features of the Next-Generation Object Detector." ArXiv abs/2408.15857 (2024).