IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 May 2024, accepted 28 May 2024, date of publication 31 May 2024, date of current version 7 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3407720

RESEARCH ARTICLE

Multi-Scale Hierarchical Contour Framework for Detecting Cluttered Threats in Baggage Security

MOHAMAD ALANSARI[1,2], ABDELFATAH AHMED[1,2], KHALED ALNUAIMI[1,2,3], DIVYA VELAYUDHAN[1,2], TAIMUR HASSAN[4], SAJID JAVED[1,2,5], MOHAMMED BENNAMOUN[6], (Senior Member, IEEE), AND NAOUFEL WERGHI[1,2,5,7], (Senior Member, IEEE)

[1]Department of Electrical Engineering, College of Computing and Mathematical Sciences, Khalifa University, Abu Dhabi, United Arab Emirates
[2]Department of Computer Science, Khalifa University, Abu Dhabi, United Arab Emirates
[3]Strategy and Planning Department, Abu Dhabi Investment Authority (ADIA), Abu Dhabi, United Arab Emirates
[4]Department of Electrical, Computer and Biomedical Engineering, Abu Dhabi University, United Arab Emirates
[5]Center for Autonomous Robotic Systems, Khalifa University, Abu Dhabi, United Arab Emirates
[6]Department of Computer Science and Software Engineering, The University of Western Australia, Perth, WA 6009, Australia
[7]Center for Cyber-Physical Systems, Khalifa University, Abu Dhabi, United Arab Emirates

Corresponding author: Mohamad Alansari (100061914@ku.ac.ae)

This work was supported in part by Khalifa University under Grant CIRA-2021-052, and in part by the Advanced Technology Research Center Program (ASPIRE) under Grant AARE20-279.

ABSTRACT Ensuring the security and safety of passengers and cargo through effective baggage screening presents a critical challenge in high-traffic environments like airports, where traditional manual processes are affected by high error rates, fatigue among security personnel, and privacy concerns. These issues underscore the urgent need for sophisticated automated systems capable of accurately detecting concealed threats. Advancements in deep learning have introduced more refined approaches to baggage screening, yet they struggle with capturing the global context of threats and challenges presented by clutter, concealment, and occlusion. Addressing these issues, this paper presents the Multi-Scale Hierarchical Transformer for X-ray Detection (MHT-X), a novel framework that surpasses conventional limitations by integrating a multi-scale contour mapping technique with a hierarchical feature extraction process using visual Transformers. MHT-X is distinctive in its ability to generate multi-level contour maps that unveil complex features of threat objects, facilitating their accurate identification. It employs a top-down approach for hierarchical feature processing, capturing vital spatial characteristics essential for precise threat localization through an efficient encoder-decoder mechanism. MHT-X represents a significant advancement in the detection of concealed and obstructed threat objects in X-ray baggage scans. By combining multi-scale contour maps with a hierarchical feature extraction process, it significantly enhances detection accuracy and outperforms traditional Convolutional Neural Network (CNN) models by effectively capturing the global context of illegal items with advanced feature representation. The proposed framework is rigorously validated on two public datasets, dubbed CLCXray, and PIDray, where it outperforms the state-of-the-art (SOTA) methods by achieving the mean average precision score of 65.26%, and 78.19%, respectively.

INDEX TERMS Aviation security, baggage X-ray scans, hierarchical feature extraction, multi-scale contours, threat detection, vision transformer (ViT).

I. INTRODUCTION

X-ray imaging stands as a fundamental tool for secure inspection, particularly critical in areas with heightened security like airports, cargo facilities, and shopping centers. It is pivotal in the detection of illegal and concealed items [1]. The urgency to swiftly identify a wide range of prohibited items, ranging from weapons to restricted liquids, particularly during times of high passenger flow, significantly complicates the screening process [2]. Despite its extensive

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

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

deployment, the manual examination of baggage is both strenuous and prone to inaccuracies, with the effectiveness of security personnel often hampered by fatigue and varying levels of experience. Research has shown that the success rate of human inspectors in identifying threats lies between 80 to 90 percent, a figure that is concerning considering the potential risks involved [3]. This backdrop has sparked a growing interest in the development of automated solutions for spotting security threats in baggage through X-ray scans [4]. However, these traditional methods of baggage surveillance still heavily depend on human intervention, leading to concerns over time efficiency and error margins [5]. To improve the accuracy in detecting forbidden items in baggage, there has been a notable shift towards the adoption of deep learning techniques [6], [7]. These methods have significantly enhanced the detection process, yet they confront distinct challenges when applied to X-ray imaging. X-ray scans, in contrast to typical color (RGB) images, are characterized by their lack of distinct textures and minimal intensity differences among overlapped objects. This absence of clear visual cues in X-ray imagery poses a substantial hurdle for conventional object detection frameworks like Mask R-CNN [8], Faster R-CNN [9], and RetinaNet [10], which were initially developed with color images in mind. As a result, these models often fall short in accurately identifying contraband within cluttered and intricate X-ray scans (as evident in [11]). The unique complexities of X-ray imagery, particularly in the context of baggage screening, underscore the need for more specialized approaches beyond traditional deep learning methodologies. Therefore, our work proposes MHT-X (Multi-Scale Hierarchical Transformer for X-ray Detection), which ingeniously integrates a multi-scale contour mapping technique with a hierarchical feature extraction process, harnessing the power of Vision Transformers (ViTs) [12]. This novel combination allows MHT-X to split input scans into multi-level contour maps that reveal the complex features of threat objects, which is a critical step toward accurate contraband identification. The ability of MHT-X lies in extracting crucial features from the multi-scale contour representations, leveraging a top-down approach in the hierarchical processing of features. Through this, it adeptly captures vital spatial features necessary for the precise localization of threats. As a result, it utilizes an efficient encoder-decoder mechanism to predict bounding boxes. MHT-X is designed to exceed the constraints of Convolutional Neural Network (CNN) models and adeptly addresses the challenge of occlusions and cluttered items in baggage X-ray scans. To summarize, the significant contributions are:

- We propose the MHT-X (Multi-Scale Hierarchical Transformer for X-ray Detection) model, a novel approach specifically tailored for detecting concealed and cluttered threat objects in baggage X-ray scans, marking a significant advancement beyond traditional detection methods.
- To the best of our knowledge, our MHT-X model is the first to integrate multi-scale contour maps with a hierarchical feature extraction process based on transformers, significantly enhancing detection accuracy by capturing detailed features across various resolutions.
- The MHT-X model leverages a hierarchical, transformer-based feature extraction module, outperforming traditional CNN models by extracting highly representative features that capture the global context of illegal items through self-attention mechanisms. This approach significantly enhances the model's ability to accurately identify threats within complex spatial configurations.
- The proposed approach is adaptable, performing consistently across various scanner types, enhancing security screening effectiveness in diverse settings.

II. RELATED WORK

Recent advancements in computer-aided X-ray baggage screening have significantly improved the detection of prohibited items. Initially, researchers relied on basic image enhancement [13] and manual features, leading to subpar outcomes [14], [15]. Later, the advent of deep learning significantly enhanced threat detection using transfer learning by capitalizing on the meaningful semantics extracted from CNNs [16], [17], [18]. This progress has inspired newer methods incorporating contour cues [19], [20], [21], [22] and material information [23], spatial and channel-wise attention [24], [25], foreground extraction [26], and multi-scale features [27], further enhancing performance. For an in-depth exploration of these evolving X-ray baggage threat recognition techniques, refer to [28] and [29].

Studies exploring modern object detection models [8], [9] for X-ray security screening found their performance lacking due to the unique challenges of this domain, such as poor contrast and object occlusion [28], [30]. To overcome these hurdles, recent advancements have focused on employing advanced techniques like structured tensors, which enhance object contours in scans and significantly improve threat detection amidst occlusion [22], [31]. Hassan e al. [22] leveraged learned features extracted from the contour representation of the X-ray images to recognize the occluded threats from cluttered scans using an encoder-decoder framework. Shafay et al. [27] leveraged information from multi-level features for identifying suspicious items using a lightweight encoder-decoder framework. Wei et al. [23] designed a module to enhance the effectiveness of state-of-the-art (SOTA) detection frameworks by leveraging low-level edge and material information. Hu et al. [24] attempted to tackle overlapping and scale variation challenges in X-ray imagery by employing spatial attention and anisotropic fusion. They leveraged multi-scale features and intelligently fused per-stage predictions to achieve better results. Bhowmik and Breckon [32] introduced a pipeline integrating segmentation and classification techniques to detect illicit baggage items. Studies [18], [33] have also integrated learned features

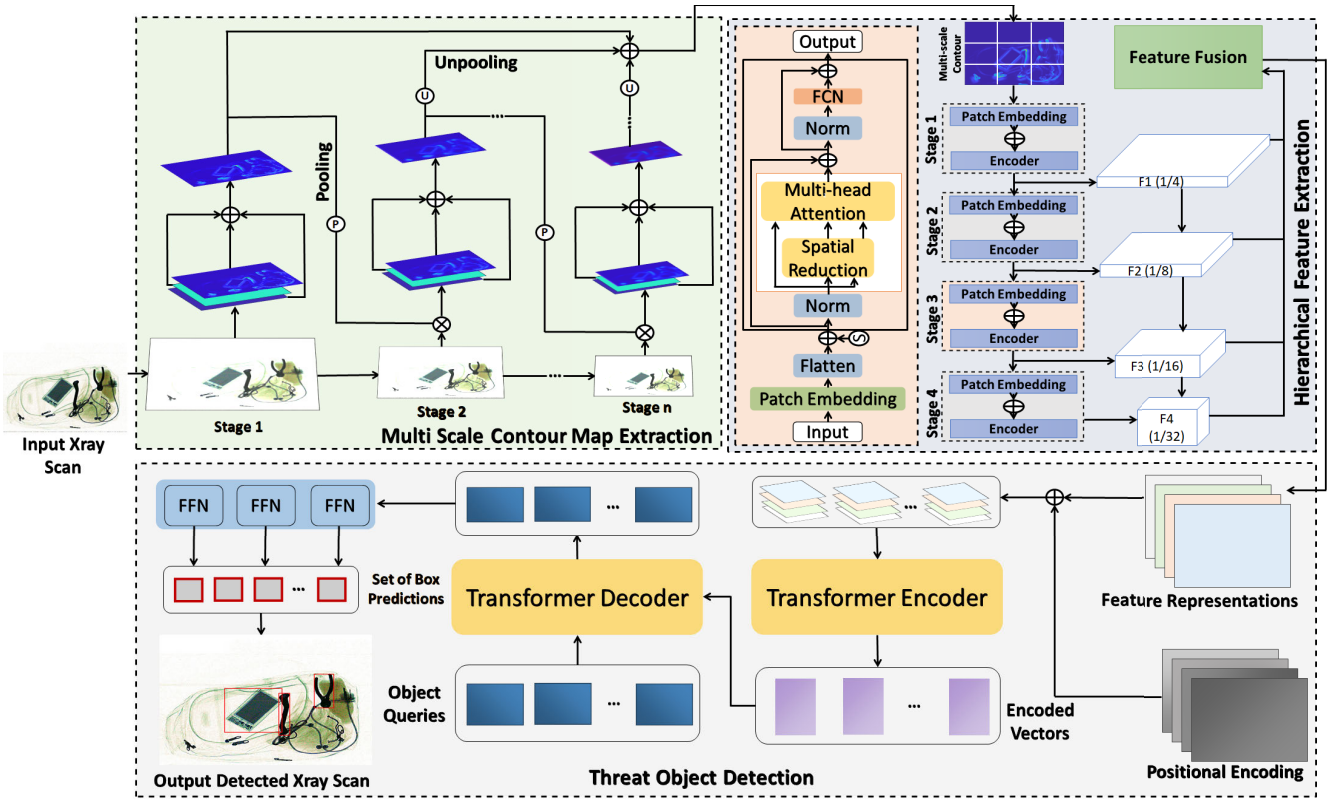


FIGURE 1. Block diagram of the proposed framework, which demonstrates a multi-stage process for identifying concealed threats in baggage X-ray scans. The method begins by extracting contour maps at multiple levels. It then proceeds through a hierarchical feature extraction approach, resulting in accurate predictions of bounding boxes using a transformer encoder-decoder architecture.

with broad learning systems to exploit faster, lightweight frameworks for baggage threat recognition. Zhu et al. [34] addressed stacking and occlusion in X-ray baggage imagery using multi-scale features and adaptive fusion of channel features. Shao et al. [26] designed an X-ray threat detection framework for extracting the target foreground by adaptively isolating items and focusing on relevant areas, significantly outperforming existing methods with limited data. Issac-Madina et al. [35] analyzed the detection performance of leading object detectors and examined the challenges in detecting smaller occluded threats. Recently, Liao et al. [20] explored a threat recognition approach that integrated feature enhancement while preserving shape and textural data, effectively suppressing noisy information using frequency domain knowledge.

Research in this field primarily focuses on occlusion and extracting key semantics from low-level features for threat identification [21], [23], with few studies tackling the class imbalance issue in X-ray baggage datasets, which necessitates specialized data sampling and algorithmic solutions [17], [36]. Miao et al. [17] used a modified loss to address class imbalance and leveraged hierarchical, refined features to localize the prohibited items. Ahmed et al. [37] introduced balanced affinity loss, integrating max-margin learning with effective sample utilization, presenting a more refined approach to classifying X-ray images.

Most of the current research in X-ray baggage threat detection focuses on CNNs, which are limited by their ability to encode only local interactions and thus struggle to capture the global context of images, a crucial factor for effectively detecting occluded threats in compact X-ray scans [25], [38]. In contrast, transformer models [12], [39], [40], widely recognized in diverse areas of computer vision and speech processing, excel in capturing this global context [41]. One recent work has delved into the potential of ViTs to identify abnormal baggage scans, especially in the context of highly imbalanced data sets [25]. This exploration has laid the groundwork for our novel approach. Building upon these insights, we introduce MHT-X (Multi-Scale Hierarchical Transformer for X-ray Detection), a transformative framework for the detection of concealed and obstructed threat objects in baggage X-ray scans, marking a significant advancement over traditional detection methods. MHT-X adeptly confronts the intricate challenges of X-ray imagery, particularly in environments characterized by dense clutter and occlusion. Central to our framework is an innovative multi-scale contour map generation process that meticulously dissects input scans into various levels to unveil distinctive transitional patterns, crucial for identifying contents within the baggage. This method facilitates the generation of detailed contour maps, pivotal for the precise delineation of objects amidst clutter. MHT-X further incorporates a

top-down hierarchical feature extraction model, extracting vital features at different resolutions from these contour maps, enabling a comprehensive scan analysis. These features are then seamlessly integrated into a transformer encoder and decoder system, enhanced by a Feed Forward Network (FFN) for the pinpoint detection of threats. The FFN predicts either a detection (class and bounding box) or a "no object" class with accuracy and precision. MHT-X transcends the constraints of conventional CNN-based approaches, setting a new benchmark in automated X-ray baggage inspection with its robust, detailed, and precise threat detection capabilities. This cutting-edge framework significantly elevates detection accuracy while substantially reducing false positives, offering an integrated solution that addresses the evolving needs of X-ray baggage security with unparalleled efficiency and effectiveness.

III. PROPOSED FRAMEWORK

In this work, we present a novel framework (MHT-X) for the efficient detection of concealed and obstructed threat objects in baggage X-ray scans. Our approach integrates a multi-scale contour feature extraction algorithm with a hierarchical threat detection system using ViT. The block diagram of this novel method is depicted in Figure 1, illustrating the process starting from the input scan, which is processed through a multi-scale contour map extraction module. These maps are crucial in representing the contours of the contraband objects accurately. They are then fed into a hierarchical structure of ViT for meticulous feature extraction. This extraction is pivotal for the accurate prediction of bounding boxes, accomplished through a transformer encoder and decoder. Our framework is designed to tackle the challenges of clutter and occlusion in X-ray imagery, enhancing detection capabilities in security scanning technology. In the subsequent section, we will explain in detail each module of the proposed framework.

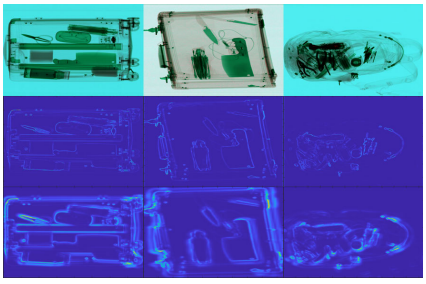


FIGURE 2. Comparison of conventional and multi-Scale contour maps extraction approaches. The first row presents the original X-ray scans. The second row demonstrates the conventional single-scale contour maps. The third row showcases the enhanced extraction of contours achieved by the proposed multi-scale contour maps module.

A. MULTI-SCALE CONTOUR MAP EXTRACTION

The generation of multi-scale contour maps involves a sophisticated process that begins with the decomposition of an input scan into a pyramid of multiple levels. Each level of this pyramid reveals unique transitional patterns in the baggage content by analyzing the distribution of orientations within the image gradients. This is achieved through a tensor pooling scheme, where transitional patterns are accentuated in N distinct image gradients, as detailed in Algorithm 1.

In this scheme, each tensor, denoted as $\Psi * (\Delta^k.\Delta^m)$ in the block matrix, is formulated as the outer product of two image gradients along with a smoothing filter Ψ. The orientation ω of each image gradient Δ^j is calculated using the formula $\omega = \frac{2\pi j}{N}$, where j varies from 0 to $N-1$. The coherent tensor, which is central to capturing the predominant orientations of baggage items, is constructed by summing the most significant tensors from the set of unique tensors, selected based on their normative values. This method not only uncovers variations in the intensity of cluttered items but also assists in forming individual contours for each item, as visually represented in Figure 2. In our approach, we emphasize the transitional patterns in Q image gradients (associated with Q directions) by computing the following $Q \times Q$ block-structured symmetric matrix in Eq. 1:

$$\begin{bmatrix} \Gamma*(\Delta^0\Delta^0) & \Gamma*(\Delta^1\Delta^0) & \dots & \Gamma*(\Delta^{Q-1}\Delta^0) \\ \Gamma*(\Delta^0\Delta^1) & \Gamma*(\Delta^1\Delta^1) & \dots & \Gamma*(\Delta^{Q-1}\Delta^1) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma*(\Delta^0\Delta^{Q-1}) & \Gamma*(\Delta^1\Delta^{Q-1}) & \dots & \Gamma*(\Delta^{Q-1}\Delta^{Q-1}) \end{bmatrix} \tag{1}$$

where Q represents the number of image gradients in different orientations within the scan, Γ refers to the Gaussian smoothing function, and the tensors are denoted by $\Gamma * (\Delta^q\Delta^q)$, representing a second-moment matrix for transition intensity maps. From Eq. 1, the multi-oriented tensor matrix is symmetric. Thus, a coherent map of the object transitions can be generated by adding the most coherent unique tensors. Moreover, each outer product in this matrix indicates the predominant orientations of changes within a specified pixel's neighborhood.

To address the limitations inherent in single-scale analysis, a multi-scale tensor strategy is employed. This strategy integrates X-ray scan transitions from the coarsest to the finest levels, thereby ensuring that every item, regardless of its low-intensity contrast with the background, is effectively emphasized. The multi-scale tensors are computed through a method of pyramid pooling up to a predefined nth level. At each level l, the decomposed image is combined, on a pixel-by-pixel basis, with the transitions obtained at the preceding $l-1$ levels. This approach guarantees that the edges of the contraband items are consistently maintained across different scales. The algorithm below outlines the steps involved in this multi-scale tensor pooling process, crucial for generating detailed and precise contour maps for effective threat detection in baggage X-ray scans.

B. HIERARCHICAL FEATURE EXTRACTION

In the field of computer vision, significant progress has been made through the development of algorithms capable of

Algorithm 1 Multi-Scale Tensor Pooling for Contour Map Generation

Input: Input X-ray scan I, Number of scales n, Number of orientations O
Output: Multi-scale Tensor M_T

M_T of size equal to I is initialized with zeros;
Set pyramid pooling factor $\pi \leftarrow 2$;
for $i \leftarrow 0$ *to* $n-1$ **do**
  if $i == 0$ **then**
    $T \leftarrow \text{ComputeTensors}(I, O)$;
    $T_C \leftarrow \text{GetCoherentTensor}(T)$;
    $M_T \leftarrow M_T + T_C$;
  else
    $I, T_C \leftarrow \text{Downscale}(I, T_C, \pi)$;
    $T \leftarrow \text{ComputeTensors}(I, O)$;
    $T_C \leftarrow \text{GetCoherentTensor}(T)$;
    $M_T \leftarrow M_T + \text{Upscale}(T_C, \pi^i)$;
  end
end

extracting features in a layered manner. These algorithms are essential for solving complex prediction tasks. The key to this advancement lies in a hierarchical approach that utilizes a top-down strategy, essential for the model's effective handling of the input contour maps. This method combines elements from both CNNs and Transformers in a unique manner.

The ViT [12] is a neural network architecture that applies self-attention mechanisms to process image patches, enabling global context understanding in computer vision tasks. Our MHT-X leveraged the ViT transformer encoder and introduces a top-down hierarchical feature extraction model. To incorporate hierarchical information, the ViT transformer encoder block is modified to include a spatial reduction block before applying the multi-head attention mechanism. The spatial reduction block is selected as a simple average pooling block, while the remainder of the ViT transformer encoder remains unaltered.

In detail, the first step involves creating patch embeddings, which convert raw data into a format ready for detailed, multi-level analysis. As the data moves through each phase of encoding, performed by spatial reduction ViT transformer encoders, it undergoes refinement and reduction in dimensionality. The feature maps at each level, labeled as F_l, become more focused. A key element of this refinement is the advanced multi-head attention mechanism, which includes spatial reduction to efficiently reduce the dimensions of the feature maps while keeping essential spatial information. This process is mathematically expressed as:

$$F_l' = \sum_{i=1}^{N} W_i \cdot \text{Attention}(\text{SR}(F_l)) \tag{2}$$

Here, F_l' indicates the transformed feature map at level l, enhanced by a combination of spatially reduced attention mechanisms, with W_i denoting the weights associated with each attention head, highlighting the selective emphasis on different data components. Reducing dimensionality is crucial for handling the computational demands associated with dense prediction models. The multi-scale feature maps, denoted as $F1$ to $F4$, are produced through four stages, each comprising patch embedding and encoder layers as depicted in Figure 1, highlighted in orange. These feature maps are hierarchically organized, with $F1$ representing the finest details and $F4$ capturing broader context. An essential step in this process is combining multi-scale feature maps, from $F1$ to $F4$, into a unified representation. This combination, essential for enriching the model's interpretation of the input, is described as:

$$F_{fusion} = \text{Fuse}(F1, F2, F3, F4) \tag{3}$$

This approach markedly enhances the model's proficiency in executing tasks such as object detection and segmentation. By merging features from different resolutions, the model gains a more holistic understanding of the input, which facilitates more precise and granular predictions.

C. THREAT OBJECT DETECTION

In our proposed method, the module for proposal extraction of threat detection revolutionizes the typical object detection process by leveraging the transformer architecture, an approach inspired by breakthroughs in natural language processing [42]. This module eschews the need for conventional, hand-designed components such as non-maximum suppression and anchor generation, instead conceptualizing object detection as a direct set prediction problem, which streamlines the overall detection process.

The method initiates with a top-down hierarchical feature extraction modified-ViT backbone that performs the feature extraction, converting the input image into a succinct feature representation. These features are subsequently fed into a transformer encoder that preserves the spatial structure of the data. The transformer decoder then takes over, predicting a set number of detections characterized by bounding boxes and class labels, which collectively constitute the output. The decoder output embedding is directed to a shared FFN for predictions, which comprises two heads: a classification head with a dense layer matching the number of object classes, and a sequential head consisting of three dense layers. Within the sequential head, two layers with 256 neurons each utilize the ReLU [43] activation function, followed by a final layer with four neurons utilizing the sigmoid function for bounding box prediction.

Our module utilizes the Hungarian matching loss function [44] to facilitate a one-to-one mapping between the predicted bounding boxes and the ground truths for rendering clear and distinct predictions. The end-to-end training capability intrinsic to this module eliminates the need for numerous task-specific adjustments and heuristic post-processing. This allows the module to learn autonomously to focus on relevant sections of the image and to interpret

the broader context, which is critical for accurately predicting bounding boxes and, consequently, for the effective detection of threats. The integration of the transformer encoder and decoder within this module marks a notable advance in the field of object detection, significantly improving the module's efficiency and accuracy in threat identification.

IV. EXPERIMENTAL SETUP

A workstation housing an Nvidia GeForce RTX 3080 GPU, an 11th Gen Intel 2.3GHz CPU, 32GB RAM, and 8GB VRAM is utilized to conduct our experiments. We trained MHT-X for 300 epochs using the AdamW [45] optimizer, with initial learning rates of 10^{-4} for the transformer and 10^{-5} for the backbone. Transformer weights were initialized using Xavier initialization [46], and the backbone used an ImageNet-pretrained [47] modified-ViT inspired from [48]. Image scaling to 500×500 and random crop augmentations were applied during training, with a transformer dropout rate of 0.1. The output from the Hierarchical Feature Extraction module for an image of this scale is a $16^2 \times 512$ dimensional tensor, which then passes through a $1^2 \times 2048$ convolutional layer to ensure compatibility with the subsequent transformer architecture. The transformer architecture comprises six encoder and six decoder layers, each with a width of 256. Sine spatial positional encodings are integrated at the attention layers within both the encoder and decoder. The decoder ingests a predefined number (100) of learned positional embeddings and further attends to the encoder output. Subsequently, each decoder output embedding is directed to a shared FFN responsible for predictions. The FFN consists of dual heads: a classification head with a dense layer matching the number of object classes, and a sequential head comprising three dense layers. The sequential head includes two layers with 256 neurons each employing the ReLU [43] activation function, followed by a final layer with four neurons using the sigmoid function for bounding box prediction.

For the classification head, we employ cross-entropy loss, while for the bounding box regression loss, we adopt a methodology outlined in [44], which combines the ℓ_1 loss and the scale-invariant Generalized Intersection over Union (GIoU) loss [49]. This choice is motivated by the fact that, unlike other detectors that predict boxes relative to initial guesses, we directly predict bounding boxes, which can lead to challenges in loss scaling. Our box loss is formulated as $\lambda_{iou}Liou(b_i, \hat{b}\sigma(i)) + \lambda L1||b_i - \hat{b}\sigma(i)||1$, where λ_{iou} and λ_{L1} are hyperparameters set to $\lambda_{iou} = 2$ and $\lambda_{L1} = 5$, respectively. Additionally, the losses are normalized by the number of objects in the batch [44].

A. DATASETS

Our experiments utilized the PIDray [50], CLCXray [51], and EDS [52] datasets, as outlined in Table 1 and subsequent subsections. We adhered to standard training/testing split protocols for each dataset, where the model was trained on the training set and subsequently evaluated on the testing set of each respective dataset.

TABLE 1. Chosen object detection under X-ray security inspection scenario datasets.

Dataset	#Total Imgs	#Training Imgs	#Validation Imgs	#Testing Imgs	#Classes
PIDray [50]	47,677	29,457	-	18,220	12
CLCXray [51]	9,565	7,664	956	956	12
EDS [52]	14,219	-	-	-	10

1) CLCXRAY

The CLCXray dataset consists of a total of 9,565 radiographic images, divided into two main subsets: 4,543 authentic images obtained from real subway environments, and 5,022 synthetic images generated by scanning manually assembled baggage. This dataset encompasses 12 distinct classes, covering a range of prohibited items. These classes include various cutting instruments, such as blades, daggers, knives, scissors, and Swiss Army knives, as well as seven types of liquid containers: cans, carton drinks, glass bottles, plastic bottles, vacuum cups, spray cans, and tins. The dataset's annotations are structured according to the COCO format, facilitating their use in object detection and image segmentation tasks. The COCO format, represented by *'json'* files, specifies how object annotations, including bounding box coordinates, category labels, and segmentation masks, are organized within a dataset, ensuring standardized labeling and metadata formatting.

2) PIDRAY

The PIDray dataset is an extensive collection specifically curated for detecting prohibited items in real-world settings, comprising a total of 47,677 images designated for both training and evaluation purposes. This dataset categorizes 12 distinct types of contraband, ranging from firearms and knives to wrenches, pliers, scissors, hammers, handcuffs, batons, sprayers, power banks, lighters, and ammunition. Each image in the dataset is annotated at both the image and instance levels, following the COCO format. The dataset is systematically partitioned, with 29,457 images (approximately 60%) allocated for training and 18,220 images (the remaining 40%) reserved for testing. The test set is further stratified based on detection complexity into three categories: 'easy,' featuring images with a single contraband item, 'hard', containing images with multiple prohibited items, and 'hidden', comprising images with deliberately concealed items to simulate more challenging detection scenarios.

3) EDS

The Endogenous Domain Shift (EDS) benchmark focuses on X-ray security inspection, where shifts among the domains primarily stem from variations in X-ray machine types, hardware parameters, and degrees of wear. EDS comprises

14,219 images, encompassing 31,654 common instances across three domains (X-ray machines), with bounding-box annotations for 10 categories including drink bottle, pressure, lighter, knife, device, power bank, umbrella, glass bottle, scissor, and laptop. The dataset's training/testing protocol entails training on one domain and testing on the other two domains. For brevity, we denote ''domain 1'', ''domain 2'', and ''domain 3'' as ''D1'', ''D2'', and ''D3'', respectively.

B. EVALUATION METRICS

Object detection model evaluation commonly relies on the mean Average Precision (mAP) metric, which encompasses various sub-metrics such as the confusion matrix, Intersection over Union (IoU), recall, and precision. We detail each sub-metric below.

The confusion matrix presents a tabular breakdown of the model's predictions compared to the ground truth labels, consisting of four quadrants:

- True Positives (TP): Instances where the model accurately identifies and localizes objects within the image.
- True Negatives (TN): Instances where the model correctly identifies background or non-object regions.
- False Positives (FP): Instances where the model incorrectly predicts the presence of an object or mislocalizes it.
- False Negatives (FN): Instances where the model fails to detect and localize objects present in the image.

IoU measures the overlap between predicted and ground truth bounding boxes, with higher values indicating better alignment. Precision assesses the model's accuracy in identifying true positives out of all positive predictions (TP+FP), calculated using IoU thresholds. Recall measures the model's ability to locate true positives out of all actual positive instances (TP+FN).

mAP is computed by averaging the Average Precision (AP) for each class (N):

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{4}$$

AP is the weighted mean of precisions at various thresholds, considering the increase in recall from the prior threshold. mAP balances precision and recall, considering both FP and FN, making it suitable for detection tasks. Our evaluation utilizes COCO metrics including mAP, mAP50, and mAP75, calculated across IoU thresholds ranging from 0.5 to 0.95, and specifically at thresholds of 0.5 and 0.75 for mAP50 and mAP75, respectively.

C. ABLATION STUDIES

The ablation studies conducted in this research are focused on five primary objectives: 1) Determining the optimal number of orientations and scaling levels for the multi-scale contour maps, 2) identifying the most suitable backbone model for our object detection model, 3) evaluating the impact of integrating the multi-scale contour maps extraction module, 4) assessing the effect of using hierarchical features versus not, and 5) examining the influence of utilizing transformer encoder/decoder versus not.

1) NUMBER OF ORIENTATIONS AND SCALING LEVELS

In our exploration of the multi-scale contours extraction method's efficacy on the CLCxray and PIDray datasets, we meticulously analyzed the impact of varying the number of orientations (N) and scaling levels (n) on detection performance. This method, fundamental to our framework, enhances the delineation of threat objects by generating detailed contour maps from X-ray scans. Our findings, distilled into a comprehensive performance evaluation, reveal that adjusting N and n parameters significantly influences the model's ability to accurately identify concealed threats. enhancing both N and n to 5 from their initial values of 2 resulted in significant enhancements in detection accuracy. Specifically, for the CLCxray dataset, there was a notable 14.10% improvement, while for the PIDray dataset, the increase was even more substantial at 16.05%. These results underscore the critical role of multi-scale contours extraction in achieving superior detection outcomes, demonstrating its effectiveness in extracting nuanced features crucial for the precise identification of contraband items within complex X-ray imagery.

TABLE 2. Detection performance (mAP) for different configurations of *N* and *n* on PIDray and CLCXray datasets.

Dataset	N	mAP (Performance)			
		$n=2$	$n=3$	$n=4$	$n=5$
PIDray	2	0.5000	0.5402	0.5803	0.5205
	3	0.5402	0.5803	0.5205	0.5606
	4	0.5803	0.5205	0.5606	0.6008
	5	0.5205	0.5606	0.6008	**0.6122**
CLCXray	2	0.5200	0.5560	0.5921	0.5281
	3	0.5560	0.5921	0.5281	0.5642
	4	0.5921	0.5281	0.5642	0.5002
	5	0.5281	0.5642	0.5002	**0.6042**

2) CHOICE OF A BACKBONE MODEL

In x-ray imagery for detecting illegal and smuggled items, common challenges such as clutter and occlusion significantly impede the performance of deep learning models. A critical and distinct challenge in this context is the multi-scale nature of these items, as they can appear in varying sizes from small to large within the images. This necessitates a model capable of effectively handling objects across multiple scales. To validate this hypothesis, we employed our detection module, integrating it with three distinct backbones: 1) The conventional DETR with a ResNet50 [53] backbone, 2) EfficientNetV2XL [54], and 3) Our proposed hierarchical feature extractor. The performances of these models were evaluated in terms of mAP on the PIDray dataset, with results documented in Table 3. Our findings indicate that incorporating hierarchical multi-scale information via the MHT-X significantly enhances performance on the PIDray

dataset, surpassing the second-best, Efficient DETR, by an average mAP increment of 3.23%. This marked improvement underscores the effectiveness of a pyramid-based architecture in addressing the multi-scale challenges posed by illegal items in X-ray images.

TABLE 3. Comparison of Multi-Class Threat Detection Performance on PIDray Dataset with Different Backbone Architectures. Bold values indicate the highest mAP scores.

Framework	Backbone	PIDray Sets			
		Easy	Hard	Hidden	Overall
DETR [44]	ResNet50 [53]	0.727	0.668	0.513	0.636
Efficient DETR	EfficientNetV2XL [54]	0.69	0.63	0.49	0.60
DETR [44]	ViT [12]	0.70	0.64	0.50	0.62
MHT-X (ours)	Hierarchical Multi-Scale Features	**0.7819**	**0.6845**	**0.5565**	**0.6743**

3) EFFECT OF MULTI-SCALE CONTOURS

The Multi-Scale Contour map extraction module is employed to accentuate the transitions in baggage content across various image gradients, oriented in N directions and up to n scaling levels. Balancing these parameters is crucial for achieving optimal contour representation and robust detection while managing computational demands.

Comparatively, Trainable Structure Tensors (TST), as presented in previous work [55], scan multiple orientations to highlight contours of occluded and cluttered contraband items, while minimizing irrelevant baggage content. In our research, we juxtaposed the Multi-Scale Contour module and TST with their raw RGB counterparts on the PIDray dataset to highlight the benefits of our approach in addressing the challenge of detecting highly cluttered and concealed illegal items in X-ray images. The comparative analysis, tabulated in Table 4, focused on mAP, reveals that our novel framework outperforms both RGB and TST configurations across Easy, Hard, and Hidden sets, as well as in the overall score. Specifically, in the Easy set, MHT-X with multi-scale contours achieved a mAP of 0.7819, compared to RGB's 0.71 and TST's 0.58. This superiority is also evident in the Hard and Hidden sets, where the our module achieved mAPs of 0.6845 and 0.5565, respectively, surpassing the other two configurations. Overall, the proposed approach with multi-scale contours demonstrated the highest performance, achieving a mAP of 0.6743.

TABLE 4. Comparison of multi-class threat detection performance on PIDray dataset with different input scan types. Bold values indicate the highest mAP scores.

Input Scan	PIDray Sets			
	Easy	Hard	Hidden	Overall
RGB	0.71	0.65	0.51	0.62
single-scale contours	0.58	0.55	0.44	0.52
Multi-scale contours	**0.7819**	**0.6845**	**0.5565**	**0.6743**

4) EFFECT OF HIERARCHICAL FEATURES INTEGRATION

The Hierarchical Feature Extraction module incorporates an average pooling block to downscale the spatial resolution of features before applying the multi-head attention mechanism, as outlined in Equation 2. These features with multi-scale resolutions are then fused to produce the final feature map, as depicted in Equation 3.

To assess the advantages of integrating multi-scale resolution features in addressing challenges such as multi-scale objects, we replaced the Hierarchical Feature Extraction module (backbone) with a simple ViT [12] and compared it with our modified ViT on the PIDray dataset. The

TABLE 5. Comparison of multi-class threat detection performance on PIDray dataset using different transformer configurations. Bold values indicate the highest mAP scores.

Framework	Transformer Encoder	Transformer Decoder	PIDray Sets			
			Easy	Hard	Hidden	Overall
MHT-X	✓	✗	0.74	0.63	0.52	0.63
MHT-X	✗	✓	0.71	0.61	0.49	0.60
MHT-X (ours)	✓	✓	**0.7819**	**0.6845**	**0.5565**	**0.6743**

TABLE 6. Comparative analysis of multi-class threat detection performance on PIDray dataset in terms of mAP. Bold values indicate the highest performance.

Framework	PIDray Sets			
	Easy	Hard	Hidden	Overall
Faster R-CNN [56]	0.633	0.572	0.421	0.542
Mask R-CNN [8]	0.647	0.590	0.438	0.558
SSD512 [57]	0.681	0.589	0.457	0.576
Cascade R-CNN [58]	0.693	0.628	0.480	0.600
Cascade Mask RCNN [59]	0.709	0.640	0.480	0.610
SDANet [50]	0.712	0.642	0.495	0.616
DETR [60]	0.727	0.668	0.513	0.636
MHT-X (ours)	**0.7819**	**0.6845**	**0.5565**	**0.6743**

TABLE 7. Comparative analysis of multi-class threat detection performance on CLCXray dataset in terms of mAP. Bold values indicate the highest performance.

Framework	mAP	mAP_{50}	mAP_{75}
SSD [57]	51.1	66.4	59.8
Cascade R-CNN [58]	58.4	71.4	68.4
FSAF [61]	55.8	70.0	66.6
NAS-FCOS [62]	57.3	72.3	67.7
FCOS+DOAM [23]	54.3	68.5	63.5
PAA [63]	58.3	71.6	68.5
ATSS+LAcls [51]	0.593	0.718	0.682
MHT-X (ours)	**0.6526**	**0.8087**	**0.7381**

TABLE 8. Comparative analysis of multi-class threat detection performance on each group of adaption of EDS dataset in terms of mAP. $D_{m\to n}$ refers to the adaptation from domain *m* to domain *n*. "SO" refers to "Source only", the Faster R-CNN model trained on the source domain only. Bold values indicate the highest performance.

Framework	$D_{1\to2}$	$D_{1\to3}$	$D_{2\to1}$	$D_{2\to3}$	$D_{3\to1}$	$D_{3\to2}$
SO [64]	42.3	53.6	41.8	55.4	52.7	53.6
DA [65]	46.3	55.6	45.0	57.5	56.1	54.6
SWDA [66]	46.9	56.5	49.7	56.7	56.6	54.8
CST [67]	46.9	54.3	49.2	55.5	56.5	52.8
CFA [68]	44.3	53.7	51.3	53.6	55.4	51.3
PSN [52]	48.3	57.6	**51.4**	57.8	58.6	54.9
MHT-X (ours)	**49.1**	**58.8**	**51.4**	**58.1**	**58.9**	**55.5**

comparative analysis, presented in Table 3, focusing on mAP, reveals that our incorporation of top-down hierarchical information (multi-scale features) outperforms the reliance on single-scale features across the Easy, Hard, and Hidden subsets, as well as in the overall score. Specifically, in the Easy, Hard, Hidden, and overall subsets, MHT-X achieved improvements in mAP of 10.47%, 6.5%, 7.82%, and 8.05%, respectively.

5) IMPACT OF TRANSFORMER ENCODER-DECODER INTEGRATION

The conventional transformer architecture [42] comprises two main components: the encoder and the decoder. The encoder is responsible for processing the input sequence and extracting relevant features, while the decoder generates output sequences based on these features. To assess the significance of each component, we conducted an ablation study within our proposed MHT-X framework, systematically removing either the encoder or the decoder and evaluating the mAP on the PIDray dataset, as summarized in Table 5. In our analysis, if either the encoder or decoder is used alone, we maintain a fixed value of six encoders or decoders, respectively. This examination offers insights into the distinct roles and contributions of each component within the transformer architecture.

The results presented in Table 5 indicate that, overall, employing only the transformer encoder yielded a performance increase compared to utilizing only the transformer decoder, with an improvement of 4.44%. However, this approach still lagged behind the utilization of both encoder and decoder by 7.03%. Upon inspecting these findings, we infer that the transformer encoder holds greater importance in image detection tasks due to its responsibility for processing input image features and capturing long-range dependencies. Conversely, the decoder's role primarily involves refining and interpreting these encoded features. Removing the encoder disrupts the extraction of features with long-range dependencies, thereby impacting the decoder's ability to effectively refine and interpret the encoded features.

V. COMPARATIVE RESULTS

A. PIDRAY

Our MHT-X model exhibits remarkable performance in the multi-class threat detection task on the PIDray dataset, as evident from the comparative analysis outlined in Table 6. When evaluated across varying levels of difficulty, Easy, Hard, and Hidden, the MHT-X consistently outperforms other leading frameworks. In the 'Easy' set, MHT-X achieves a mAP of 0.7819, surpassing the next best model, DETR [60], which scores 0.727. Similarly, in the 'Hard' set, our model leads with an mAP of 0.6845, compared to DETR's [60] 0.668. The superiority of proposed framework is further underscored in the 'Hidden' set, where it attains an mAP of 0.5565, significantly higher than DETR's [60] 0.513. The proposed approach achieves a composite mAP of 0.6743, which is notably higher than any other competing framework, marking it as the best-performing model in this study. This clearly demonstrates the advanced capabilities of the MHT-X in effectively detecting threats across various complexity levels in X-ray imagery, solidifying its status as a leading solution in this domain.

B. CLCXRAY

In the domain of multi-class threat detection on the CLCXray dataset, our MHT-X model demonstrates exceptional performance, as detailed in Table 7. The model's effectiveness is evident across all measured metrics, including mAP, mAP50, and mAP75. For the overall mAP, MHT-X achieves an impressive score of 0.6526, significantly outpacing the nearest competitor, ATSS+LAcls [51], which registers an mAP of 0.593. This trend of superior performance is maintained in the more specific metrics of mAP50 and mAP75, where MHT-X scores 0.8087 and 0.7381, respectively. These scores are notably higher than those of ATSS+LAcls, which achieves 0.718 and 0.682 in the same categories. These results firmly establish the proposed framework as the leading framework in threat detection on the CLCXray dataset, with its highest recorded values in all three key performance indicators. The model's dominance in this comparative analysis underscores its advanced capability in accurately identifying threats across various detection thresholds, making it a highly effective tool for security and surveillance applications in X-ray imagery analysis.

C. EDS

Our MHT-X model demonstrates SOTA performance in multi-class threat detection on the EDS dataset, as illustrated by the comparative analysis presented in Table 8. Across all evaluated scenarios, MHT-X consistently surpasses other leading frameworks. On average, MHT-X achieves an mAP of 0.553, approximately 1% higher than the average mAP of the second-best performing model, PSN [52]. This underscores the advanced capabilities of MHT-X in effectively detecting threats across diverse X-ray imagery domains, solidifying its position as a leading solution in this field.

D. QUALITATIVE RESULTS

For visualization, we provide qualitative evaluations of the proposed framework on PIDray, CLCXray, and EDS datasets in Figure 3. In this figure, the first row shows examples of PIDray dataset, the second row shows examples of CLCXray dataset, and the third row shows samples from EDS dataset. Here, we can appreciate how accurately the proposed scheme has picked different cluttered suspicious items in (b)-(f). Moreover, we can also observe the extracted gun in (a) even with different orientation. Such examples demonstrate the proposed framework's capacity in picking the cluttered items from the PIDray, CLCXray, and EDS datasets.

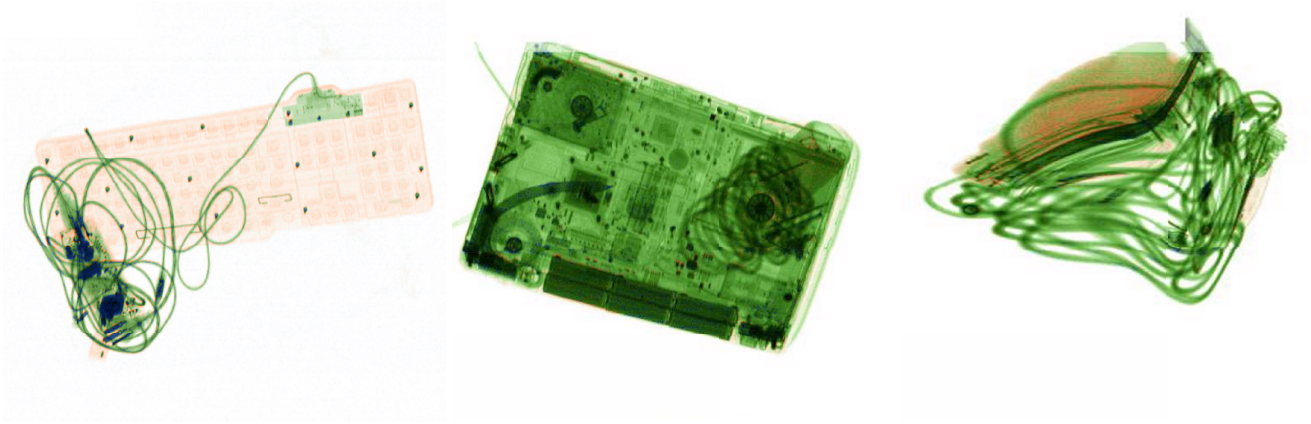
In Figure 4, we report examples of failure cases encountered during the testing. In all examples, we can see that the proposed framework could not detect the contraband items.

FIGURE 3. Qualitative results of three benchmarks where the first row shows examples of the PIDray dataset ordered by complexity as follows: (a) easy, (b) hard, and (c) hidden, the second row shows examples of the CLCXray dataset, and the third row shows examples of the EDS dataset ordered as follows: (a) D_1, (b) D_2, and (c) D_3.

However, such cases are observed in highly occluded scans, where it is difficult, even for a human observer, to distinguish the items' regions properly.

E. MODEL COMPLEXITY

A comparative analysis of different methods in terms of model complexity and run-time performance, for an image size of 224 × 224 is presented in Table 9. The comparison is based on inference time, the number of parameters (in millions), and Multiply-Accumulate Operations (MACs, in billions). MHT-X showcases exceptional efficiency with the lowest MACs at 5.91G, indicating high computational efficiency. Additionally, it achieves competitive inference time (0.14 seconds) despite having the highest number



(a) (b) (c)

FIGURE 4. Failed detection cases by our MHT-X on PIDray (a) easy, (b) hard, and (c) hidden testing sets.

of parameters (100.52M), highlighting a trade-off between model complexity and processing efficiency. On the other hand, methods like YOLOv3 and YOLACT demonstrate a more balanced approach. YOLOv3 has the quickest inference time (0.023 seconds) with moderate complexity, while YOLACT also maintains a lower inference time (0.036 seconds) and MACs (59.2G). Other methods, including MS R-CNN, RetinaNet, and Mask R-CNN, show varied performance with higher MACs and varying inference times, indicating different trade-offs between efficiency and complexity. Overall, MHT-X's position is unique, excelling in computational efficiency but at the cost of longer processing times indicating a potential trade-off between complexity and efficiency.

TABLE 9. Model complexity and run-time performance comparison of different methods for image size of 224 × 224.

Method	Inference time (sec)	# Params(M)	MACs(G)
MS R-CNN [69]	0.156	44.5	141.1
RetinaNet [10]	0.033	34.0	151.5
CIE-Net	0.072	**33.4**	18.3
YOLOv3 [70]	**0.023**	63.1	78.6
Mask R-CNN [8]	0.141	44.4	134.3
YOLACT [71]	0.036	34.7	59.2
HTC [72]	0.311	83.2	120.1
MHT-X	0.14	100.52	**5.91**

We believe that in high-sensitivity environments like airports, prioritizing the accuracy of detection over inference time is critical for ensuring the highest level of security. Accurate identification of potential threats is paramount to effectively mitigate risks and maintain safety for passengers and staff. While minimizing inference time is desirable, it cannot come at the expense of compromising the accuracy of detection, which is essential for preventing security breaches and safeguarding public welfare.

VI. CONCLUSION

This paper proposes the Multi-Scale Hierarchical Transformer for X-ray Detection (MHT-X), a novel framework designed to enhance automated baggage screening systems. MHT-X addresses key challenges such as occlusion, clutter, and the global context of threats by integrating a multi-scale contour mapping technique with hierarchical feature extraction using ViT. This proposed method is effective and has outperformed different existing SOTA frameworks. However, our model requires more data and correctly labeled data to work effectively in a fully supervised manner. Additionally, the training complexity of MHT-X, with its substantial large number of parameters, poses challenges, especially for smaller-scale projects. Moreover, MHT-X's requirement for a fixed number of output slots limits its flexibility in scenarios with variable numbers of objects, leading to potential inaccuracies in cluttered scenes or when the number of objects exceeds the model's predefined output size. Future work could explore leveraging MHT-X for segmentation tasks and investigating the integration of Vision language models to further enhance its capabilities.

ACKNOWLEDGMENT

(Mohamad Alansari, Abdelfatah Ahmed, and Khaled Alnuaimi contributed equally to this work.)

REFERENCES

[1] D. Saavedra, S. Banerjee, and D. Mery, "Detection of threat objects in baggage inspection with X-ray images using deep learning," *Neural Comput. Appl.*, vol. 33, pp. 7803–7819, Jan. 2021.

[2] T. Hassan, S. Akcay, M. Bennamoun, S. Khan, and N. Werghi, "A novel incremental learning driven instance segmentation framework to recognize highly cluttered instances of the contraband items," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 11, pp. 6937–6951, Nov. 2022, doi: 10.1109/TSMC.2021.3131421.

[3] S. Michel, S. M. Koller, J. C. De Ruiter, R. Moerland, M. Hogervorst, and A. Schwaninger, "Computer-based training increases efficiency in X-ray image interpretation by aviation security screeners," in *Proc. Int. Carnahan Conf. Secur. Technol.*, 2007, pp. 201–206.

[4] I. Uroukov and R. Speller, "A preliminary approach to intelligent X-ray imaging for baggage inspection at airports," *Signal Process. Res.*, vol. 4, pp. 1–11, Jan. 2015.

[5] M. Mendes, A. Schwaninger, and S. Michel, "Can laptops be left inside passenger bags if motion imaging is used in X-ray security screening?" *Frontiers Hum. Neurosci.*, vol. 7, p. 654, Oct. 2013.

[6] X. Yu, W. Yuan, and A. Wang, "X-ray security inspection image dangerous goods detection algorithm based on improved YOLOv4," *Electronics*, vol. 12, no. 12, p. 2644, 2023.

[7] D. Mery and C. Pieringer, "Deep learning in X-ray testing," in *Computer Vision for X-Ray Testing: Imaging, Systems, Image Databases, and Algorithms*. Cham, Switzerland: Springer, 2021, pp. 275–336, doi: 10.1007/978-3-030-56769-9_7.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[9] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[11] M. D. Rouzi, B. Moshiri, M. Khoshnevisan, M. A. Akhaee, F. Jaryani, S. S. Nasab, and M. Lee, "Breast cancer detection with an ensemble of deep learning networks using a consensus-adaptive weighting method," *J. Imag.*, vol. 9, no. 11, p. 247, 2023.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[13] Z. Chen, Y. Zheng, B. Abidi, D. Page, and M. Abidi, "A combinational approach to the fusion, de-noising and enhancement of dual-energy X-ray luggage images," in *Proc. CVPR-Workshops*, 2006, pp. 1–6.

[14] M. Baştan, M. R. Yousefi, and T. M. Breuel, "Visual words on baggage X-ray images," in *Proc. Int. Conf. Comput. Anal. Images Patterns* Cham, Switzerland: Springer, 2011, pp. 360–368.

[15] D. Mery, E. Svec, and M. Arias, "Object recognition in baggage inspection using adaptive sparse representations of X-ray images," in *Image and Video Technology*. Cham, Switzerland: Springer, 2015, pp. 709–720.

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

[17] C. Miao, L. Xie, F. Wan, C. Su, H. Liu, J. Jiao, and Q. Ye, "SIXray: A large-scale security inspection X-ray benchmark for prohibited item discovery in overlapping images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2114–2123.

[18] D. Velayudhan, T. Hassan, A. H. Ahmed, E. Damiani, and N. Werghi, "Baggage threat recognition using deep low-rank broad learning detector," in *Proc. IEEE 21st Medit. Electrotechnical Conf. (MELECON)*, Jun. 2022, pp. 966–971.

[19] T. Hassan, S. Akcay, B. Hassan, M. Bennamoun, S. Khan, J. Dias, and N. Werghi, "Cascaded structure tensor for robust baggage threat detection," *Neural Comput. Appl.*, vol. 35, no. 15, pp. 11269–11285, 2023.

[20] H. Liao, B. Huang, and H. Gao, "Feature-aware prohibited items detection for X-ray images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2023, pp. 1040–1044.

[21] T. Hassan, M. Shafay, S. Akçay, S. Khan, M. Bennamoun, E. Damiani, and N. Werghi, "Meta-transfer learning driven tensor-shot detector for the autonomous localization and recognition of concealed baggage threats," *Sensors*, vol. 20, no. 22, p. 6450, 2020.

[22] T. Hassan, S. Akcay, M. Bennamoun, S. Khan, and N. Werghi, "Tensor pooling-driven instance segmentation framework for baggage threat recognition," *Neural Comput. Appl.*, vol. 34, no. 2, pp. 1239–1250, 2022.

[23] 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*, 2020, pp. 138–146.

[24] B. Hu, "Multi-label X-ray imagery classification via bottom-up attention and meta fusion," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 4–7.

[25] D. Velayudhan, A. H. Ahmed, T. Hassan, M. Bennamoun, E. Damiani, and N. Werghi, "Transformers for imbalanced baggage threat recognition," in *Proc. IEEE Int. Symp. Robotic Sensors Environ. (ROSE)*, Nov. 2022, pp. 1–7.

[26] 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 Recognit.*, vol. 122, Feb. 2022, Art. no. 108261.

[27] M. Shafay, T. Hassan, D. Velayudhan, E. Damiani, and N. Werghi, "Deep fusion driven semantic segmentation for the automatic recognition of concealed contraband items," in *Proc. SoCPaR*, 2020, pp. 550–559.

[28] D. Velayudhan, T. Hassan, E. Damiani, and N. Werghi, "Recent advances in baggage threat detection: A comprehensive and systematic survey," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 1–38, 2022.

[29] J. Wu, X. Xu, and J. Yang, "Object detection and X-ray security imaging: A survey," *IEEE Access*, vol. 11, pp. 45416–45441, 2023.

[30] S. Akcay and T. P. Breckon, "An evaluation of region based object detection strategies within X-ray baggage security imagery," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1337–1341.

[31] T. Hassan, M. Bettayeb, S. Akçay, 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 Process. (ICIP)*, Oct. 2020, pp. 2016–2020.

[32] N. Bhowmik and T. P. Breckon, "Joint sub-component level segmentation and classification for anomaly detection within dual-energy X-ray security imagery," in *Proc. 21st IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2022, pp. 1463–1467.

[33] M. Shafay, A. Ahmed, T. Hassan, J. Dias, and N. Werghi, "Programmable broad learning system for baggage threat recognition," *Multimedia Tools Appl.*, vol. 83, no. 6, pp. 1–18, 2023.

[34] X. Zhu, J. Zhang, X. Chen, D. Li, Y. Wang, and M. Zheng, "AMOD-Net: Attention-based multi-scale object detection network for X-ray baggage security inspection," in *Proc. 5th Int. Conf. Comput. Sci. Artif. Intell.*, 2021, pp. 27–32.

[35] B. K. S. Isaac-Medina, S. Yucer, N. Bhowmik, and T. P. Breckon, "Seeing through the data: A statistical evaluation of prohibited item detection benchmark datasets for X-ray security screening," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 524–533.

[36] R. Tao, Y. Wei, H. Li, A. Liu, Y. Ding, H. Qin, and X. Liu, "Over-sampling de-occlusion attention network for prohibited items detection in noisy X-ray images," 2021, *arXiv:2103.00809*.

[37] A. Ahmed, D. Velayudhan, T. Hassan, M. Bennamoun, E. Damiani, and N. Werghi, "Enhancing security in X-ray baggage scans: A contour-driven learning approach for abnormality classification and instance segmentation," *Eng. Appl. Artif. Intell.*, vol. 130, Apr. 2024, Art. no. 107639.

[38] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Intriguing properties of vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 23296–23308.

[39] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[40] D. Velayudhan, A. Ahmed, T. Hassan, N. Gour, M. Owais, M. Bennamoun, E. Damiani, and N. Werghi, "Autonomous localization of X-ray baggage threats via weakly supervised learning," *IEEE Trans. Ind. Informat.*, vol. 20, no. 4, pp. 6563–6572, 2024, doi: 10.1109/TII.2023.3348838.

[41] D. Velayudhan, A. Ahmed, T. Hassan, M. Bennamoun, E. Damiani, and N. Werghi, "Context-aware transformers for weakly supervised baggage threat localization," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2023, pp. 3538–3542.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 2–6.

[43] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.

[44] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[45] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[46] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, vol. 9, Y. W. Teh and M. Titterington, Eds. Chia Laguna Resort, Sardinia, Italy, May 2010, pp. 249–256.

[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[48] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, 2022.
[49] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.
[50] 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. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5392–5401.
[51] 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. Inf. Forensics Security*, vol. 17, pp. 998–1009, 2022.
[52] R. Tao, H. Li, T. Wang, Y. Wei, Y. Ding, B. Jin, H. Zhi, X. Liu, and A. Liu, "Exploring endogenous shift for cross-domain detection: A large-scale benchmark and perturbation suppression network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 21157–21167.
[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[54] M. Tan and Q. Le, "EfficientNetv2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.
[55] T. Hassan and N. Werghi, "Trainable structure tensors for autonomous baggage threat detection under extreme occlusion," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Nov. 2020, pp. 3–6.
[56] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
[57] J. Jeong, H. Park, and N. Kwak, "Enhancement of SSD by concatenating feature maps for object detection," 2017, *arXiv:1705.09587*.
[58] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.
[59] M. Wu, H. Yue, J. Wang, Y. Huang, M. Liu, Y. Jiang, C. Ke, and C. Zeng, "Object detection based on RGC mask R-CNN," *IET Image Process.*, vol. 14, no. 8, pp. 1502–1508, 2020.
[60] A. Ahmed, M. Alansari, K. Alnuaimi, D. Velayudhan, T. Hassan, and N. Werghi, "Detection transformer framework for recognition of heavily occluded suspicious objects," in *Proc. IEEE Int. Conf. Comput. Intell. Virtual Environments Meas. Syst. Appl. (CIVEMSA)*, Jun. 2023, pp. 1–6.
[61] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 840–849.
[62] N. Wang, Y. Gao, H. Chen, P. Wang, Z. Tian, C. Shen, and Y. Zhang, "NAS-FCOS: Fast neural architecture search for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11940–11948.
[63] K. Kim and H. S. Lee, "Probabilistic anchor assignment with iou prediction for object detection," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 355–371.
[64] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.
[65] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3339–3348.
[66] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6949–6958.
[67] G. Zhao, G. Li, R. Xu, and L. Lin, "Collaborative training between region proposal localization and classification for domain adaptive object detection," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 86–102.
[68] Y. Zheng, D. Huang, S. Liu, and Y. Wang, "Cross-domain object detection through coarse-to-fine feature adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13763–13772.
[69] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6402–6411.
[70] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[71] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9156–9165.
[72] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "Hybrid task cascade for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4969–4978.

MOHAMAD ALANSARI received the B.Sc. and M.Sc. degrees in electrical and computer engineering from Khalifa University, United Arab Emirates, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. His research interest includes computer vision in autonomous robotics.

ABDELFATAH AHMED received the B.Sc. degree in electrical and electronic engineering from the University of Sharjah, in 2020, and the M.Sc. degree in electrical and computer engineering from Khalifa University, United Arab Emirates, in 2022, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. His research interests include computer vision, deep learning, and speech signal processing.

KHALED ALNUAIMI received the B.Tech. degree (Hons.) in computer programming from the Royal Melbourne Institute of Technology (RMIT) and the M.Sc. degree in computational data science from Khalifa University, United Arab Emirates, in 2023, where he is currently pursuing the Ph.D. degree in engineering systems and management. He is also a Research Specialist with the Strategy and Planning Department, Abu Dhabi Investment Authority (ADIA), for 12 years. His research interests include computer vision, deep learning, and large language models.

DIVYA VELAYUDHAN received the B.Tech. degree in electronics and communication engineering from Cochin University of Science and Technology, India, and the M.Tech. degree in electronics and communication engineering from the University of Kerala, India. She is currently pursuing the Ph.D. degree in electrical and computer engineering with Khalifa University. Her research interests include signal processing, image analysis, and computer vision.

TAIMUR HASSAN received the B.S. degree in computer engineering from Bahria University, Islamabad, Pakistan, in 2013, the M.S. degree in computer engineering from the University of Engineering and Technology (UET), Taxila, Pakistan, in 2015, and the Ph.D. degree in computer engineering from the National University of Sciences and Technology (NUST), Islamabad, in 2019. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Abu Dhabi University, United Arab Emirates. Prior to that, he was a Postdoctoral Fellow with the Khalifa University Center for Autonomous Robotic Systems (KUCARS) and the Center for Cyber-Physical Systems (C2PS), Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, United Arab Emirates. He has worked on many local and foreign-funded research projects as a principal investigator, a co-principal investigator, and a lead scientist/engineer. His research interests include robotic vision, medical imaging, deep learning, signal processing, and computer vision. He was a recipient of various national and international awards.

SAJID JAVED received the B.Sc. degree in computer science from the University of Hertfordshire, U.K., in 2010, and the master's and Ph.D. degrees in computer science from Kyungpook National University, Republic of Korea, in 2017. He is currently a Faculty Member with Khalifa University (KU), United Arab Emirates. Prior to that, he was a Research Fellow with KU, from 2019 to 2021, and the University of Warwick, U.K., from 2017 to 2018. His research interests include visual object tracking in the wild, multi-object tracking, background-foreground modeling from video sequences, moving object detection from complex scenes, and cancer image analytics, including tissue phenotyping, nucleus detection, and nucleus classification problems. His research themes involve developing deep neural networks, subspace learning models, and graph neural networks.

MOHAMMED BENNAMOUN (Senior Member, IEEE) is currently a Professor with the Department of Computer Science and Software Engineering, The University of Western Australia (UWA), Crawley, WA, Australia. He is also a Researcher in computer vision, machine/deep learning, robotics, and signal/speech processing. He has published four books (available on Amazon), one edited book, one Encyclopedia article, 14 book chapters, more than 170 journal articles, more than 280 conference publications, and 16 invited and keynote publications. He won the Best Supervisor of the Year Award from The Queensland University of Technology (QUT), in 1998. He received the Award for Research Supervision from UWA, in 2008 and 2016, and the Vice-Chancellor Award for Mentorship, in 2016.

NAOUFEL WERGHI (Senior Member, IEEE) received the Ph.D. degree in computer vision from the University of Strasbourg, Strasbourg, France, in 1996. He was a Research Fellow with the Division of Informatics, The University of Edinburgh, Edinburgh, U.K., and a Lecturer with the Department of Computer Sciences, University of Glasgow, Glasgow, U.K. He has also been a Visiting Professor with the Department of Electrical and Computer Engineering, University of Louisville, Louisville, KY, USA. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates. His main research interests include image analysis and interpretation, where he has been leading several funded projects in the areas of biometrics, medical imaging, and intelligent systems.