Autonomous Localization of X-Ray Baggage Threats via Weakly Supervised Learning

Divya Velayudhan, *Graduate Student Member, IEEE*, Abdelfatah Ahmed, Taimur Hassan, *Member, IEEE*, Neha Gour, *Member, IEEE*, Muhammad Owais, *Member, IEEE*, Mohammed Bennamoun, *Senior Member, IEEE*, Ernesto Damiani, *Senior Member, IEEE*, and Naoufel Werghi, *Senior Member, IEEE*

***Abstract*—Autonomous X-ray baggage security screening has shown significant strides recently, proving itself a viable solution to the flaws in manual screening, thanks to advancements in deep learning. However, these data-hungry techniques feed on extensively annotated data involving strenuous labor, impeding their advances in baggage screening. Consequently, we present a context-aware transformer for weakly supervised localization to relieve the annotation burden and provide visual interpretability that aids screeners in threat recognition and researchers in identifying the pitfalls of existing systems. The proposed approach can generalize and localize different types of contraband with only cost-effective binary labels without explicit training on item detection. Context extraction block, integrated into the dual-token framework, generates threat-aware context maps, while the token scoring block focuses on minimizing partial activations. Experimental results surpass state of the art (SOTA) methods in terms of classification and localization accuracies. Furthermore, we analyze failures to determine current vulnerabilities and provide new insights for future research.**

***Index Terms*—Aviation, machine learning, threat identification, threat localization, X-ray baggage security.**

I. Introduction

THE rapidly growing global air passenger traffic (estimated to exceed eight billion in 2037 [1]) exacerbates the challenges of aviation security in coping with the evolving risks posed by ingeniously concealed security threats while maintaining faster passenger throughput. As a deterrent, authorities worldwide have reinforced their security measures, in which baggage screening plays an inevitable role in identifying concealed contraband from X-ray scans [2]. However, baggage monitoring is heavily reliant on human expertise, where scans are scrutinized by security personnel to expose illegal items from densely stacked baggage imagery. Hence, the current procedure is not only error-prone but also necessitates constant vigilance to identify the contraband within mere seconds amidst occlusion from high-density objects, overlapping contours, uneven scales, and high inter-class variability [3].

Automated X-ray baggage threat identification systems have been proposed as a viable solution to overcome these pitfalls. Researchers have pursued numerous methods to accomplish the goal, and recent breakthroughs in computer vision aided by deep learning have yielded promising results [4]. Furthermore, in light of suboptimal performance by contemporary detection and segmentation models [5], recent research works in the domain have leveraged low-level edge cues and contour representations of baggage scans [6] to cope with the poor contrast, lack of texture, and occlusion in X-ray imagery [3]. Despite tremendous advances, these data-hungry schemes rely on excessive amounts of well-annotated training data that are expensive to procure. Studies report that dense annotation of security threat instances is laborious and demands skilled operators (taking over 3 min for a single scan) [7], [8].

Within the broader computer vision community, weakly supervised localization has been widely embraced to relieve the burden by exploring cost-effective weak supervisory signals in the form of image labels rather than demanding instance-level bounding boxes and dense pixel labels [9], [14]. In simpler terms, weakly supervised object localization (WSOL) aims to localize an object of interest using only image-level labels during training that merely confirm the presence of the object category. Furthermore, WSOL provides visual interpretability, which is crucial for safety-critical applications, such as baggage security screening, where both performance and explainability are vital. Visual reasoning aids security personnel in locating the threats in the scans effectively. Also, it enables the interpretation of decisions made by the systems, which are crucial for trust-building and deployment [15]. Moreover, it helps minimize potential bias

Manuscript received 9 June 2023; revised 30 October 2023; accepted 14 December 2023. Date of publication 17 January 2024; date of current version 4 April 2024. This work was supported in part by Khalifa University under Grant CIRA-2021-052 and in part by Advanced Technology Research Center Program under Grant ASPIRE AARE20-279. Paper no. TII-23-2107. *(Corresponding author: Divya Velayudhan.)*

Divya Velayudhan, Abdelfatah Ahmed, Neha Gour, Muhammad Owais, Ernesto Damiani, and Naoufel Werghi are with the Center of Secure Cyber-Physical Systems (C2PS), and the Department of Computer Sciences, Khalifa University, Abu Dhabi 127788, UAE (e-mail: divyavelayudhan@gmail.com; 100059689@ku.ac.ae; neha.gour@ku.ac.ae; muhammad.owais@ku.ac.ae; ernesto.damiani@ku.ac.ae; naoufel.werghi@ku.ac.ae).

Taimur Hassan is with the Department of Electrical, Computer, and Biomedical Engineering, Abu Dhabi University, Abu Dhabi 59911, UAE (e-mail: taimur.hassan@adu.ac.ae).

Mohammed Bennamoun is with the Department of Computer Science and Software Engineering, The University of Western Australia, Perth, WA 6907, Australia (e-mail: mohammed.bennamoun@uwa.edu.au).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2023.3348838.

Digital Object Identifier 10.1109/TII.2023.3348838

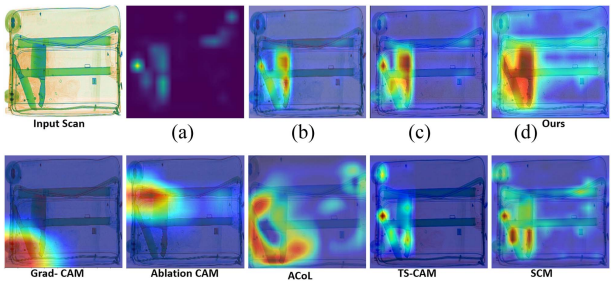

Fig. 1. Visualization of baggage threat localization. The first row displays the output from different blocks of the proposed approach. (a) Context map extracted from CEB. (b) Map overlaid on the input scan, (c) smoothed context map, and (d) final threat localization map. The bottom row compares the results with five different approaches, Grad-CAM [9], Ablation CAM [10], adversarial complementary learning (ACoL) [11], TS-CAM [12], and SCM [13].

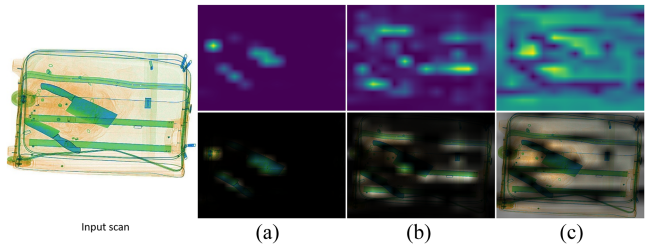

Fig. 2. Visualization of attention maps (and corresponding maps overlaid on the original image) from randomly chosen heads of the transformer to demonstrate that different heads focus on different semantic regions. Column (a) emphasizes the *knives*, while Column (c) emphasizes the *benign* areas. In column (b), attention is on the baggage's knife-like metal bands. Please zoom in for better visualization.

in training and allows the researchers to identify the pitfalls of the system and develop more robust frameworks.

However, the widely favored approaches in weakly supervised learning, being predominantly based on convolutional neural nets (CNN), are constrained to localized regions and fail to capture spatially distant contexts [9]. On the other hand, visual transformers [16], [17] have gained popularity due to their ability to learn global features by exploiting long-range dependencies between semantic concepts. Recently, Gao et al. [12] proposed token semantic coupled attention map (TS-CAM), a pioneering work in WSOL using transformers, in which they attempt to leverage semantic maps from patch tokens for localizing the object of interest. However, the focus is often restricted to a few semantically dense regions, as observed in Fig. 1.

Notwithstanding these advances, WSOL has not yet been widely explored in baggage security threat recognition. This is mainly attributed to the additional challenges for WSOL in this domain.

1) *Occlusion:* Threat items might be fully or partially occluded by high-density objects rendering them indistinguishable.
2) *Heavily cluttered background:* Precise localization of prohibited items from cluttered and compactly packed baggage scans is challenging due to noisy activation maps.
3) *Limited priors:* The unpredictability of baggage contents restricts the availability of prior knowledge for effective localization of threats (unlike in natural images where *food* is more likely to be associated with *plate* and not *sky*).

Toward this goal, we investigate weakly supervised threat localization using transformers by leveraging their ability to model long-range spatial interactions in effectively localizing concealed prohibited items from baggage scans. Furthermore, studies have shown that transformers favor shape over texture (unlike CNNs) and are more robust against occlusion [18]. Consequently, qualifying as an ideal candidate for baggage threat localization from X-ray imagery (since they are texture-less).

Transformers have demonstrated undeniable dominance in several areas [19], thanks to their multiheaded attention that enables them to focus on multiple semantic areas (as shown in Fig. 2). However, the attention is not targeted class-wise [12]. In other words, linking specific attention heads to specific semantic categories or regions is challenging [12], [20]. Furthermore, it is to be noted that in transformers, the class token captures the semantic relationships between different classes and the background, thus learning both category-specific and generic features. The single-class token architecture can, thus, lead to noisy activations, especially in densely stacked baggage scans where overlapping threats and normal objects share similar semantics. Hence, it is essential to redesign the architecture to encode the threat-specific semantics that will enable the effective localization of illegal items in baggage scans. Furthermore, transformer attention focuses on global semantic-rich regions [as seen in Fig. 2(a), where it focuses on the knife blade regions rather than the entire threat object] and fails to capture local structures and boundaries, resulting in partial activation. To address these limitations, we propose a context-aware transformer (CAT) with dual-token architecture that can generalize well to different types of contraband and localize them from only binary image labels (*threat* versus *normal*) by capturing the object-level context of threat items within baggage. The dual-class tokens (threat and normal class tokens) enable the model to learn class-specific interactions with the input patches, thus modeling the global semantics of the concealed threats from the baggage scans. Moreover, we have employed a class-specific training strategy to ensure that the class tokens learn specific semantics (explained in Section III). Furthermore, we have integrated a context map extraction block (CEB) to obtain the object-level semantics of the threat items and a token scoring block (TSB) to expose local features and other relevant occluded regions.

The contributions can be summarized as follows.

1) The first attempt in weakly supervised baggage security threat localization from cost-effective binary image labels (merely confirming the presence of contraband) without explicit training on item detection.
2) The proposed dual-token CAT can generalize well to different types of contraband by capturing their object-level semantics.
3) Experimental results on two challenging public X-ray security screening datasets demonstrate that the proposed

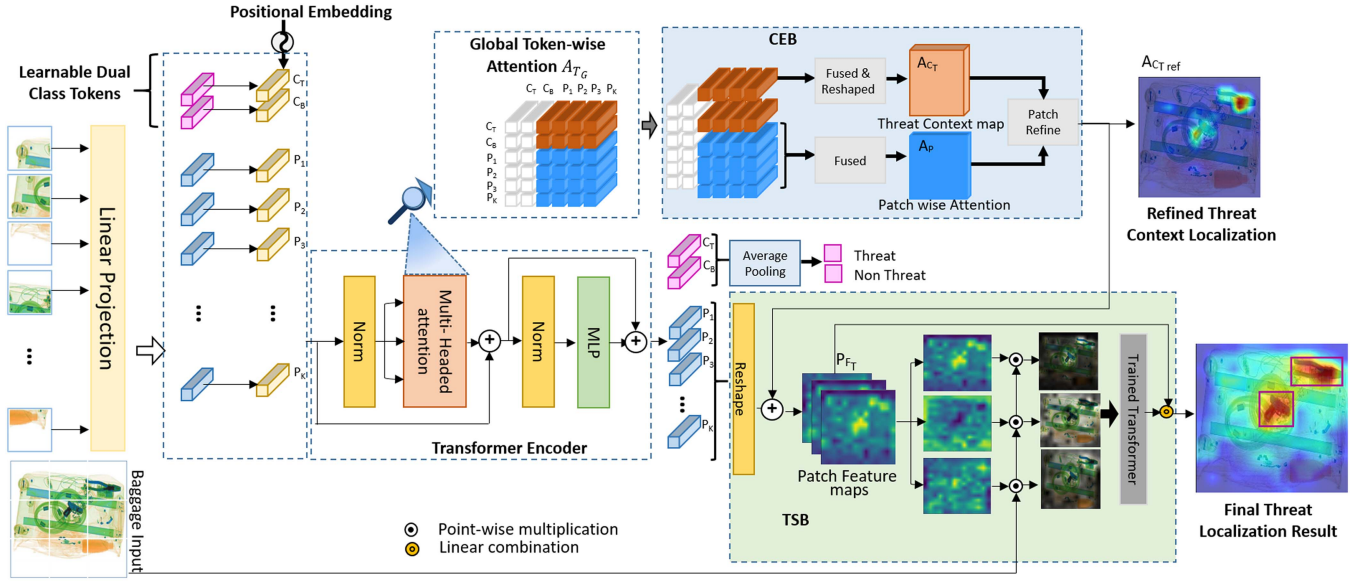

Fig. 3. Overview of the proposed CAT that employs learnable dual-class tokens to localize concealed contraband from only binary labels (*threat versus normal*). CEB captures the global semantics of the prohibited items by leveraging the semantic similarities between the threat-specific class token C_T and the patch tokens. After refining, the context map is passed to TSB to expose other relevant occluded regions.

approach surpasses state of the art (SOTA) methods in terms of classification and localization accuracies.

4) Furthermore, we have provided a detailed analysis of the failure patterns to identify the vulnerabilities of current approaches and offer new insights for future research.

II. Related Works

A. X-Ray Baggage Threat Recognition

Advances in deep learning have shown promise in baggage threat classification [21], detection [2], and segmentation [6]. While transfer learning approaches [5] attempt to improve performance, novel approaches address issues, such as class imbalance [4] and occlusion [2]. Object detection models [5], attention mechanisms [22], and contour cues [2], [6] have also been employed to enhance the results. Chouai et al. [23] used adversarial autoencoders to identify prohibited items in X-ray scans. Wei et al. [24] studied multiscale intermediate features for threat detection. Bhowmik et al. [25] proposed a combined segmentation-classification pipeline to identify occluded illegal items.

B. Weakly Supervised Object Localization

Class activation maps (CAMs) [26] and gradient-based approaches like Grad-CAM [9] have been widely embraced but struggle to highlight entire or integral object parts. Methods like adversarial erasing [11] have been explored to address these limitations. Nevertheless, CNN-based approaches are limited by localized interactions and fail to expand the activations beyond discriminative regions [12], as opposed to the global cues learned by vision transformers [16], [17]. The seminal work by Gao et al. [12] infused transformers with CAMs to highlight relevant object areas. Bai et al. [13] further improved TS-CAM results by dynamic integration of semantic relations. WSOL in baggage threat identification has received limited attention, with Miao et al. [4] being the only study that employed WSOL to analyze the robustness of their framework.

III. Proposed Methodology

This section introduces the proposed architecture of the CAT, illustrated in Fig. 3. Also, it gives a detailed description of CEB and TSB used for threat localization, as well as the implemented training strategy.

A. Overview of CAT

Consider a dataset $D = \{(x_1, y_1), \ldots, (x_M, y_M)\}$ comprised of normal and abnormal (i.e., containing threat items) baggage scans, where $y_i \in [0, 1]$ denotes the binary image labels. A sample image x of resolution $W \times H$ is split into K patches, with $K = N \times N, N = W/s$, and s denoting the patch height/width. The patches $x_{p_n} \in \mathbb{R}^{s\times s\times 3}, n = 1, 2, \ldots, K$ are then transformed into patch embeddings, $x_n \in \mathbb{R}^{K\times D}$, where D indicates the embedding dimension and $\mathcal{F}()$ represents the projection onto embedding subspace in (1). Class tokens $x_{\text{CL}} \in \mathbb{R}^{2\times D}$ are then affixed to the patch embeddings, where $x_{\text{CL}} = [x_{C_T}; x_{C_B}]$) as shown in (1). Here, it is important to highlight the dual-token architecture of the proposed framework as opposed to the standard design. The tokens x_{C_T} and x_{C_B} capture the semantic elements corresponding to the threat and benign (normal) baggage items distinctively, which can be leveraged to learn the global context of concealed contraband. After that, the tokens are updated with positional embeddings $x_{\text{pos}} \in \mathbb{R}^{(2+K)\times D}$ to yield the input tokens $x_{\text{in}} \in \mathbb{R}^{(2+K)\times D}$, as in (1)

$$x_{\text{in}} = [x_{C_T}; x_{C_B}; \mathcal{F}(x_{P_1}); \mathcal{F}(x_{P_2}); \cdots \mathcal{F}(x_{P_K})] \oplus x_{\text{pos}} \quad (1)$$

$$= [C_T; C_B; P_1; P_2; \cdots P_K]\,. \quad (2)$$

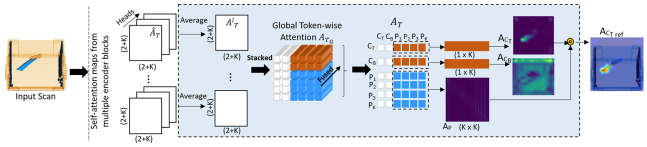

Fig. 4. Detailed schematics of CEB. Please zoom in for better visualization.

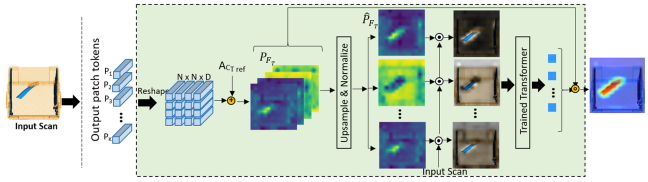

Fig. 5. Detailed schematics of Patch TSB. Please zoom in for better visualization.

These tokens are fed through S stacked transformer blocks, each comprising multiheaded attention [16] and multilayer perceptron layers. As the training progresses, the threat-specific class token C_T [as shown in (2)] learns semantic interactions specific to the prohibited items, which can be leveraged to obtain the global context of the concealed contraband, as detailed in subsequent sections.

B. Context Map Extraction Block

The proposed CEB is tasked with extracting the global semantics of the concealed prohibited items within the baggage imagery to generate the threat-aware context map. CEB exploits the long-range dependencies learned by the self-attention blocks within the transformer encoder toward this goal. Specifically, as the tokens x_{in} propagate through the transformer encoder, it is split into queries Q, keys K, and values V using linear layers [16], where $Q, K, V \in \mathbb{R}^{(2+K)\times D_k}$ and $D_k = D/k$ (k representing the number of attention heads). Inside the multihead attention layer, the conventional dot-product attention technique [16] [as in (3)] is used to compute the attention between Q and K by each of the heads, which is used as weights in combining the input tokens. Assuming Q^l and K^l represent the query and key features from the lth encoder layer, we aggregate the token-wise attentions $\hat{A_T}$ denoted in (4), which are then averaged over the k heads to obtain $A_T^l \in \mathbb{R}^{(2+K)\times(2+K)}$, as illustrated in Fig. 4. A_T^l yielded from multiple encoder layers are stacked to yield the global token-wise similarity map A_{T_G}, and then summed across the encoder layers to generate $A_T \in \mathbb{R}^{(2+K)\times(2+K)}$. In this study, we have leveraged the attention maps derived from the last L encoder blocks to obtain a better localization performance. The optimal value of L has been determined experimentally (see Section V)

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right)V \tag{3}$$

$$\hat{A_T} = \text{softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right). \tag{4}$$

A_T encapsulates the pair-wise attention between the input tokens and can be leveraged to learn the interactions between the threat-specific class token and patch tokens. From the orange-tinged squares, which represent the attention between class tokens and patch tokens, we can derive the threat-relevant context map A_{C_T} [shown in Fig. 1(a)], by leveraging the attention corresponding to the class token C_T. $A_{C_T} \in \mathbb{R}^{1\times N\times N}$ captures the global context of the threat items by reshaping the attention between C_T and the patch tokens $(P_1, P_2 \cdots P_K)$, as shown in Fig. 4, and is used for localizing the concealed threats. Additionally, Fig. 4 also depicts A_{C_B}, which captures the attention between C_B and the patch tokens that focus primarily on the benign areas. The context map A_{C_T} is later refined by exploiting the patch-wise attention learned by the self-attention layers within the encoder. This is a simple approach to smooth the context map over the adjacent local areas based on their semantic affinities. Assuming Q^l and K^l represent the query and key features from the lth encoder layer, we aggregate the patch-wise attentions excluding the class tokens, which are averaged over the k heads to yield $A_p^l \in \mathbb{R}^{K\times K}$. The patch-wise attentions A_p^l are then summed across the S encoder blocks to yield a consolidated attention map A_p, which is used for smoothing the context map A_{C_T} using matrix multiplication. A_{C_T} is flattened prior to the multiplication as denoted by the reshape operator $\Delta^{K\times 1}$ in (5). Afterward, it is reshaped into a 2-D tensor (denoted by $\Delta^{N\times N}$), yielding the refined context map $A_{C_{T\text{ref}}}$

$$A_{C_{T\text{ref}}} = \Delta^{N\times N}\left[A_p \cdot \Delta^{K\times 1}\left(A_{C_T}\right)\right]. \tag{5}$$

Furthermore, it can be observed from the qualitative results (see Fig. 7) that the refined context maps have better continuity than raw context maps A_{C_T}.

C. Patch TSB

The proposed patch TSB [illustrated in Fig. 5] is responsible for exposing additional pertinent and occluded threat object regions by employing a perturbation technique. TSB expertly adapts Score-CAM [27] for vision transformers and helps to reveal other salient parts while preserving the context captured by the CEB. Score-CAM was proposed to localize objects by grasping the relevance of the feature maps in CNNs. However, in transformers, Score-CAM often highlights irrelevant background regions and semantically similar benign regions, leading to noisy localization, as detailed in Section V.

Consider the patch tokens $\{P_1, P_2, \ldots, P_K\}$ at the output of the encoder block, excluding the class tokens $\{C_T, C_B\}$. The patch tokens are reshaped and transposed into a feature tensor $P_F \in \mathbb{R}^{N\times N\times D}$, where each feature map $P_F^d, d \in \{1, 2, \ldots, D\}$ emphasizes distinct semantic elements. The smoothed context map $A_{C_{T\text{ref}}}$ obtained from CEB is then added to P_F to suppress the activation of background elements and other benign items. Thus yielding $P_{F_T} \in \mathbb{R}^{N\times N\times D}$, which is then upsampled to the size of the original input x and normalized [as shown in (7)]

$$P_{F_T} = A_{C_{T\text{ref}}} \oplus P_F \tag{6}$$

$$\hat{P_{F_T}} = \frac{P_{F_T} - \min(P_{F_T})}{\max(P_{F_T}) - \min(P_{F_T})}. \quad (7)$$

The normalized maps $\hat{P_{F_T}}$ are overlaid using element-wise multiplication on the scan x to obtain partially masked images. By doing so, we emphasize different regions in the input scan based on feature maps obtained from the output tokens. The masked images are then simply forward passed through the trained dual-token transformer model with the objective of determining the impact of retaining only the highlighted regions on the target score. The forward passing yields the scores on the threat category for the partially masked images, which are converted to weights via softmax. The weights are subsequently used to linearly combine the respective feature maps, thus yielding the final localization result. The localization maps are transformed into binary masks by retaining only the pixels that exceed a preset threshold. Then, the bounding boxes are drawn to localize the object, as shown in [26].

D. Dual Token Training Method

The output tokens generated by the final encoder block comprise patch tokens and two class tokens C_T and C_B. Establishing a one-to-one relationship between each class token and its corresponding ground truth labels is crucial. The final class tokens C_T and C_B are averaged along the embedding dimension by passing through a pooling layer to obtain the scores of the threat and benign classes, as

$$y(c) = \frac{1}{D}\sum_{l}^{D} C_{\text{Tok}}(c, l), c \in [0, 1] \quad (8)$$

where $C_{\text{Tok}}(c, l)$ is lth element of the cth token, along the D dimension. The scores and one-hot encoded labels are used for training the framework using binary cross entropy, thus facilitating the model to learn distinctive semantic correlations specific to the categories.

IV. Datasets and Implementation Details

This section details the experimental settings, including a summary of the datasets utilized and the performance evaluation measures employed to assess the proposed approach.

A. Datasets

The proposed weakly supervised threat localization framework was evaluated on two challenging X-ray baggage security datasets: security inspection X-ray (SIXray) [4] and Compass-XP [28]. SIXray and Compass-XP are the only public datasets in X-ray baggage security screening that contain both positive (threat) and normal (benign) scans.

SIXray dataset, the largest baggage security benchmark, is a challenging dataset with extremely obscure and cluttered baggage scans. It comprises 1 059 231 pseudocolored scans, of which only 8929 scans include threats. The dataset includes six classes of prohibited objects: guns, scissors, knives, wrenches, pliers, and hammers. The dataset is also challenging in terms of occlusion, with less than 1% marked as positive (Threat), and has high intracategorical variations in terms of scale and viewpoint. It is organized into three subdivisions: SIXray10, SIXray100, and SIXray1000. We have used SIXray10 in our experiments. Furthermore, the baggage scans are collected from different scanners, which raises additional challenges.

Compass-XP dataset is unique with different representations (low-energy and high-energy X-ray, density, color, grayscale, and RGB images) for the same baggage scan and is highly imbalanced. It is comprised of paired images, both photographic and X-ray images. The dataset includes a total of 11 568 scans, which includes 1643 pairs of benign images and 258 pairs of threat images (comprising 69 types of threat objects), with very few samples of each threat category. We used 80% of the scans for training, per the dataset protocol.

B. Implementation

The proposed framework for weakly supervised baggage threat localization was built using the DeiT-S backbone [17], pretrained on ImageNet, with $K = 196$, $s = 16$, $D = 384$, $S = 12$, and six attention heads. The dual-class tokens in the proposed architecture were initialized with the ImageNet weights of the original class token. The input scans were resized to 224×224 for all experiments. Simple data augmentation techniques, horizontal, and vertical flipping, were employed during training. The training was carried out for 15 epochs with a learning rate of 2e$-$5 and a batch size of 8. The framework was implemented with PyTorch using Python 3.8 on an Intel(R) Core(TM) i7-10700K@ 3.80 GHz processor with NVIDIA GeForce RTX 3060 Ti.

C. Evaluation Metrics

Following the prior works in the domain [11], [12], we adopt the following metrics for evaluating the proposed approach.

1) GT-Known localization accuracy (GT-K Loc.): Localization is counted as positive if the intersection-over-union (IoU) between the ground truth and predicted bounding boxes exceeds the fixed threshold of 50%.
2) Top localization accuracy (Top Loc.): If correctly classified and the IoU between the ground truth and predicted bounding boxes exceeds the fixed threshold of 50%, then the localization is counted as positive.
3) MaxBoxAccV2: Irrespective of the classification results, the localization accuracy is averaged across different IoU thresholds (30%, 50%, 70%) to yield MaxBox AccV2.
4) Classification accuracy (Cls Acc): Represents the classification performance.

In addition, we have also computed localization metric (Loc) as in [4]. Loc is considered positive if the maximal response coincides with one of the ground truth boxes.

V. Experimental Analysis and Results

We have compared our framework with SOTA strategies in WSOL, as well as other relevant baggage threat localization approaches. In addition, we present several ablative studies to demonstrate the efficacy of the proposed architecture.

TABLE I
COMPARATIVE ANALYSIS ON SIXRAY

Method	Backbone	Cls Acc	GT-K Loc.	Top Loc.	MaxBoxAccV2	Loc
*Grad-CAM (CVPR 17) [9]	ResNet50	93.2	20.21	17.71	19.2	-
*Ablation CAM (WACV 20) [10]	ResNet50	93.2	20.08	20.21	17.9	-
*ACoL (CVPR 18) [11]	VGG16	88.3	22.39	18.12	23.3	-
*TS-CAM (ICCV 21) [12]	Deit-S	85.9	28.34	24.82	25.9	-
*SCM (ECCV 22) [13]	Deit-S	80.6	32.78	28.05	33.8	-
CHR [4]	Inception-V3	79.4	-	-	-	63.5
Ours	Deit-S	**96.6**	**34.63**	**30.35**	**35.7**	**89.2**

The bold values signify the best results.

TABLE II
COMPARATIVE ANALYSIS ON COMPASS-XP

Method	Backbone	Cls Acc	GT-K Loc.	Top Loc.	MaxBoxAccV2
*Grad-CAM (CVPR 17) [9]	ResNet50	88.3	37.25	35.29	28.0
*Ablation CAM (WACV 20) [10]	ResNet50	88.3	33.30	27.45	26.1
*ACoL (CVPR 18) [11]	VGG16	86.7	41.20	39.21	29.3
*TS-CAM (ICCV 21) [12]	Deit-S	89.1	49.01	45.09	47.6
*SCM (ECCV 22) [13]	Deit-S	81.4	54.91	41.18	59.6
Ours	Deit-S	**94.7**	**76.47**	**70.59**	**65.3**

The bold values signify the best results.

TABLE III
COMPARATIVE ANALYSIS OF MODEL COMPLEXITY

Method	Model	# Parameters (M)	MACs (G)	Inference time (sec) (for threat localization)
Grad-CAM [9]	ResNet50	23.51	4.13	0.34
Ablation CAM [10]	ResNet50	23.51	4.13	5.90
ACoL [11]	VGG16	138.3	15.5	2.05
TS-CAM [12]	Deit-S	21.67	4.25	1.05
SCM [13]	Deit-S	21.67	4.25	1.15
Ours	CAT	21.67	4.26	1.07

A. Comparative Performance Analysis

Table I showcases the performance of the proposed CAT against state-of-art approaches in WSOL on SIXray. Within the context of X-ray baggage screening, CHR presented by Miao et al. [4] is the only research work that investigated weakly supervised threat localization and evaluated the localization efficacy of their framework using the Loc metric. Hence, we have also computed the Loc metric for a fair comparison, outperforming CHR by ∼25%, demonstrating the competence of the proposed approach in recognizing threat objects. In addition, we have also compared with SOTA approaches, which include gradient-based (Grad-CAM [9]), gradient-free (Ablation CAM [10]), adversarial erasing (ACoL [11]), and transformer-based (TS-CAM [12], SCM [13]) techniques. The results for these approaches, shown in Table I, were obtained by training them on the SIXray dataset using the respective released codes and thus denoted by an asterisk (*).

Overall, transformer-based methods lead the scoreboard thanks to their ability to retain long-range semantic relations, which is crucial in localizing the threats. The proposed approach outperforms TS-CAM by a large margin of 6.33%, 5.53%, and 9.8% in terms of GT-K Loc, Top Loc, and MaxBoxAccV2, respectively. Furthermore, compared with SCM, the leading competitor, our approach leads by 1.85%, 2.3%, and 1.9% in terms of GT-K Loc, Top Loc, and MaxBoxAccV2, respectively. Nonetheless, SCM performs inadequately in terms of Cls Acc, scoring only 80.6%, compared to the 96.6% accuracy achieved by our approach.

Similarly, Table II compares the performance on Compass-XP. The results of the SOTA methods, marked "*" in Table II, were obtained using their respective publicly released codes and training them on the Compass-XP dataset [28]. The proposed approach achieved remarkable performance across all metrics, outperforming the best competitive method, SCM, by large margins of ∼29%, ∼37%, and ∼12% in terms of GT-K Loc, Top Loc, and MaxBoxAccV2, respectively. The inadequate performance of other approaches on Compass-XP is primarily attributed to their inability to model the object-level context (as can be observed in Fig. 6). TSCAM [12] fails to capture the object-level context, resulting in partial activation, while SCM [13] performs better due to its capability to capture contextual and spatially coherent threat object regions. However, SCM yields poor classification results, drastically lowering the Top Loc results. On the contrary, the proposed CAT surpasses other approaches in terms of Cls Acc by a large margin on both SIXray and Compass-XP.

Furthermore, to provide a comprehensive understanding of the performance of our proposed framework, we compared and contrasted the model complexity in terms of the number of parameters, multiply accumulate operations (MACs), and inference time. Table III shows that our approach has a relatively lower number of parameters, and while the inference time is slightly greater than Grad-CAM, it is important to highlight that the localization accuracy of our proposed method significantly outperforms all these approaches. This demonstrates that our approach balances model complexity and localization accuracy, offering a competitive solution for X-ray baggage threat localization.

Fig. 6 depicts the localization maps using different techniques on both SIXray (top five rows) and Compass-XP (bottom three rows). We have used purple bounding boxes to highlight the baggage threats in the input scans on the left-most column. It can be observed that only the proposed CAT succeeds in localizing the knife that overlays the metal band in the top row, as well as all three threats in the second row. Similarly, all methods fail to localize both guns in the third row. Furthermore, the last two rows demonstrate that CAT localizes the entire object, unlike its competitors.

B. Significance of CEB and TSB

CEB compiles the global contextual information and generates the threat-aware context map, while the TSB exposes other relevant occluded object regions and also focuses on the local areas, thus minimizing partial activations. Fig. 7 illustrates the extracted global context map, smoothed context map derived from CEB, and the final result after TSB, demonstrating the roles of the two modules. In the top row, CEB captures the two guns and the knife but fails to identify the second occluded knife, which is localized by TSB. Similarly, in the bottom row, TSB exposes the complete threat objects, which were partially localized by CEB.

Furthermore, to analyze the relevance of CEB and TSB, we have assessed the performance of the framework both quantitatively and qualitatively by excluding each of the modules. Table IV reports the GT-K Loc and Top Loc accuracies with CEB and TSB alone on both SIXray and Compass XP. It may be noted that using CEB alone can yield comparative performance, being

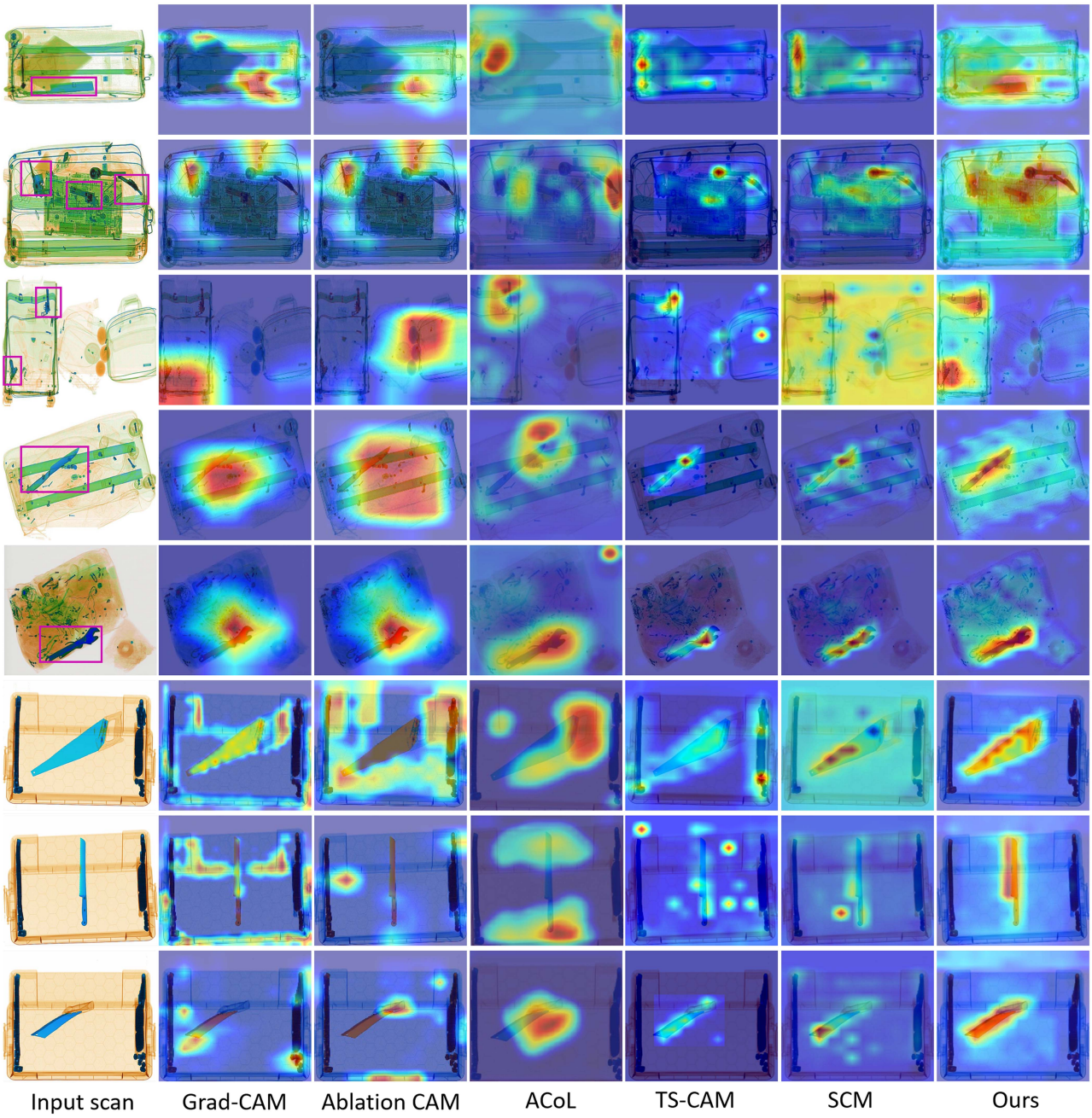

Fig. 6. Visualization of baggage threat localization using different methods on SIXray [4] and Compass-XP datasets [28].

TABLE IV
COMPARATIVE ANALYSIS OF THE SIGNIFICANCE OF CEB AND TSB IN THE PROPOSED METHOD

Dataset	Method	GT-K Loc	Top Loc
Compass-XP	CEB Only	68.6	62.7
	TSB Only	64.7	58.8
SIXray	CEB Only	32.1	27.6
	TSB Only	22.3	20.4

able to capture relevant threat semantics. However, it sometimes fails to localize occluded threats and entire threat objects, as shown in Fig. 8. There are five threats in the top row, and without TSB, CEB misses the second gun (shown in purple) due to its unique orientation. In the second row, CEB fails to capture the second occluded knife (shown with a purple bounding box) without TSB. On the other hand, using TSB alone increases the background noise, highlighting semantically comparable regions. However, incorporating both modules together yields a better result, as illustrated in Fig. 8.

C. Optimal Number of Encoder Blocks

As detailed in Section III-B, CEB leverages the attention between the class tokens and patch embeddings from the last L encoder blocks to generate the threat-aware context map. For determining the optimal number of encoder blocks, we have

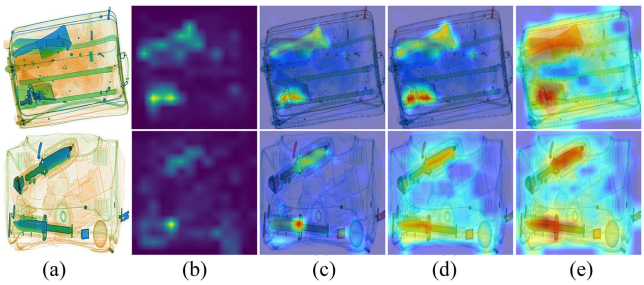

Fig. 7. Visualization of localization maps at different stages of the proposed approach. (a) Baggage scans. (b) Threat-aware context maps, (c) Overlaid context map on the input scan. (d) Smoothed context map from CEB, and (e) Final threat localization map after TSB.

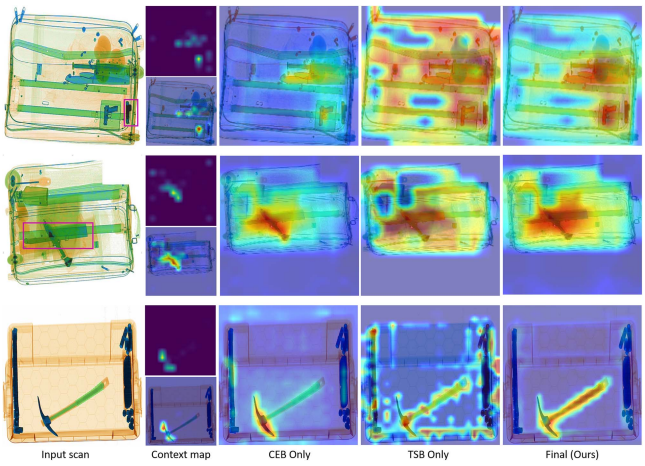

Fig. 8. Qualitative analysis to study the significance of CEB and TSB. Input scans are shown in first column, while raw context maps are shown in next column. Third column shows the refined context maps obtained from CEB. Fourth column shows the output with only TSB. Proposed framework output (with CEB and TSB) is in last column.

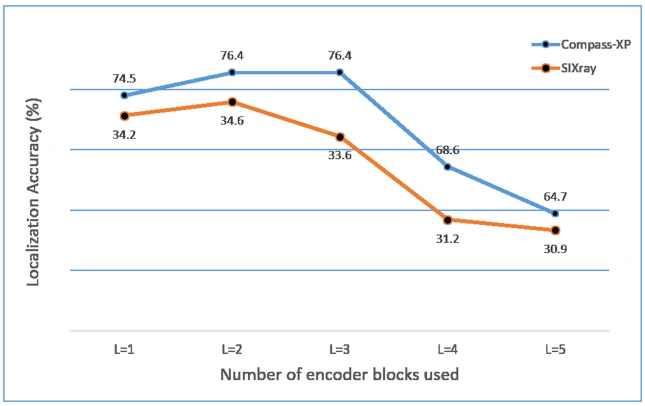

Fig. 9. Localization performance analysis by varying the number of encoder layers.

stacked and summed the attention from multiple higher order transformer encoder blocks and assessed the variation in performance. Fig. 9 analyzes the performance using the top L encoders and shows that the localization accuracy slightly increases and then drops as more encoder layers are added, indicating that lower order encoder blocks learn generalized representations and amplify the background activations. We have used the top two layers in our experiments based on the results on SIXray and Compass-XP.

TABLE V
COMPARATIVE ANALYSIS OF CLASS SCORE PREDICTION STRATEGIES ON PERFORMANCE

	SIXray			Compass-XP		
	Cls Acc	F1	GT-K Loc.	Cls Acc	F1	GT-K Loc.
Dense Layer (MLP)	94.3	83.2	32.38	93.8	85.1	70.59
Average pooling	96.6	85.7	34.63	94.7	86.8	76.47

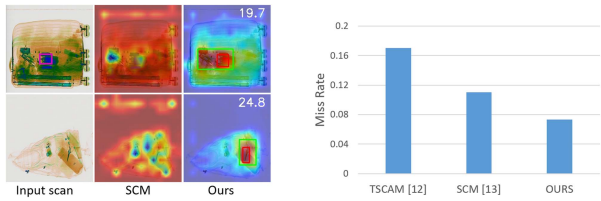

Fig. 10. (Left) Sample scans with effective threat localization despite lower IoU values. Please zoom in. (Right) Comparative analysis of Miss rate (missed threats).

D. Class Score Prediction Strategies

Furthermore, we assessed the effect of different strategies in converting the class tokens to class scores on both datasets using Cls Acc, F1 score, and GT-K Loc. Table V reveals that GT-K Loc accuracy significantly improves when we utilize average pooling as opposed to linear projection through a multilayer perceptron, with an improvement of 2.2% and 5.8% on SIXray and Compass-XP, respectively. Furthermore, adding a dense layer to the model increases complexity in terms of model parameters, reaffirming our initial design rationale.

VI. DISCUSSIONS AND FUTURE DIRECTIONS

Despite the overall notable performance of the proposed CAT, Table I shows comparatively lesser quantitative results on SIXray than Compass-Xp (Table II). Visual analysis showed that it was due to the activation of semantically related background pixels in highly cluttered imagery in SIXray [4]. Sample scans in Fig. 10 (left) demonstrate effective threat localization but yield IoU below 50%, impacting localization performance. The proposed method highlights the sharp knife-like implement along with the scissors in the top row and also localizes the scissors in the bottom row. In contrast, SCM, the next leading competitor, fails and classifies the scans as *Normal*. However, the activation of the background lowers the IoU, yielding zero localization. Furthermore, it may be noted that within the context of baggage security screening, identifying concealed threats is more crucial than determining object boundaries precisely.

Furthermore, for a detailed analysis of the missed threat instances, we computed the miss rate, given as $\frac{\text{Number of missed threat instances}}{\text{Total number of baggage threats}}$. The proposed approach yields the lowest miss rate of 0.073% among its best competitors, as shown in Fig. 10 (right).

We also investigate the failed localization (missed threats) cases to aid future research studies in baggage screening. Sample scans where the threats were not localized are presented in Fig. 11, along with results from other methods for comparison.

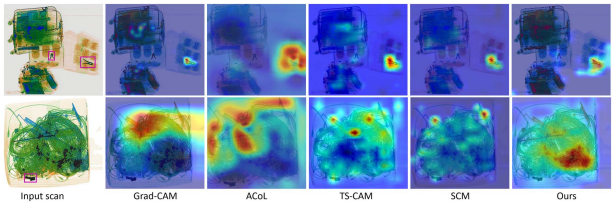

Fig. 11. Failed cases (threats were missed). Please zoom in.

All the methods missed one of the pliers in the top row, while they failed to locate the gun from the heavily cluttered scan in the bottom row. Detailed analysis revealed that the missed threats were often very small and heavily occluded. For instance, the ratio of the threat object area to the total scan area for the samples in Fig. 11 was 0.38% and 0.73%. A recent study [29] has also compared several detection models and ascertained that false negatives tend to occupy a smaller area and are quite challenging. Our analysis shows that transformers are better suited for detecting small baggage threats due to the limitations of downsampling and pooling in vanilla networks [30]. Further transformers are better capable of leveraging the shape information from texture-less X-ray scans. The integration of the context extraction block (CEB) and Patch TSB within our proposed framework further enhances the model's context-awareness, contributing to the model's ability to focus on critical areas within the scan, aiding in the localization of threats while effectively handling occlusion. In addition, other studies have also quantified the substantial performance improvement attained by transformers [29]. Furthermore, boosting feature resolution and multiscale-aware training strategies can enhance the efficacy of baggage threat detection performance.

VII. Conclusion

In this work, we propose CAT for localizing X-ray baggage threats using only binary image labels, without explicit training on threat detection. The proposed framework can generalize well to different types of contraband by capturing the object-level semantics of the threat items. The integrated CEB generates threat-aware context maps, while the patch TSB exposes other relevant occluded threat objects. The experimental results on two challenging public datasets demonstrate the efficacy of the proposed approach. Furthermore, a detailed analysis of failure patterns is provided to identify the vulnerabilities of current approaches, aiming to provide new insights for future research in the domain.

In conclusion, our research represents a significant advancement in the field of X-ray baggage security screening, particularly in real-world settings where the emergence of new security threats necessitates the labor-intensive task of annotating thousands of samples of these threats to train robust frameworks, straining security operations and impeding the ability to effectively address evolving threats. Our work presents a promising alternative by leveraging cost-effective weak supervisory image labels, thereby reducing the burden of dense annotations. This enables faster response and adaptability to new threat categories in a rapidly changing security landscape. Furthermore, transparency and interpretability provided by our approach contribute to trust-building and the seamless deployment of security systems. Furthermore, the proposed framework offers the potential to be applied to various domains beyond X-ray baggage security screening. For instance, in fields such as industrial defect localization, where annotating data is time-consuming and laborious, the CAT framework can effectively leverage weak labels to pinpoint abnormalities. This demonstrates the scalability and broader applicability of our approach, making it a valuable tool in multiple safety-critical and industrially significant domains.

References

[1] A. de Juniac, "Aviation security amid evolving threats," 2019. [Online]. Available: https://www.iata.org/en/pressroom/pressroom-archive/2019-speeches/2019-02-27-01/

[2] 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.*, 2020, pp. 2016–2020.

[3] 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, 2022, Art. no. 165.

[4] C. Miao et al., "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.*, 2019, pp. 2119–2128.

[5] 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.*, 2017, pp. 1337–1341.

[6] T. Hassan and N. Werghi, "Trainable structure tensors for autonomous baggage threat detection under extreme occlusion," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 257–273.

[7] R. Tao et al., "Towards real-world X-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10923–10932.

[8] 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.*, 2021, pp. 5412–5421.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[10] S. Desai and H. G. Ramaswamy, "Ablation-CAM: Visual explanations for deep convolutional network via gradient-free localization," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 983–991.

[11] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang, "Adversarial complementary learning for weakly supervised object localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1325–1334.

[12] W. Gao et al., "TS-CAM: Token semantic coupled attention map for weakly supervised object localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2886–2895.

[13] H. Bai, R. Zhang, J. Wang, and X. Wan, "Weakly supervised object localization via transformer with implicit spatial calibration," in *Proc. 17th Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, Springer, 2022, pp. 612–628.

[14] H. Liu, Z. Liu, W. Jia, D. Zhang, and J. Tan, "A novel imbalanced data classification method based on weakly supervised learning for fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 18, no. 3, pp. 1583–1593, Mar. 2022.

[15] S. Singla, B. Nushi, S. Shah, E. Kamar, and E. Horvitz, "Understanding failures of deep networks via robust feature extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12853–12862.

[16] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[17] 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.

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

[19] X. Zang, G. Li, and W. Gao, "Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval," *IEEE Trans. Ind. Informat.*, vol. 18, no. 12, pp. 8776–8785, Dec. 2022.

[20] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, and D. Xu, "Multi-class token transformer for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4310–4319.

[21] D. Mery, E. Svec, M. Arias, V. Riffo, J. M. Saavedra, and S. Banerjee, "Modern computer vision techniques for X-ray testing in baggage inspection," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 47, no. 4, pp. 682–692, Apr. 2017.
[22] 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.
[23] M. Chouai, M. Merah, J.-L. Sancho-Gómez, and M. Mimi, "Supervised feature learning by adversarial autoencoder approach for object classification in dual X-ray image of luggage," *J. Intell. Manuf.*, vol. 31, pp. 1101–1112, 2020.
[24] Y. Wei, Z. Zhu, H. Yu, and W. Zhang, "AFTD-net: Real-time anchor-free detection network of threat objects for X-ray baggage screening," *J. Real-Time Image Process.*, vol. 18, no. 4, pp. 1343–1356, 2021.
[25] 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.*, 2022, pp. 1463–1467.
[26] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.
[27] H. Wang et al., "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 24–25.
[28] M. Caldwell and L. D. Griffin, "Limits on transfer learning from photographic image data to X-ray threat detection," *J. X-ray Sci. Technol.*, vol. 27, no. 6, pp. 1007–1020, 2019.
[29] B. Issac-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*, 2023, pp. 524–533.
[30] G. Chen et al., "A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 52, no. 2, pp. 936–953, Feb. 2022.

Divya Velayudhan (Graduate Student Member, IEEE) received the B.Tech. degree in electronics and communication engineering from the Cochin University of Science and Technology, Kochi, India, in 2005, and the M.Tech. degree in electronics and communication engineering from the University of Kerala, Thiruvananthapuram, India, in 2016. She is currently working toward the Ph.D. degree in electrical and computer engineering with Khalifa University, Abu Dhabi, United Arab Emirates.

Her research interests include signal processing, image analysis, and computer vision.

Abdelfatah Ahmed received the B.Sc. degree in electrical and electronic engineering in 2020 from the University of Sharjah, Sharjah, United Arab Emirates, and the M.Sc. degree in electrical and computer engineering in 2022 from Khalifa University, Abu Dhabi, United Arab Emirates, where he is currently working toward the Ph.D. degree in electrical and computer engineering.

His research interests include computer vision, deep learning, and speech signal processing.

Taimur Hassan (Member, IEEE) received the Ph.D. degree in computer engineering from the National University of Sciences and Technology, Islamabad, Pakistan, in 2019.

He is currently an Assistant Professor with Abu Dhabi University, Abu Dhabi, United Arab Emirates. His research interests lie in the fields of computer vision, medical imaging, and robotics.

Dr. Hassan was a recipient of many national and international awards throughout his career.

Neha Gour (Member, IEEE) received the B.Tech. degree from Jabalpur Engineering College, Jabalpur, India, in 2013 and the M.Tech. degree and Ph.D. degrees from the Indian Institute of Information Technology, Design, and Manufacturing, Jabalpur, India, in 2016 and 2022, respectively.

She is currently a Postdoctoral Fellow with the Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, United Arab Emirates, specializing in deep learning-based medical image analysis, pattern recognition, and computer vision.

Muhammad Owais (Member, IEEE) received the B.S. and M.S. degrees in computer engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2014 and 2016, respectively, and the Ph.D. degree in electronics and electrical engineering from Dongguk University, Seoul, South Korea, in 2022.

He is currently a Postdoctoral Fellow with the Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, United Arab Emirates, specializing in deep learning-based medical image analysis, biomedical 2-D/3-D imaging data processing, pattern recognition, and image recognition.

Mohammed Bennamoun (Senior Member, IEEE) received the M.Sc. degree in control theory from Queen's University, Kingston, Ontario, Canada, and the Ph.D. degree in computer vision from the Queensland University of Technology, Brisbane, Australia.

He is currently a Winthrop Professor with the Department of Computer Science and Software Engineering, The University of Western Australia (UWA), Perth, WA, Australia, where he is a Researcher in Computer Vision, Machine/Deep Learning, Robotics, and Signal/Speech Processing. He has authored or coauthored four books, 14 book chapters, more than 200 journal articles, more than 270 conference publications, and 16 keynote publications.

Mr. Bennamoun was the recipient of more than 65 competitive research grants from the Australian Research Council and numerous other Government, UWA, and Industry Research Grants.

Ernesto Damiani (Senior Member, IEEE) received the M.Sc. degree in electronics engineering from Universita degli Studi di Pavia, Pavia, Italy, in 1987, and the Ph.D. degree in computer science from Universita degli Studi di Milano, Milan, Italy in 1994.

He is currently a Director of the Center for Cyber-Physical Systems (C2PS), Khalifa University, Abu Dhabi, United Arab Emirates, and a Full Professor with the Università degli Studi di Milano, Milan, Italy. He has more than 670 publications listed on DBLP. His research interests include cyber-physical systems, AI/ML, cybersecurity, big data, and cloud/edge processing.

Dr. Damiani is considered among the most proliic European computer scientists.

Naoufel Werghi (Senior Member, IEEE) received the Habilitation and Ph.D. degrees in computer vision from the University of Strasbourg, Strasbourg, France, in 2000.

He is currently a Full Professor with the Computer Science Department, Khalifa University, Abu Dhabi, United Arab Emirates. He is the theme leader of the Artificial Intelligence and Big Data Pipelines theme in the Cyber-Physical Security System Center (C2PS), Khalifa University. His research interests span computer vision and machine learning, where he has been leading several funded projects related to biometrics, medical imaging, remote sensing, surveillance, and intelligent systems.