Performance Evaluation of Segment Anything Model with Variational Prompting for Application to Non-Visible Spectrum Imagery

Yona Falinie A. Gaus[1], Neelanjan Bhowmik[1], Brian K. S. Isaac-Medina[1], Toby P. Breckon[1, 2]
Department of {[1]Computer Science, [2]Engineering}, Durham University, UK

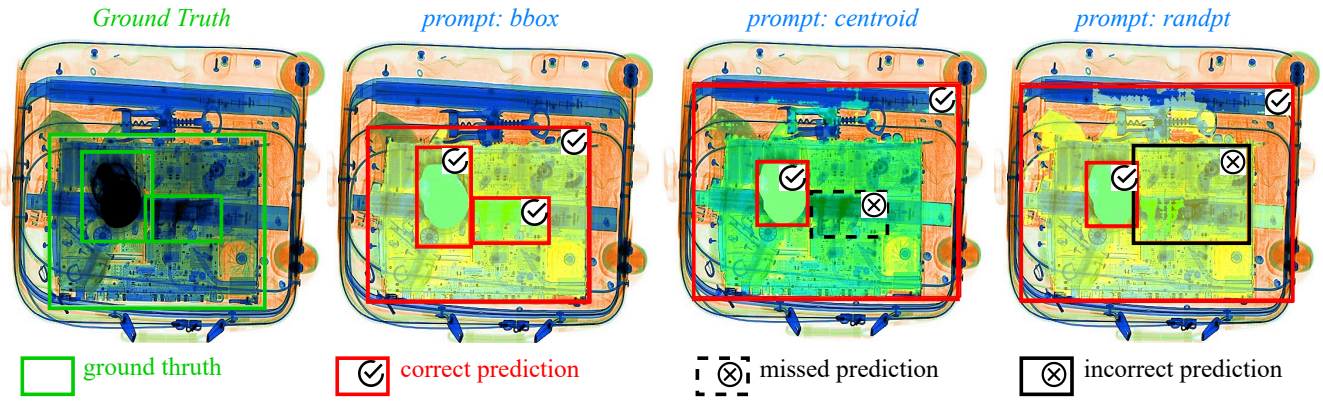

Figure 1. We propose evaluating three prompting strategies (bounding box - *bbox*, *centroid*, random point - *randpt*) to assess the effectiveness of the Segment Anything Model applied to X-ray and infrared imagery for identifying objects of interest. The bbox prompt yields superior segmentation results, while the other two prompting strategies demonstrate notably higher incorrect/missed predictions.

Abstract

The Segment Anything Model (SAM) is a deep neural network foundational model designed to perform instance segmentation which has gained significant popularity given its zero-shot segmentation ability.SAM operates by generating masks based on various input prompts such as text, bounding boxes, points, or masks, introducing a novel methodology to overcome the constraints posed by dataset-specific scarcity. While SAM is trained on an extensive dataset, comprising more than $11M$ *images, it mostly consists of natural photographic (visible band) images with only very limited images from other modalities. Whilst the rapid progress in visual infrared surveillance and X-ray security screening imaging technologies, driven forward by advances in deep learning, has significantly enhanced the ability to detect, classify and segment objects with high accuracy, it is not evident if the SAM zero-shot capabilities can be transferred to such modalities beyond the visible spectrum. For this reason, this work comprehensively assesses SAM capabilities in segmenting objects of interest in the X-ray and infrared imaging modalities. Our approach reuses and preserves the pre-trained SAM with three different prompts, namely bounding box, centroid and random points. We present several quantitative and qualitative results to showcase the performance of SAM on selected datasets. Our results show that SAM can segment objects in the X-ray modality when given a box prompt, but its performance varies for point prompts. Specifically, SAM performs poorly in segmenting slender objects and organic materials, such as plastic bottles. Additionally, we find that infrared objects are also challenging to segment with point prompts given the low-contrast nature of this modality. Overall, this study shows that while SAM demonstrates outstanding zero-shot capabilities with box prompts, its performance ranges from moderate to poor for point prompts, indicating that special consideration on the cross-modal generalisation of SAM is needed when considering use on X-ray and infrared imagery.*

1. Introduction

In the domain of security, the strategic deployment of advanced imaging technologies holds pivotal significance, contributing significantly to safeguarding national borders,

airports, public facilities, transportation systems, and national infrastructure. Infrared-band camera imagery is well established within visual surveillance, offering extensive applications in target detection, visual tracking, behaviour analytics, home monitoring, and automotive environment perception [9, 31–33, 36, 49, 51]. In the domain of X-ray imaging, X-ray security screening stands as a widely utilised method in aviation and broader transportation sectors, detecting prohibited items by scrutinising X-ray images of baggage, freight, and postal items [18, 54, 64].

The recent rise of Convolutional Neural Networks (CNN) [29] has revolutionised visual tasks significantly advancing the state-of-the-art in image understanding technologies. Within object detection, most efforts have focused on detecting objects-of-interest in standard colour imagery by using multi-stage [22, 61, 65], single-stage [47, 59, 62] and transformer-based [11, 60] detectors. These aforementioned object detection based CNN methods rely heavily on architectures that have been trained on large-scale colour imagery datasets such as ImageNet [14]. Introducing CNN to object detection within infrared and X-ray imagery is significantly hindered by the absence of such annotated datasets of the same scale and variety [30, 52, 63].

To address the challenges in the field of infrared imagery analysis, methods such as transfer learning [19], generation of pseudo-RGB equivalents [15], and domain adaptation [43] have been utilised to enhance existing CNN models and establish public benchmarks for research. Gaus et al. [19] concentrate on detecting objects in infrared imagery by leveraging a transfer learning approach, where the knowledge obtained from the visible spectrum is transferred to the infrared domain. Devaguptapu et al. [15] employ image-to-image translation techniques to create pseudo-RGB versions of infrared images. These pseudo-RGB are then processed using CNN models for object detection in infrared imagery. Munir et al. [43] introduce a method of self-supervised domain adaptation through an encoder-decoder transformer network to develop a robust infrared image object detector in autonomous driving.

This methodology parallels efforts in the broader domain of X-ray security imaging, where several benchmarks for security inspection have been developed [41, 42, 55, 57]. Public datasets such as GDXray [41], SIXray [42], PIDray [57] or OPIXray [55] have been released, where the main goal is to advance the developments of prohibited item detection in X-ray images using CNN based methods [3, 12, 20, 37, 50, 54].

The performance of these CNN-based models heavily depends on the availability of suitable infrared and X-ray imagery datasets with sufficient object annotations, diversity and scale, which has often been lacking compared to visible imagery resources. A common solution to address this issue involves pre-training, which effectively utilizes a limited volume of target dataset (X-ray and infrared imagery). However pre-training on datasets beyond visible imagery could lead to dataset bias, potentially misaligning with the idiosyncratic characteristics of the target dataset, which significantly diverges from visible datasets [25]. For example, X-ray datasets consist of semi-transparent transmission imagery, where objects appear translucent and blend visually from front to back, unlike visible images where foreground objects occlude those in the background. Conversely, infrared images are not influenced by variations in visible spectrum illumination and shadows, illustrating the unique challenges and differences each imaging modality presents when compared to standard visual datasets. To address these challenges, the development of comprehensive datasets enriched with detailed annotations in non-visible spectrum imagery becomes essential. This approach serves not only to enhance model training and performance but also to ensure broader applicability and effectiveness across varying imaging modalities. In this context, developing foundation models [8] and zero-shot learning [56] techniques can significantly alleviate the common challenges for datasets of different modalities. Foundation models are neural networks that undergo training on an extensive body of data, utilising innovative learning and prompting strategies that generally bypass the need for conventional supervised training labels. This approach enhances their capability to apply zero-shot learning to entirely new datasets across diverse settings.

Whilst foundation models have revolutionised the field of natural language processing [10, 27, 45], the Segment Anything Model (SAM) [28] has demonstrated promising zero-shot segmentation capabilities across multiple datasets of natural images. Therefore, to address the issue of demanding requirement for extensive annotated datasets in non-visible spectrum, this work examines the application of SAM and its effects on identifying objects of interest under X-ray (PIDray [38], CLCXray [58], DBF6 [1]) and infrared imagery datasets (FLIR [17]). Utilising the variational prompting capabilities of SAM (bounding box, centroid, and random points), we conduct a thorough quantitative and qualitative analysis of the segmentation results produced by SAM (Fig. 1). We aim to pave the way for utilising SAM to enhance the segmentation of object-of-interest beyond visible spectrum imagery through this evaluative research.

2. Literature Review

Research works on non-visible spectrum imagery have witnessed increased attention in the literature. In this context, infrared imaging is progressively gaining traction across various fields, driven by the decreasing size and cost of its sensors. This trend has positioned it as a preferred choice for applications in visual surveillance and autonomous driv-

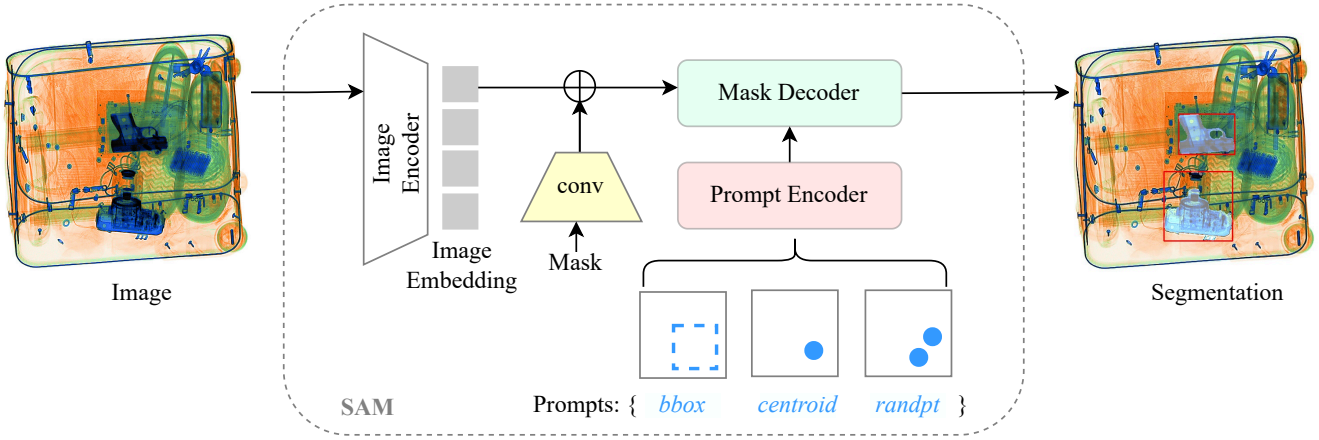

Figure 2. Given an input image, Segment Anything Model (SAM) initiates the process by generating image embeddings via an image encoder. These embeddings are subsequently interactively queried by variational prompts (bounding box, centroid, and random points) in order to generate precise segmentation masks for the objects of interest.

ing [4, 15, 19, 35, 44, 48]. Moreover, there has been a steady rise in research focused on object-based prohibited items [6, 7, 54] and anomaly detection [1, 5, 18] within X-ray baggage security imagery. For a comprehensive review of X-ray security screening, readers can refer to the works of [2, 46, 53].

Key public datasets support the utilisation of CNN-based object detection beyond the visible spectrum. For Infrared imagery, pivotal resources include the FLIR ADAS dataset [17] and the Multispectral KAIST dataset [24], while for X-ray imaging, popular datasets such as GDXray [41], SIXray [42], PIDray [57], and OPIXray [55] play a crucial role. Both datasets are designed to emphasize the advancement of CNN-driven object detection systems by offering detailed annotations beyond the visible spectrum in both Infrared and visible imagery.

The effectiveness of object detection methods depends significantly on the presence of annotated and labelled Infrared and X-ray images. Such a universal model can be achieved via foundational models such as SAM, showcasing its remarkable performance across various medical segmentation tasks [13, 23, 39, 40]. The investigations into SAM performance across a diverse array of medical imagery demonstrated that while SAM achieves commendable segmentation results on targets with clear boundaries, it struggles significantly with typical medical subjects that have weak boundaries or exhibit low contrast [23, 40]. While prior methods utilised the standard SAM directly for segmentation tasks, MedSAM [39] has adopted a distinct strategy by fine-tuning SAM on a large dataset containing over one million medical image-mask pairs. The results demonstrate that MedSAM significantly enhances the segmentation performance and outperforms specialist models that were trained from the same modality [26].

Drawing on the success of SAM in medical imagery, this study seeks to provide a thorough examination into the efficacy of SAM in segmenting imagery of non-traditional modalities, specifically Infrared and X-ray images, without the need for re-training or fine-tuning. Our study explores the zero-shot capabilities of SAM across three public X-ray datasets [3, 57, 58] targeted on prohibited items such as firearms, knives, hammers, etc., alongside one public infrared dataset [17] where the detected objects include pedestrians, cars, bicycles, and similar entities. We aspire that this preliminary investigation offers insights into the performance of SAM beyond the visible spectrum, potentially looking into its applicability for generating high-quality annotations in infrared and X-ray imagery. Our goal is to evaluate whether SAM could facilitate the curation and offer detailed annotation of new datasets beyond the visible spectrum, fostering further advancements in the field.

3. Methodology

We propose SAM to produce high-quality zero-shot segmentation masks for datasets beyond the visible spectrum. In Section 3.1, we first briefly review the architecture of SAM, followed by Section 3.2 which addresses our primary application of prompting capabilities of SAM to generate segmentation masks for infrared and X-ray modality images.

3.1. SAM architecture

The Segment Anything Model (SAM) is a foundation model that has achieved promising zero-shot segmentation performance, trained on a large visible imagery dataset. It is done by isolating specific objects within an image based on user-defined prompts. These prompts can vary from a single point, a full mask, a bounding box or text. SAM mainly consists of three modules, as depicted in Fig. 2. The first

module, the image encoder, is composed of a Vision Transformer (ViT) [16] backbone for image feature extraction, resulting in image embedding in a spatial size of 64×64. The second module, the prompt encoder, encodes the interactive positional information derived from input points, boxes, or masks, to provide for the mask decoder. The third module, the mask decoder, consists of a two-layer transformer-based decoder which takes both the extracted image embedding with the concatenated output and prompt tokens for final mask prediction. The core principle of SAM lies in its ability to show strong zero-shot generalisation to new data without the necessity for additional training, since it was trained progressively on the large-scale Segment Anything 1 Billion (SA-1B) dataset, which contains over 1 billion automatically generated masks ($400\times$ more masks than any existing segmentation datasets [21, 34]) and 11 million images.

3.2. SAM application

A key feature of SAM in the second module, as explained in Section 3.1, is the selection of the appropriate segmentation prompts (Fig. 2). While automatic mask generators without manual prompts can be derived, this work focuses on isolating only particular objects of interest, rather than segmenting every object present. Therefore, an auto-prompt approach does not align with this task.

We propose that SAM can be employed through two different prompting conditions for segmenting objects of interest beyond the visible spectrum. First, we use points as prompts. In this setup, a series of specific points within the object of interest in the image is provided to guide the processing of SAM. We provide two types of point prompts. The first type is *centroid*, where we defined the point as the centre of the mask given at each object. The second type is *randpt*, where we defined the prompt as two random points inside the mask given at each object. In addition, bounding box prompts are tested as input prompts for each object of interest, akin to security inspections. This is conducted by using the ground truth bounding box given for each image. The prompting techniques used in this work are:

- **SAM-bbox**: where we employed ground truth bounding box as prompt.
- **SAM-centroid**: where we defined SAM-centroid as the mass centre of the ground truth mask as the prompt.
- **SAM-randpt**: as known as SAM-random point, where we used two random points or coordinates inside the ground truth mask as the prompt.

In each of these approaches, SAM is directly utilised on the selected datasets, from infrared to X-ray security imagery, without any re-training or fine-tuning specific to those datasets. Moreover, all parameters are set to the default values [28]. When multiple masks corresponding to various regions or structures within the image are generated, we select the mask that exhibits the highest overlap with the ground-truth mask for evaluating the segmentation.

4. Experimental Setup

This section presents the used datasets and the implementation details of our experiments.

4.1. Datasets

The following datasets are used in our evaluation:

PIDray [57]: this X-ray imagery dataset comprises a comprehensive collection of prohibited items, encompassing 12 distinct classes: *baton, bullet, gun, hammer, handcuffs, knife, lighter, pliers, power bank, scissors, sprayer, and wrench*. With a total of $29,457$ training images sourced from various environments, including airports, subway stations, and railway stations, this dataset offers a diverse and realistic representation of real-world scenarios.

CLCXray [58]: this X-ray imagery dataset offers a substantial dataset featuring overlapping objects sourced from real-life scenarios, with a particular emphasis on hazardous liquids, thus broadening the scope of threat object research. The dataset comprises $7,652$ X-ray training images, a combination of real subway scenes and synthetically generated through manual bag design simulations. It encompasses 12 categories, including five types of cutters (*blade, dagger, knife, scissors, swiss army knife*) and seven types of liquid containers (*can, carton drink, glass bottle, plastic bottle, vacuum cup, spray can, tin*).

DBF6 [1]: this dataset comprises conventional pseudo-colour X-ray security images captured by a Smith Detection dual-energy scanner, featuring four views. It includes six object classes: *firearm, firearm part, knife, camera, ceramic knife, and laptop*, with a total of $8,100$ training images. Each object is meticulously annotated with segmentation masks across all views, enabling accurate identification, and is assigned a local index for seamless tracking. The dataset encompasses images depicting single objects as well as complex scenarios with multiple objects, providing diverse and challenging samples for analysis and training.

FLIR [17]: this infrared imagery dataset offers meticulously annotated single-channel grayscale infrared images covering various object classes. These images are captured in clear-sky conditions, encompassing both day (60%) and night (40%) settings. The Infrared imagery is acquired using a FLIR Tau2 camera, renowned for its Long Wave Infrared Cameras (LWIR), with a high resolution of 640×512 pixels. For our experiments, we primarily focus on the training set (totalling $7,859$ images) with three key object classes: *Person, Bicycle, and Car*.

4.2. Implementation Details

We use the original implementation of SAM [28] with a ViT-H [16] backbone without further training or fine-tuning

Table 1. PIDray: Average Recall comparison using IoU types: {*Bbox*} with various IoU thresholds.

	Prompt	$AR_{[IoU=0.50:0.95]}$	$AR_{[IoU=0.50]}$	$AR_{[IoU=0.75]}$	$AR_{S\,[IoU=0.50:0.95]}$	$AR_{M\,[IoU=0.50:0.95]}$	$AR_{L\,[IoU=0.50:0.95]}$
Bbox	↳*bbox*	**0.767**	**0.972**	**0.855**	**0.767**	-	-
	↳*centroid*	0.393	0.613	0.401	0.393	-	-
	↳*randpt*	0.456	0.687	0.469	0.456	-	-

Table 2. CLCXray: Average Recall comparison using IoU type: {*Bbox*} with various IoU thresholds.

	Prompt	$AR_{[IoU=0.50:0.95]}$	$AR_{[IoU=0.50]}$	$AR_{[IoU=0.75]}$	$AR_{S\,[IoU=0.50:0.95]}$	$AR_{M\,[IoU=0.50:0.95]}$	$AR_{L\,[IoU=0.50:0.95]}$
Bbox	↳*bbox*	**0.797**	**0.992**	**0.894**	**0.704**	**0.697**	**0.801**
	↳*centroid*	0.597	0.821	0.644	0.512	0.477	0.596
	↳*randpt*	0.282	0.423	0.292	0.544	0.295	250

Table 3. DBF6: Average Recall comparison using IoU types: {*Bbox, Segm*} with various IoU thresholds.

	Prompt	$AR_{[IoU=0.50:0.95]}$	$AR_{[IoU=0.50]}$	$AR_{[IoU=0.75]}$	$AR_{S\,[IoU=0.50:0.95]}$	$AR_{M\,[IoU=0.50:0.95]}$	$AR_{L\,[IoU=0.50:0.95]}$
Bbox	↳*bbox*	**0.660**	**0.978**	**0.726**	**0.449**	**0.621**	**0.745**
	↳*centroid*	0.399	0.703	0.398	0.281	0.364	0.460
	↳*randpt*	0.400	0.696	0.400	0.249	0.314	0.475
Segm	↳*bbox*	**0.537**	**0.912**	**0.533**	**0.320**	**0.474**	**0.572**
	↳*centroid*	0.394	0.715	0.387	0.229	0.346	0.427
	↳*randpt*	0.432	0.742	0.441	0.224	0.195	0.459

Table 4. FLIR: Average Recall comparison using IoU type: {*Bbox*} with various IoU thresholds.

	Prompt	$AR_{[IoU=0.50:0.95]}$	$AR_{[IoU=0.50]}$	$AR_{[IoU=0.75]}$	$AR_{S\,[IoU=0.50:0.95]}$	$AR_{M\,[IoU=0.50:0.95]}$	$AR_{L\,[IoU=0.50:0.95]}$
Bbox	↳*bbox*	**0.606**	**0.991**	**0.627**	**0.546**	**0.688**	**0.784**
	↳*centroid*	0.286	0.606	0.239	0.266	0.303	0.226
	↳*randpt*	0.282	0.565	0.250	0.231	0.353	0.322

to assess its performance for non-visible band datasets. We evaluate on the training partition of each dataset since they allow for more statistically significant results. For each experiment, the prompts (*bbox, centroid* and *randpt*) are obtained from the ground truth datasets. The centroid is calculated as the mean of all ground truth mask vertices while the random points are obtained via Monte Carlo sampling of points within the bounding box and testing whether these points lie inside the polygon defining the ground truth mask until the desired number of random points is achieved (in our experiments, two random points). We report bounding box average recall (AR) in all our experiments by comparing the bounding box from the predicted mask against the ground truth bounding boxes. For DBF6, segmentation AR is also reported since ground truth masks are available. Additionally, we report the Recall for different intersection-over-union (IoU) thresholds and the mean IoU for each prompt/dataset pairs to evaluate the quality of the predictions. All experiments were run using an NVIDIA 3090Ti GPU.

5. Results

The resulting metrics for each dataset are summarised in Table 1 to 4. We compiled our performance on the datasets across X-ray and infrared imagery, with varying IoU thresholds. The results are reported in terms of AR in two modes, bounding box (Bbox) mode and segmentation mask (Segm) mode. Note that only the DBF6 dataset has segmentation mask ground truth, meanwhile, the other three public datasets chosen only provide bounding box ground truth. The results are reported under three prompts, (*bbox*, *centroid*, *randpt*), as explained in Section 3.2.

Having compiled a dataset across X-ray and infrared imagery, we noted that the segmentation efficacy of SAM is quantitatively influenced by the chosen prompting technique. For instance, across all datasets, bbox prompt shows superior results on object segmentation tasks on all IoU evaluation metrics, indicating that bbox prompting allows strong features combination within that particular area, by covering the entire object, making it more efficient in these instances.

We compare the result of the AR according to IoU evaluation criteria as shown in Table 1 and Fig. 3 (*left*) for PIDray dataset [57]. For each of the given prompts, it consistently shows that AR decreases as IoU thresholds become stricter. At lower IoU thresholds, it is easier for SAM to have high AR because the criteria for correct detection are more lenient. As the IoU threshold increases, requiring more precise overlap, SAM ability to capture all class targets without also increasing false positives becomes more challeng-

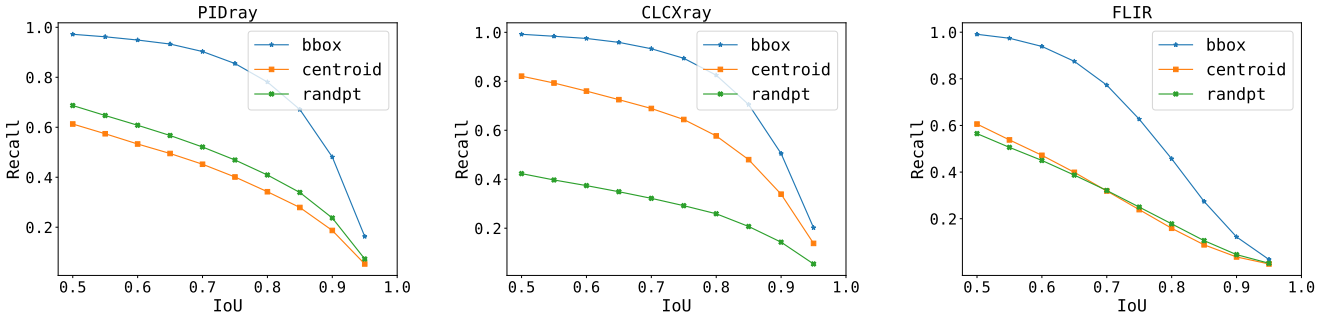

Figure 3. Recall performance using variational prompting strategies across different IoU thresholds and IoU type: *Bbox*.

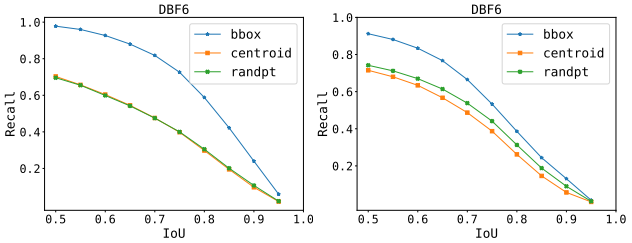

Figure 4. DBF6: Recall performance using variational prompting strategies across different IoU thresholds and IoU types: *Bbox* (left), *Segm* (right).

ing, leading to lower AR. It can be seen that whilst bbox prompt gives higher results, randpt gives slightly better results than the centroid point. The distinct class of PIDray, as explained in Section 4.1, give more advantages to SAM with randpt prompt, where having more points inside the target class is effectively better at segmenting compared to centroid point, with a small margin.

We further analyse the AR based on IoU evaluation criteria, as presented in Table 2 and Fig. 3 (*middle*) for the CLCXray dataset [58]. This analysis reveals a consistent trend with the PIDray results; however, the centroid prompt notably outperforms the random point prompt by a significant margin, unlike the PIDray findings (Fig. 3 (*left*)). In this regard, the critical influence of material composition may influence the effectiveness of prompt selection. The CLCXray dataset has a class-wise combination which consists of organic material (carton drinks, plastic bottles) as well as metallic material (spray cans, tin), contrasting with the PIDray dataset that exclusively consists of metallic composition, as discussed in Section 4.1. The presence of clutter, especially involving organic materials, poses greater challenges for the random point prompt due to the less distinctive features of these objects. This suggests that increasing the number of prompt points does not automatically enhance segmentation performance. Instead, strategically placing a prompt at the centre of the target object significantly improves results, as evidenced in Fig. 3 (*middle*).

In our final analysis of X-ray imagery, we focus on the DBF6 dataset results, detailed in Table 3 and illustrated in Fig. 4. This evaluation considers two modes: bounding box (Bbox) and segmentation (Segm). While the bbox prompt maintains consistent performance in the Bbox mode, aligning with the patterns observed in Fig. 3, we notice a marginal decline in its effectiveness in the Segm mode. Although SAM can generate bounding boxes from segmentation masks, this approach proved to be less efficient, primarily due to difficulty associating the bbox prompt with Segm mode of SAM. Conversely, the point-based prompting methods, both centroid and randpt, demonstrated stable performance across both modes, differing only slightly. This consistency indicates the robust adaptability of point prompts to varying segmentation tasks within X-ray imagery analysis.

Table 5. Mean IoU for each prompt/dataset pairs.

Prompt	DBF6 (Box)	DBF6 (Segm)	PIDRay	CLCXray	FLIR
↳*bbox*	0.808 ± 0.130	0.730 ± 0.171	0.849 ± 0.131	0.870 ± 0.104	0.779 ± 0.110
↳*centroid*	0.603 ± 0.288	0.591 ± 0.287	0.577 ± 0.299	0.728 ± 0.255	0.534 ± 0.235
↳*randpt*	0.606 ± 0.283	0.621 ± 0.281	0.634 ± 0.280	0.440 ± 0.337	0.547 ± 0.262

In our analysis of infrared imagery, as presented in Table 4 and depicted in Fig. 3 (*right*), we observe that while the bbox prompt generally results in higher AR, there is a notable decrease in AR performance as the IoU threshold increases for all types of prompts. This trend suggests SAM's limited ability to understand the characteristics of class-specific infrared imagery, given its training predominantly on natural images. We propose that fine-tuning the SAM model with an infrared imagery dataset could significantly enhance its segmentation accuracy and effectiveness, providing more robust quantitative results.

Table 5 presents the mean IoU for each prompt/dataset pair. Overall, it is observed that the bounding box prompts usually lead to a good bounding box prediction, meaning that SAM can segment the object inside the proposed bounding box. This is still confirmed for the DBF6 segmentation dataset, with a relatively high mean mask IoU. Among the bounding box datasets, FLIR obtains the lowest mean IoU, indicating its challenging nature to SAM, which is explained by the low contrast of the objects against the background (see Fig. 7). From the centroid and random point prompts, it is observed that, generally, two ran-

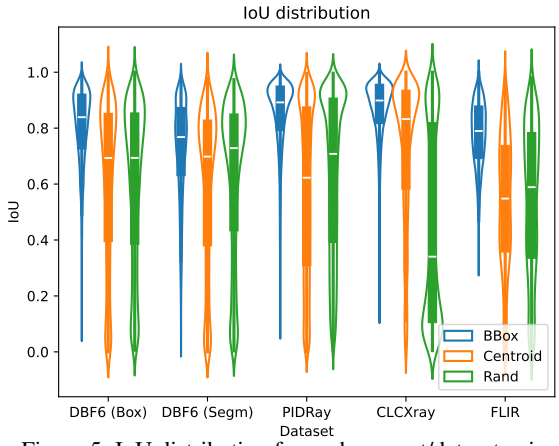

Figure 5. IoU distribution for each prompt/dataset pair.

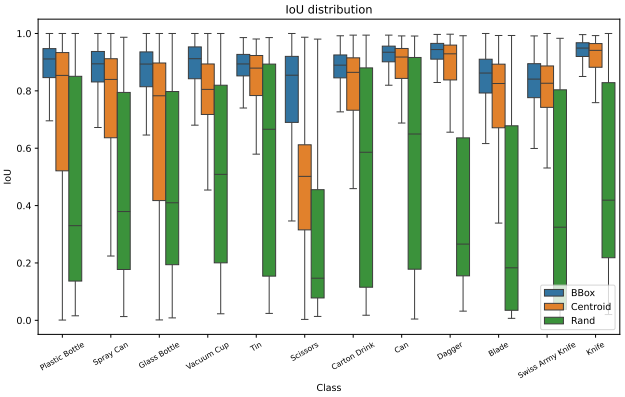

Figure 6. Class-wise IoU distribution on the CLCXray dataset.

dom points lead to a better performance than the centroid. While this might seem counter-intuitive, it indicates that the centroid is not always the most significant point, while two random points within the object lead to more cues for SAM. The contrary is noted for the CLCXray, where several objects consist of thin objects and bottles containing liquids [58], which are difficult to capture by an X-ray machine and two random points might still be confused with the background. These trends are further confirmed in the IoU distribution shown in Fig. 5. It is noted that the three types of prompts generally yield a good segmentation mask for DBF6 and PIDRay, where the test objects are metallic. On the other hand, the CLCXray shows a lower performance with a high density of low IoU objects for the random points prompt. To further investigate this, Fig. 6 shows the class-wise IoU distribution of the CLCXray dataset. It is seen that the best-performing classes for the random point prompt are the tin and the cans, which are metallic and easily segmented. On the other hand, the worst performances are obtained for thin objects (scissors, blades and daggers) and organic material objects (such as plastic bottles), where the choice of the random point might significantly impact the prediction. Finally, it is also observed that SAM has the poorest performance on the FLIR dataset when using point prompts, which is again attributed to the low contrast of the dataset.

The qualitative analysis, which examines the variance in prompts and their alignment with ground truths across different datasets, is illustrated in Figs. 1 and 7. For the PIDray dataset (Fig. 7, 1^{st} *row*), we observe that point-based prompts often extend beyond the actual object boundaries, leading to an increased occurrence of false positives compared to the bounding box prompts. In the case of the CLCXray dataset (Fig. 7, 2^{nd} *row*), which features a high degree of overlap between organic and metallic materials, strategically placing a prompt directly at the centre of the target significantly improves bounding box precision compared to random placement. The randpt prompt particularly struggles with smaller objects, often exceeding their boundaries. Regarding the DBF6 dataset (Fig. 7, 3^{rd} *row*), point-based prompts generally achieve better segmentation than bounding box prompts, although they tend to generate more false positives for classes with smaller objects. In the FLIR dataset (Fig. 7, 4^{th} *row*), while the bounding box prompts offer superior outcomes, point-based prompts fail to accurately localise distinct classes such as pedestrians and cars, likely due to the significant domain shift between the infrared imagery encountered here and the visible band imagery upon which SAM was trained.

6. Conclusion

This work presents a thorough assessment of the Segment Anything Model (SAM) performance for images beyond the visible spectrum. We evaluate SAM within three types of prompts, namely bounding box, centroid and random points, on three X-ray security imagery datasets and an infrared surveillance dataset. Our results suggest that while SAM exhibits a great capability to segment objects when given a bounding box prompt, its performance drops when given point prompts. Specifically, it is observed that SAM extends objects beyond their boundaries in X-ray images, with particular difficulty for objects based on organic materials (which might be confused with the background). Additionally, the low-contrast characteristic of infrared images and the different appearance of the objects impose a significant challenge on SAM, with poor segmentation performance using the centroid and random point prompts. Future directions may include fine-tuning SAM to the assessed image modalities to increase its segmentation performance. This would allow to streamlining the dataset annotation processes, reducing the reliance on manual labelling. Such advancements could facilitate the creation and enrichment of datasets across various image modalities, significantly broadening the scope and utility of machine learning applications in areas where data collection and annotation have traditionally posed challenges.

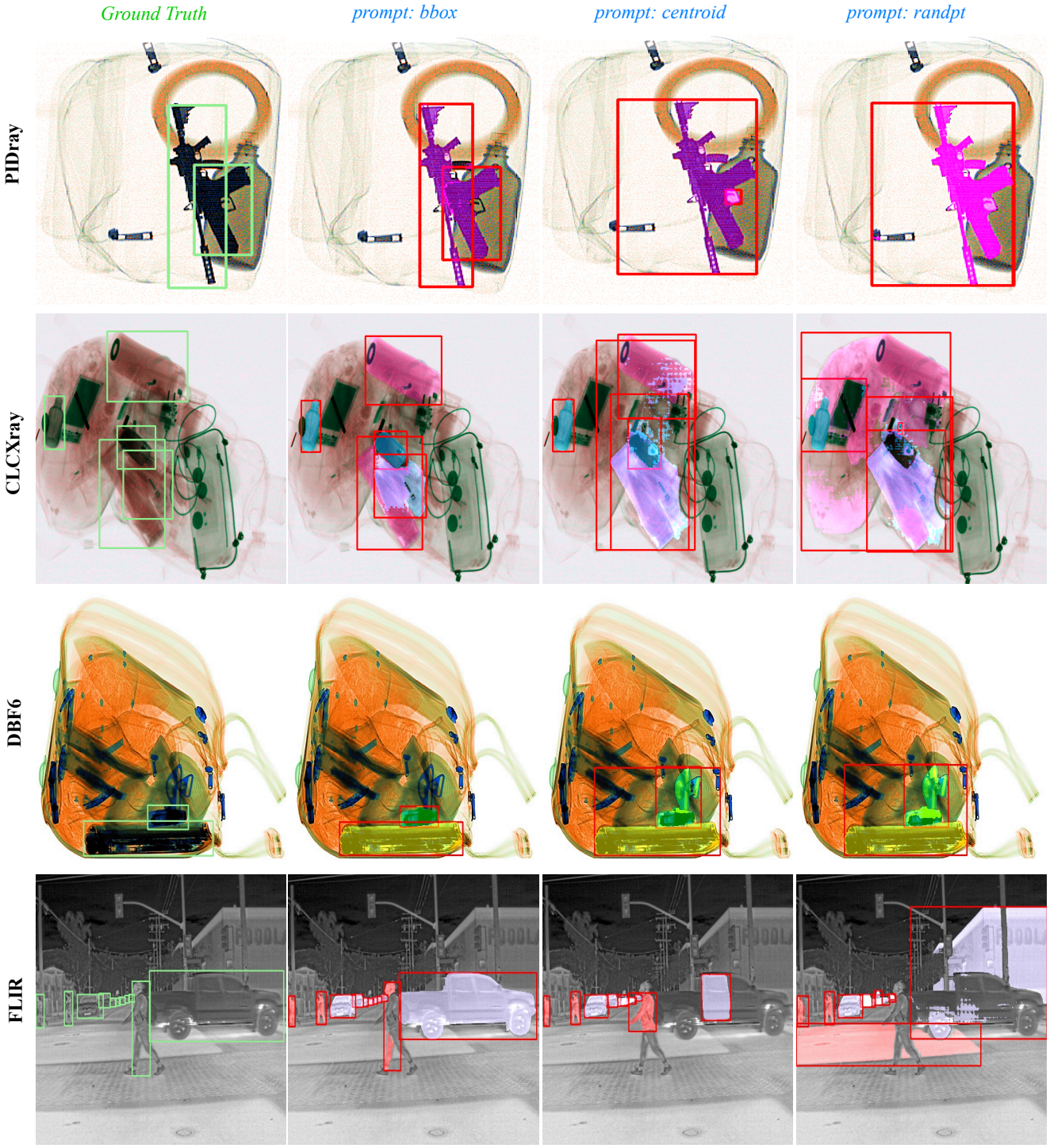

Figure 7. The segmentation results obtained by SAM, utilising variational prompting strategies, are examined across PIDray, CLCXray, DBF6, and FLIR datasets. Notably, the prompt bbox consistently yields the most accurate segmentations. However, the other two prompting strategies occasionally encounter challenges, particularly in scenarios where objects are overlapped and cluttered, as observed in the X-ray datasets.

References

[1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, pages 622–637. Springer, 2018. 2, 3, 4

[2] Samet Akcay and Toby Breckon. Towards automatic threat detection: A survey of advances of deep learning within x-ray security imaging. *Pattern Recognition*, 122:108245, 2022. 3

[3] Samet Akcay, Mikolaj E Kundegorski, Chris G Willcocks, and Toby P Breckon. Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery. *IEEE transactions on information forensics and security*, 13(9):2203–2215, 2018. 2, 3

[4] Neelanjan Bhowmik, Jack W Barker, Yona Falinie A Gaus, and Toby P Breckon. Lost in compression: the impact of lossy image compression on variable size object detection within infrared imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 369–378, 2022. 3

[5] N. Bhowmik, Y.F.A. Gaus, S. Akcay, J. W. Barker, and T. P. Breckon. On the impact of object and sub-component level segmentation strategies for supervised anomaly detection within x-ray security imagery. In *18th IEEE Int. Conf. on Machine Learning and Applications (ICMLA 2019)*. IEEE, December 2019. 3

[6] N. Bhowmik, Y.F.A. Gaus, and T.P. Breckon. On the impact of using x-ray energy response imagery for object detection via convolutional neural networks. In *Proc. Int. Conf. on Image Processing*, pages 1224–1228, 2021. 3

[7] N. Bhowmik, Wang. Q., Y.F.A. Gaus, M. Szarek, and T.P. Breckon. The good, the bad and the ugly: Evaluating convolutional neural networks for prohibited item detection using real and synthetically composite x-ray imagery. In *Proc. British Machine Vision Conf. Workshops*, pages 1–8. BMVA, 2019. 3

[8] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 2

[9] Raluca Brehar and Sergiu Nedevschi. Pedestrian detection in infrared images using hog, lbp, gradient magnitude and intensity feature channels. In *Proc. Conf. on Intelligent Transportation Systems*, pages 1669–1674. IEEE, 2014. 2

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. European Conf. on Computer Vision*, pages 213–229, 2020. 2

[12] An Chang, Yu Zhang, Shunli Zhang, Leisheng Zhong, and Li Zhang. Detecting prohibited objects with physical size constraint from cluttered x-ray baggage images. *Knowledge-Based Systems*, 237:107916, 2022. 2

[13] Dongjie Cheng, Ziyuan Qin, Zekun Jiang, Shaoting Zhang, Qicheng Lao, and Kang Li. Sam on medical images: A comprehensive study on three prompt modes. *arXiv preprint arXiv:2305.00035*, 2023. 3

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. Conf. Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[15] Chaitanya Devaguptapu, Ninad Akolekar, Manuj M Sharma, and Vineeth N Balasubramanian. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In *Proc. Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 3

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *Int. Conf. of Learning Representations*, 2021. 4

[17] FLIRSystems. FLIR Thermal Datasets for Algorithm Training. https://www.flir.co.uk/oem/adas/dataset/. 2, 3, 4

[18] Y.F. A. Gaus, N. Bhowmik, S. Akçay, P.M. Guillen-Garcia, J.W. Barker, and T.P. Breckon. Evaluation of a dual convolutional neural network architecture for object-wise anomaly detection in cluttered x-ray security imagery. In *Proc. Int. Joint Conf. on Neural Networks*, 2019. 2, 3

[19] Yona Falinie A Gaus, Neelanjan Bhowmik, Brian KS Isaac-Medina, and Toby P Breckon. Visible to infrared transfer learning as a paradigm for accessible real-time object detection and classification in infrared imagery. In *Counterterrorism, Crime Fighting, Forensics, and Surveillance Technologies IV*, volume 11542, page 1154205, 2020. 2, 3

[20] Bangzhong Gu, Rongjun Ge, Yang Chen, Limin Luo, and Gouenou Coatrieux. Automatic and robust object detection in x-ray baggage inspection using deep convolutional neural networks. *IEEE Transactions on Industrial Electronics*, 68(10):10248–10257, 2020. 2

[21] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 4

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proc. Int. Conf. on Computer Vision*, pages 2980–2988, 2017. 2

[23] Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *Medical Image Analysis*, 92:103061, 2024. 3

[24] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proc. of the Conf. on computer vision and pattern recognition*, pages 1037–1045, 2015. 3

[25] Brian KS Isaac-Medina, Seyma Yucer, Neelanjan Bhowmik, and Toby P Breckon. Seeing through the data: A statistical evaluation of prohibited item detection benchmark datasets for x-ray security screening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 524–533, 2023. 2

[26] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 3

[27] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 2

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 4

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton.

Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2

[30] M.E. Kundegorski, S. Akcay, M. Devereux, A. Mouton, and T.P. Breckon. On using feature descriptors as visual words for object detection within x-ray baggage security screening. In *Proc. Int. Conf. on Imaging for Crime Detection and Prevention*, pages 1–6, November 2016. 2

[31] Mikolaj E Kundegorski, Samet Akçay, Grégoire Payen de La Garanderie, and Toby P Breckon. Real-time classification of vehicles by type within infrared imagery. In *Optics and Photonics for Counterterrorism, Crime Fighting, and Defence XII*, volume 9995, pages 266–281. SPIE, 2016. 2

[32] Mikolaj E Kundegorski and Toby P Breckon. A photogrammetric approach for real-time 3d localization and tracking of pedestrians in monocular infrared imagery. In *Optics and Photonics for Counterterrorism, Crime Fighting, and Defence X; and Optical Materials and Biomaterials in Security and Defence Systems Technology XI*, volume 9253, page 92530I, 2014.

[33] Mikolaj E Kundegorski and Toby P Breckon. Posture estimation for improved photogrammetric localization of pedestrians in monocular infrared imagery. In *Optics and Photonics for Counterterrorism, Crime Fighting, and Defence XI; and Optical Materials and Biomaterials in Security and Defence Systems Technology XII*, volume 9652, pages 114–125. SPIE, 2015. 2

[34] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 4

[35] Eun Ju Lee, Byoung Chul Ko, and Jae-Yeal Nam. Recognizing pedestrian's unsafe behaviors in far-infrared imagery at night. *Infrared Physics & Technology*, 76:261–270, 2016. 3

[36] Jianfu Li, Weiguo Gong, Weihong Li, and Xiaoying Liu. Robust pedestrian detection in thermal infrared imagery using the wavelet transform. *Infrared Physics & Technology*, 53(4):267–273, 2010. 2

[37] Zhongqiu Liu, Jianchao Li, Yuan Shu, and Dongping Zhang. Detection and recognition of security detection object based on yolo9000. In *2018 5th International Conference on Systems and Informatics (ICSAI)*, pages 278–282. IEEE, 2018. 2

[38] Bowen Ma, Tong Jia, Min Su, Xiaodong Jia, Dongyue Chen, and Yichun Zhang. Automated segmentation of prohibited items in x-ray baggage images using dense de-overlap attention snake. *IEEE Transactions on Multimedia*, pages 1–1, 2022. 2

[39] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 3

[40] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023. 3

[41] Domingo Mery, Vladimir Riffo, Uwe Zscherpel, German Mondragón, Iván Lillo, Irene Zuccar, Hans Lobel, and Miguel Carrasco. Gdxray: The database of x-ray images for nondestructive testing. *Journal of Nondestructive Evaluation*, 34(4):42, 2015. 2, 3

[42] Caijing Miao, Lingxi Xie, Fang Wan, Chi Su, Hongye Liu, Jianbin Jiao, and Qixiang Ye. Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2119–2128, 2019. 2, 3

[43] Farzeen Munir, Shoaib Azam, and Moongu Jeon. Sstn: Self-supervised domain adaptation thermal object detection for autonomous driving. In *2021 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 206–213. IEEE, 2021. 2

[44] Min Peng, Chongyang Wang, Tong Chen, and Guangyuan Liu. Nirfacenet: A convolutional neural network for near-infrared face identification. *Information*, 7(4):61, 2016. 3

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[46] Mehdi Rafiei, Jenni Raitoharju, and Alexandros Iosifidis. Computer vision on x-ray data in industrial production and security applications: A comprehensive survey. *IEEE Access*, 11:2445–2477, 2023. 3

[47] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[48] Iain Rodger, Barry Connor, and Neil M Robertson. Classifying objects in lwir imagery via cnns. In *Electro-Optical and Infrared Systems: Technology and Applications XIII*, volume 9987, page 99870H, 2016. 3

[49] Arnold WM Smeulders, Dung M Chu, Rita Cucchiara, Simone Calderara, Afshin Dehghan, and Mubarak Shah. Visual tracking: An experimental survey. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1442–1468, 2013. 2

[50] Malarvizhi Subramani, Kayalvizhi Rajaduari, Siddhartha Dhar Choudhury, Anita Topkar, and Vijayakumar Ponnusamy. Evaluating one stage detector architecture of convolutional neural network for threat object detection using x-ray baggage security imaging. *Rev. d'Intelligence Artif.*, 34(4):495–500, 2020. 2

[51] Michael Teutsch, Thomas Muller, Marco Huber, and Jurgen Beyerer. Low resolution person detection with a moving thermal infrared camera by hot spot classification. In *Proc. of the Conf. on Computer Vision and Pattern Recognition Workshops*, pages 209–216, 2014. 2

[52] D. Turcsany, A. Mouton, and T.P. Breckon. Improving feature-based object recognition for x-ray baggage security screening using primed visual words. In *Proc. Int. Conf. on Industrial Technology*, pages 1140–1145. IEEE, February 2013. 2

[53] Divya Velayudhan, Taimur Hassan, Ernesto Damiani, and Naoufel Werghi. Recent advances in baggage threat detection: A comprehensive and systematic survey. *ACM Comput. Surv.*, 55(8), dec 2022. 3

[54] Thomas W Webb, Neelanjan Bhowmik, Yona Falinie A Gaus, and Toby P Breckon. Operationalizing convolutional neural network architectures for prohibited object detection in x-ray imagery. In *Proc. Int. Conf. on Machine Learning and Applications*, pages 610–615, 2021. 2, 3

[55] Yanlu Wei, Renshuai Tao, Zhangjie Wu, Yuqing Ma, Libo Zhang, and Xianglong Liu. Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module. In *Proc. Int. Conf. on Multimedia*, MM '20, page 138–146, 2020. 2, 3

[56] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018. 2

[57] Libo Zhang, Lutao Jiang, Ruyi Ji, and Heng Fan. Pidray: A large-scale x-ray benchmark for real-world prohibited item detection. *arXiv preprint arXiv:2211.10763*, 2022. 2, 3, 4, 5

[58] Cairong Zhao, Liang Zhu, Shuguang Dou, Weihong Deng, and Liang Wang. Detecting overlapped objects in x-ray security imagery by a label-aware mechanism. *IEEE Transactions on Information Forensics and Security*, 17:998–1009, 2022. 2, 3, 4, 6, 7

[59] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. In *Proc. of the Conf. on computer vision and pattern recognition*, pages 840–849, 2019. 2

[60] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *Proc. Int. Conf. on Learning Representations*, 2021. 2

[61] Z. Cai and N. Vasconcelos. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. 2

[62] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proc. Int. Conf. on Computer Vision*, pages 2999–3007, 2017. 2

[63] D. Mery, V. Riffo, I. Zuccar, and C. Pieringer. Automated x-ray object recognition using an efficient search algorithm in multiple views. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition Workshops*, pages 368–374, 2013. 2

[64] 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. on Computer Vision and Pattern Recognition*, pages 2114–2123, 2019. 2

[65] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 2