ISSN: 2664-9640

SE-YOLO: Refining Fine-Grained Feature Extraction for Detecting Hazardous Materials in X-ray Security Inspection Images

Fuquan Qin, and Yan Wei*

College of Computer Science and Information Sciences, Chongqing Normal University, Huxi, 401331, Chongqing, China

*Corresponding Author

Abstract

Detecting hazardous materials in X-ray security images is crucial for public safety and property protection. Currently, the identification of such materials in X-ray luggage scans heavily relies on manual inspection by security officers. To enhance the accuracy of automated detection, this paper introduces a novel SE-YOLO model based on the YOLOv8 architecture. Firstly, SCDConv is implemented to remove convolutional strides and pooling operations by integrating channel spatial-to-depth layers and strideless convolution layers. Secondly, an EMMA attention module is developed to partition feature information along the channel dimension, fostering interaction and fusion of multidimensional data in the fea- ture space through parallel branch structures. Lastly, a high-resolution detection head is designed to facilitate more precise category and confidence predictions. Experimental results on the SIXray dataset demonstrate that the SE-YOLO model achieves a detection accuracy of 92.3%, a 2.1% improvement over the YOLOv8 model.

Keywords

Dangerous Goods Detection; Yolov8; Deep Learning; Attention Mechanism.

1. Introduction

In recent years, the escalating threats of terrorism and illicit goods have highlighted the indispensable role of X-ray security inspection technology in the realm of safety. Widely deployed in public venues like airports, train stations, and border checkpoints, X-ray security inspection images serve the crucial purpose of detecting hazardous materials to ensure public security. However, the task of identifying hazardous items in these images presents numerous challenges. This is primarily attributed to the complex shapes, densities, and materials of hazardous items, making them indistinguishable from the surrounding environment in X-ray scans. Additionally, X-ray security inspec- tion images often contain a plethora of interfering elements such as luggage, clothing, and metallic objects, which can obscure or distort the characteristics of small haz- ardous objects. Furthermore, the resolution and quality of X-ray security inspection images can significantly impact detection accuracy. Addressing these challenges and enhancing the precision of hazardous material detection in X-ray security inspection images remains an ongoing and formidable task.

Detecting hazardous items in X-ray security inspection images involves two pri- mary methodologies: traditional machine learning and deep learning approaches. Traditional machine learning methods rely on manual feature extraction and classifier utilization for hazardous item detection. However, these traditional approaches are lim- ited as they often require significant manual intervention and involve complex feature engineering. Their effectiveness diminishes when dealing with X-ray security inspec- tion images containing complex features. With the rise of deep learning, numerous detection methods based on deep

ISSN: 2664-9640

learning have been introduced to enhance security screening effectiveness and reduce the workload of human operators[16]. Compared to traditional machine learning algorithms, neural networks benefit from deeper layers and larger feature spaces, allowing them to be trained on large datasets to achieve more expressive results [22]. Deep learning-based object detection algorithms can be classified into two categories: two-stage algorithms and single-stage algorithms. Two- stage algorithms first generate region proposals and then classify them, resulting in longer detection times and poorer real-time performance. Representative algorithms of two-stage algorithms include R-CNN[4], Fast-RCNN [3], and Faster R-CNN [18]. Single-stage algorithms, such as YOLO (you only look once, YOLO)[17] and SSD (single-shot multi-box detector)[11] series algorithms, directly predict object positions and categories using convolutional neural networks, offering faster detection speeds and better real-time performance.

With the development and refinement of various algorithms, there has been continuous improvement in the accuracy and speed of detecting hazardous items in X-ray security inspection images. However, existing algorithms often lack optimization for detecting small-sized hazardous items in security inspection images. This deficiency results in models consistently neglecting the detection capability for small-sized hazardous items, thus limiting further enhancement in average detection accuracy. In response to the inadequacy of existing models in detecting small-sized hazardous items, this paper proposes a novel X-ray image hazardous item detection algorithm-SE-YOLO. This paper contributes in the following ways:

• SCDConv is devised to enhance the extraction capability of fine-grained feature information. By eliminating convolutional strides and pooling operations through the Channel Spatial to Depth (SCD) layer, SCDConv reduces the loss of fine-grained information during convolution operations.

• The EMMA attention module facilitates the interaction and fusion of multidimensional information in the channel space through parallel branch structures, thereby augmenting the model's multiscale feature extraction capability.

• To enable the model to capture small-sized hazardous items more effectively, a high-resolution detection head is designed. This detection head enables the model to make more optimized category and confidence predictions.

2. Related Work

2.1. X-ray Security Inspection Image Detection

Currently, detection in X-ray security inspection images is predominantly categorized into traditional machine learning methods and deep learning methods. Traditional machine learning methods involve manual feature design and the utilization of machine learning classifiers for feature classification. These methods have achieved significant progress. For instance, Zhu et al. [31] utilized low-level features such as color, texture, shape, and edge features in X-ray images to extract higher-level features for contraband detection. Turcsany et al. [23] employed Support Vector Machine (SVM) and Speeded- Up Robust Features (SURF) to construct a visual bag-of-words model, clustering feature descriptors to generate visual words, and employing them for contraband iden- tification in dual-energy X-ray images. Kundegorski et al.[8]explored various feature point descriptors as variants of visual words in Bag-of-Visual-Words (BoVW)representation schemes for image-based threat detection in luggage security X-ray images. Vukadinovic et al.[24] constructed a Support Vector Machine (SVM) classifier based on Local Binary Patterns (LBP) texture features. However, X-ray hazardous item detection algorithms based on traditional machine learning methods necessitate manual configuration of detectors and classifiers, resulting in high algorithmic complexity and low accuracy.

ISSN: 2664-9640

Deep learning methods exhibit stronger feature representation capabilities than traditional methods, facilitating automatic learning and extraction of higher-level semantic features. To address challenges in X-ray baggage image security detection, Dong et al.[2] proposed an enhanced YOLOv5 network model for contraband detection. They introduced a convolutional attention module to enhance the network's extraction of deep important features of contraband items, utilized the Mixup data augmentation strategy to simulate detection scenarios with highly overlapped and occluded items, and employed a weighted bounding box fusion algorithm to optimize redundant prediction boxes. Miao et al.[15] proposed a model based on an improved capsule network (DMF and SE Capsule) for contraband detection in X-ray images. This model enhances traditional capsule networks with feature enhancement (dilated convolution multi-scale feature fusion, DMF) and feature selection (squeezeand-excitation block, SE) modules. Wu et al.[27] introduced a mask self-attention mechanism on top of YOLOv5, enhancing its feature representation capacity.They also introduced the Quality Focal Loss function to effectively alleviate class imbalance issues. Yu et al.[29] redesigned the path aggregation network (PANet) module of YOLOv4 using deformable convolutions. They also introduced the Focal-EIOU loss function to address severe loss value oscillations when handling low-quality samples. Ma et al.[12] embedded learnable Gabor convolution layers into the network's lower layers and designed a Spatial Attention (SA) mechanism to weight the output features of Gabor convolution layers. They utilized the Global Context Feature Extraction (GCFE) module to extract multiscale global context information of objects and proposed the Dual-Scale Feature Aggregation (DSFA) module to fuse these global features with features from another layer. Su et al.[19] constructed a module for rotation and occlusion removal (DROM). They employed edge, color, and Oriented FAST and Rotated BRIEF (ORB) features to generate integrated feature maps. Zhang et al.[30] proposed a Deformable Attention Module (MAM), connecting corresponding backbone output feature layers with Large Kernel Attention (LKA) blocks to better focus on effective feature information in feature maps using the adaptive selection feature of the self-attention module. They replaced the Feature Pyramid Network (FPN) with a Path Aggregation Network (PAN) and added Conv-MLP blocks to the self-bottom-up feature fusion part of the PAN network to reduce the loss of some low-level details. Wang et al.[26] introduced an object detection algorithm based on the SSD model, using an improved HarDNet network as the backbone network and introducing multi-scale feature fusion and attention mechanisms to improve detection accuracy. Li et al.[10] improved the upsampling module by embedding channel convolution self-attention and spatial convolution self-attention in the CARAFE structure and reinforced the new upsampling operator to capture dependencies between longdistance features. The aforementioned studies have enhanced neural network models by improving the network structure, introducing attention mechanisms, and employing new feature fusion techniques, thereby making them more suitable for various detection tasks.

2.2. Fine-grained Feature Extraction

As the prevalence of scenarios necessitating small object detection grows, there is a burgeoning interest in refining feature extraction techniques for nuanced analysis. Fine-grained feature extraction aims to distill intricate details from images or videos, facilitating precise identification and understanding of small objects or scenes. Inspired by the Bidirectional Feature Pyramid Network (BiFPN), Huang et al.[7] introduced the Small Target Detection Layer (STPL) into the YOLOv5 framework to identify minor surface defects on steel wires. Li et al.[20] integrated the Channel Attention (CA) module, extending the reach of shallow features in the original FPNet and bolstering small object detection capabilities. Hu et al.[6] revamped the backbone structure of their object recognition algorithm with the Deformable ConvNets v2 module and global attention mechanism, curbing feature loss during network processing and enhancing sensitivity to small-scale objects. Xiong et al.[28] employed soft pooling to fortify feature extraction networks, mitigating the loss of crucial edge information in small objects

ISSN: 2664-9640

inherent in traditional downsampling techniques. They amalgamated learnable parameters in the feature fusion process to ensure the preservation of small object information amid larger object features during fusion. Additionally, they introduced an auxiliary feature extraction layer for comprehensive capture of essential shallow information about small objects. Ma et al.[13] utilized an efficient channel attention mechanism to extract backbone features and combined it with the Expanded Scale Feature Pyramid Network to streamline computation and enrich small object detection capabilities. Cheng et al.[1] employed the k-means++ algorithm for more accurate anchor box initialization and replaced standard convolution blocks with full-dimensional dynamic convolutions in the backbone network to enhance feature extraction for small objects. Moreover, they inserted a Global Attention Mechanism (GAM) into the neck network to focus on global information extraction, effectively addressing sparse object feature challenges. Incorporating WiseIoU (WIoU) mitigates harmful gradients from low-quality annotation data, thereby improving small object detection accuracy. Through attention mechanism integration, feature extraction network optimization, and anchor box selection algorithm refinement, these research efforts have bolstered object detection algorithms' ability to extract fine-grained features and improve small object detection accuracy.

3. Theoretical Framework

3.1. YOLOv8

YOLOv8 stands as a refinement built upon the foundation of YOLOv5. Illustrated in Figure 1, the YOLOv8 model incorporates the concept of Cross-Stage Partial (CSP) components[25] and partitions the overarching model architecture into backbone, neck, and head networks. Similar to its predecessor, YOLOv5, YOLOv8 offers models of different scales (N/S/M/L/X) based on scaling coefficients to cater to various real-world scenarios. YOLOv8 adopts a novel decoupled head structure, segregating the classification and detection heads to enable more focused optimization and training. The head segment of YOLOv8 encompasses multiple target detection networks with varying resolutions for detecting objects of different sizes, featuring resolutions of 20×20, 40×40, and 80×80, respectively. This study introduces a novel detection head with a resolution of 160×160 to enhance the model's capacity in detecting small hazardous objects. Furthermore, YOLOv8 transitions from anchor-based methods to anchor-free approaches. In terms of loss computation, YOLOv8 employs the Task Aligned Assigner positive sample allocation strategy and integrates Distribution Focal Loss to guide the model towards prioritizing object features. This loss calculation strategy empowers the model to glean essential object characteristics, thereby enhancing detection accuracy.

3.2. SE-YOLO

In this research, we propose an enhanced X-ray baggage inspection algorithm, SEYOLO, based on the YOLOv8n architecture, as illustrated in Fig. 1. SE-YOLO primarily focuses on improving the feature extraction and fusion capabilities for smallsized hazardous objects. To achieve this, we devised the SCDConv module to enhance the backbone network's ability to extract fine-grained features. Moreover, we strengthened the neck network's multi-dimensional feature fusion capability by integrating the EMMA attention module following the convolutional operations. Finally, a novel highresolution detection head was designed to enhance the recognition capability for small objects.

ISSN: 2664-9640

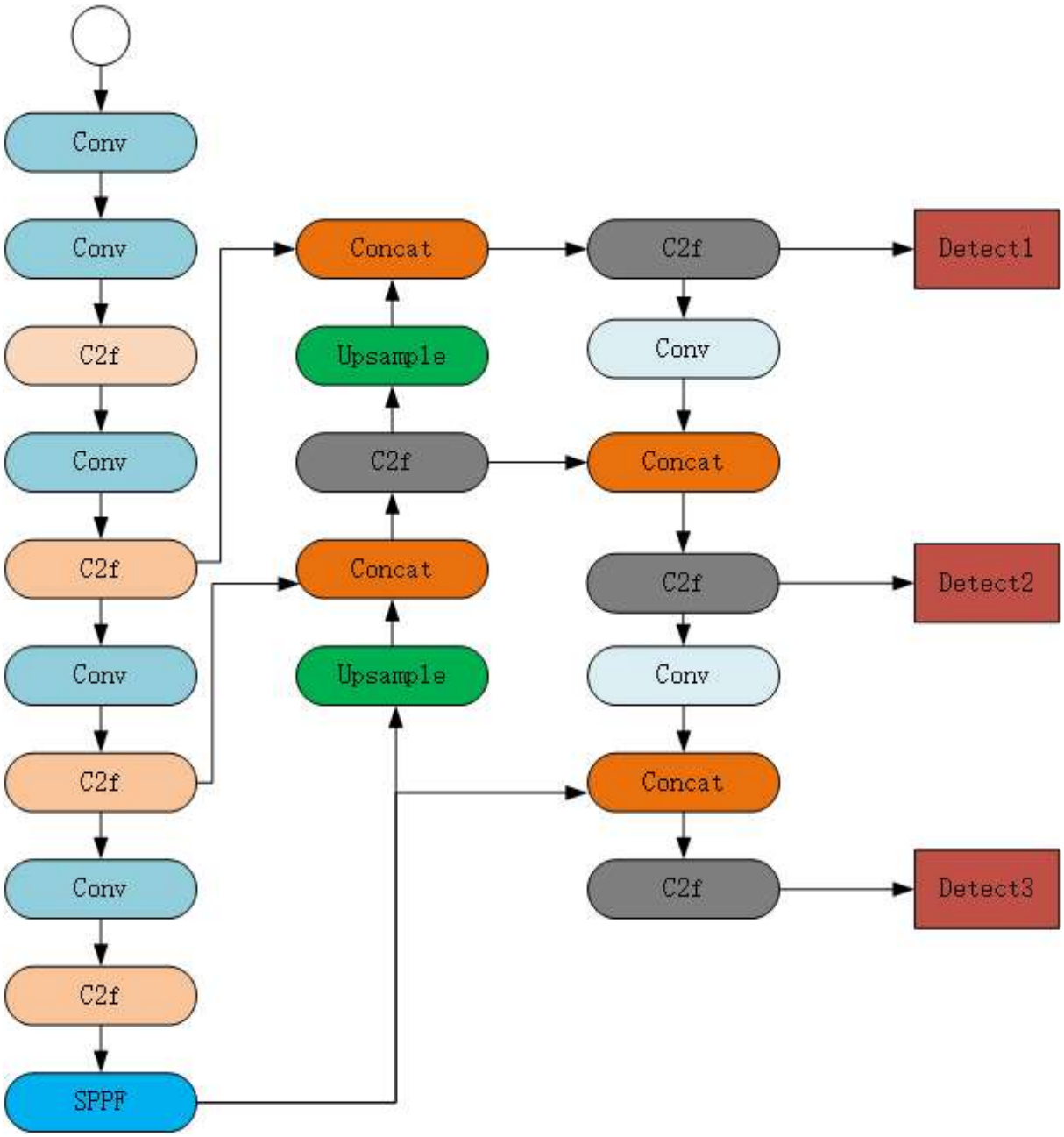

Fig. 1 YOLOv8

3.2.1. SCDConv

In conventional convolutional operations, information loss occurs due to the presence of convolution strides and pooling layers during feature extraction. While this loss may be acceptable for larger objects, it significantly hampers the detection of small objects by exacerbating the difficulty in detecting objects with limited pixels. To mitigate the impact of information loss and insufficient feature learning in detecting small hazardous items, this study was inspired by the SPDConv module[21] and devised a module named SCDConv. SPDConv addresses fine-grained feature loss by eliminating convolution strides and pooling layers. Illustrated in Fig. 2, the SPDConv structure replaces traditional convolution operations with spatial-to-depth (SPD) layers and replaces pooling layers with strideless convolution layers. This substitution scheme aids in preserving fine-grained information and enhancing feature learning capability.

ISSN: 2664-9640

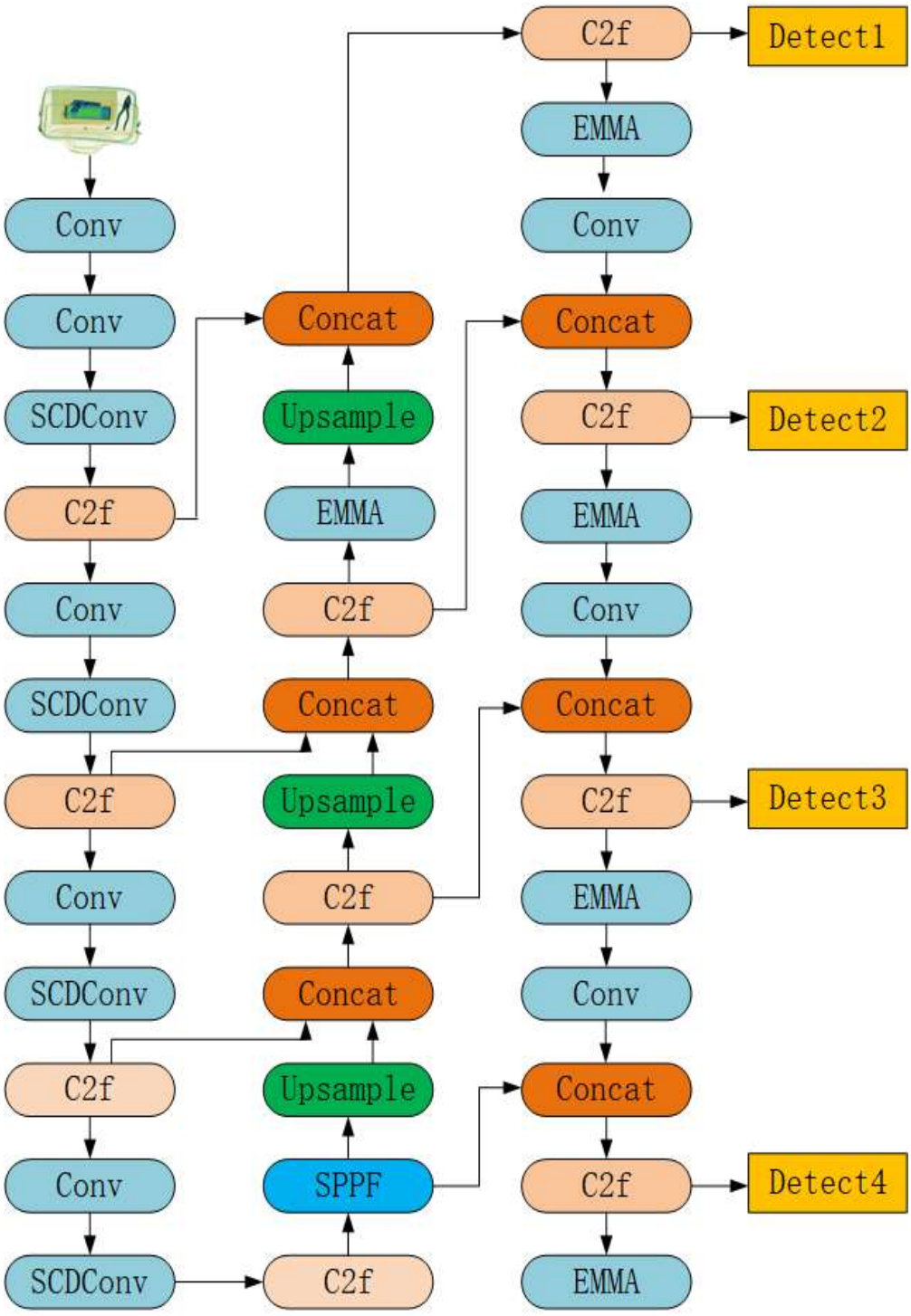

Fig. 2 SE-YOLO

Following each SPD layer, a strideless convolution operation is introduced to minimize the increase in the number of channels in the augmented convolutional layers, leveraging learnable parameters. In Fig. 3(a)(b)(c), four sub-maps with shapes of (S/2,S/2,C1) are generated, resulting in a twofold downsampling of X. Subsequently, these sub-feature maps are concatenated along the channel dimension to yield a feature map X', with reduced spatial dimensions by a scale factor and increased channel dimensions by a factor of 2. The strideless convolution operation added after the SPD layer adjusts the channel dimension of the feature map X' to C2, aiming to preserve all feature information to the maximum extent.

ISSN: 2664-9640

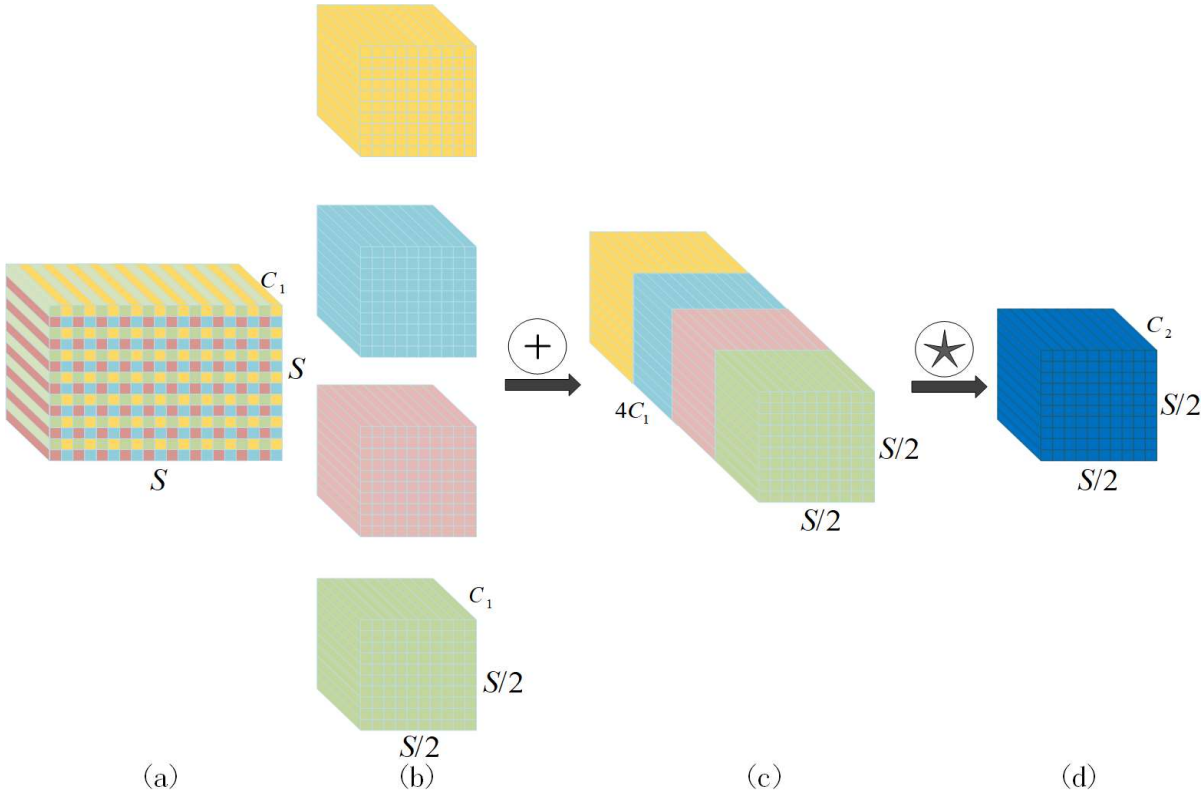

Fig. 3 SPDConv

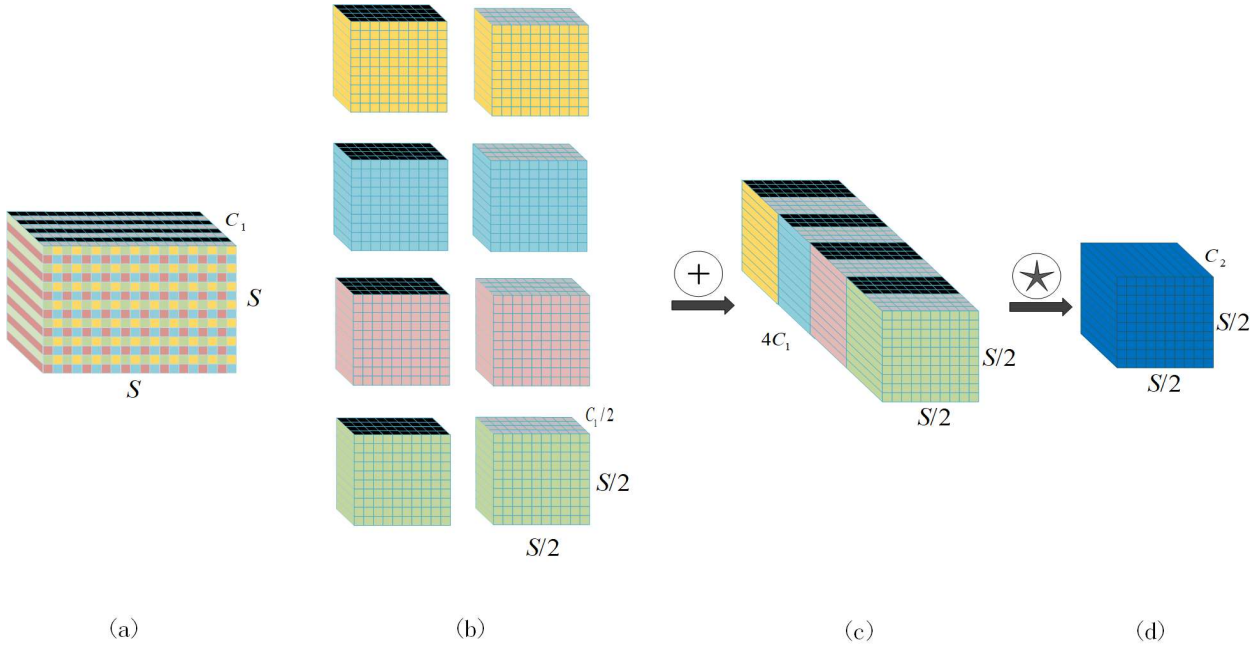

Fig. 4 SCDConv

Derived from SPDConv, SCDConv's structure is depicted in Fig. 4. In a similar vein to SPDConv, SCDConv also conducts downsampling of channel dimension information. Feature information is segregated into spatial and channel dimensions, and their interaction is facilitated through the Channel-Space to Depth (SCD) layer. This design effectively captures multidimensional details and enhances the extraction of fine-grained features. Illustrated in Fig. 4(a)(b)(c), with a scale factor of 2, eight sub-maps are derived, each with shapes of (S/2, S/2, C1/2), leading to a twofold downsampling of X. These sub-feature maps are subsequently concatenated along the

ISSN: 2664-9640

channel dimension to yield a new feature map X'. Unlike SPDConv, the channel dimension of the feature map X' obtained by SCDConv is reorganized based on the channel dimensions of the eight sub-maps.

Through the incorporation of SCDConv, the model proficiently processes feature mappings and mitigates the degradation of fine-grained information. The amalgamation of sub-maps and the adoption of stride-free convolution methodologies concurrently reduce spatial dimensionality while augmenting the model's feature expression capabilities.

3.2.2. EMMAattention

Beyond merely mitigating feature information loss, fostering the interaction of longdistance feature information across diverse dimensions emerges as a pivotal strategy for enhancing network model detection capabilities. To facilitate the comprehensive interaction of dimension information and amplify feature representation, this paper introduces an improved module, termed the Efficient Multi-channel Multi-scale Attention (EMMA) module, building upon the Efficient Multi-scale Attention (EMA) module[9]. The EMMA module reconfigures partial channels into batch dimensions and partitions the channel dimension into multiple sub-features, ensuring homogeneous distribution of spatial semantic features within each feature group. Initially, global information is encoded to derive channel weights for each parallel branch, followed by further amalgamation of feature information from two parallel branches through cross-dimensional interactions. Furthermore, the EMMA module applies grouping operations to each channel, thereby segmenting the feature space along the channel dimension.

Illustrated in Fig. 5, the EMMA module incorporates "X Avg Pool" and "Y Avg Pool," representing global pooling operations along the one-dimensional horizontal and vertical axes, respectively. Within this module, the input is segregated based on channel dimensions and processed through distinct branches. While one branch executes one-dimensional global pooling, the other employs 3x3 convolutions for feature extraction. Subsequently, the outputs from these branches undergo modulation via sigmoid functions and normalization operations. They are then merged through the cross-dimensional interaction module to capture pairwise relationships at the pixel level. Following final sigmoid modulation, the output feature maps either amplify or attenuate the original input features, culminating in the ultimate output.

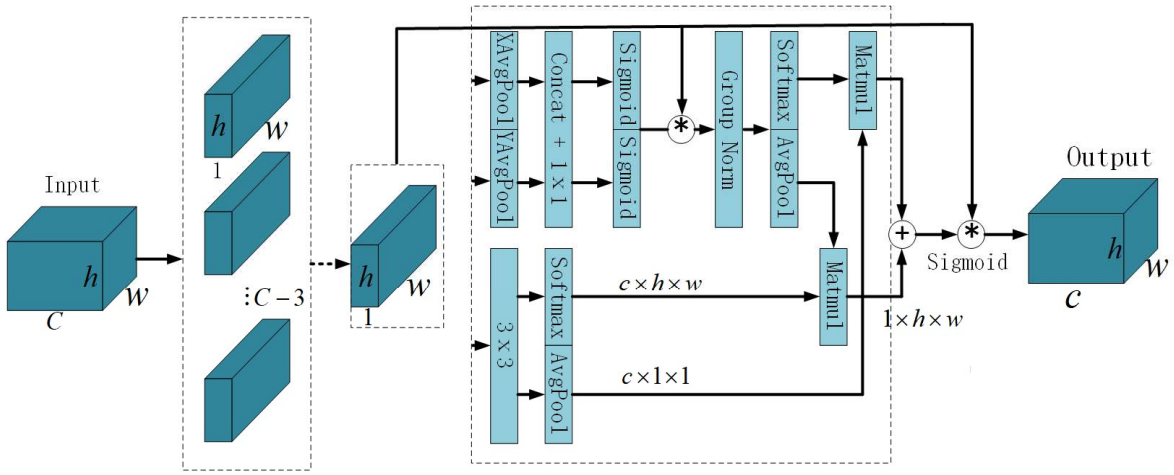

Fig. 5 EMMA

The Efficient Multi-channel Multi-scale Attention (EMMA) module enhances feature representation through the restructuring of channel dimensions, utilization of cross-

ISSN: 2664-9640

dimensional interactions to capture pixel-level relationships, and integration of global information within parallel branches to facilitate information exchange across various dimensions. This design significantly augments the model's performance and expressive prowess. The EMMA module proficiently captures feature correlations and adjusts weights to enhance or diminish the contribution of different features, thereby yielding a more distinctive feature representation.

3.2.3. Small Object Detection Head

Given the small dimensions of object samples and the significant downsampling factor of the YOLOv8 model, its native version exhibits a compromised ability to detect small objects. This limitation arises from the difficulty of deeper feature maps in effectively capturing the features of such objects. In the original model, characterized by an input image size of 640×640 and a minimum detection resolution of 80×80, the receptive field of each grid measures 8×8. Consequently, if both the height and width of objects in the original image are less than 8 pixels, the original network faces challenges in discerning the object feature information within the grid.

Consequently, this paper proposes a technique to enhance the detection performance of small objects by introducing an additional small object detection layer into the existing network architecture. This layer introduces a 160×160 scale for small object detection, incorporating an additional fusion feature layer and an extra detection head to augment the semantic information and feature representation capability of small objects. As depicted in Figure 2, the 80×80 scale feature layer from the fifth layer of the Backbone network is integrated with the upsampling feature layer from the Neck network. Following the application of C2f and upsampling techniques, a profound semantic feature layer containing detailed small object features is produced. Subsequently, this deep semantic feature layer is merged with the shallow positional feature layer from the third layer of the Backbone network, supplementing and completing a fusion feature layer scaled at 160×160, aimed at expressing the semantic features and positional data of small objects. Finally, this fusion feature layer is directed into an additional detection head with a resolution of 160×160 via C2f. The 160×160 resolution facilitates the model in generating 4×4 scale grids, thereby improving the detection efficiency for objects in the original image with dimensions smaller than 8 pixels.

The addition of the Head section allows the feature information of small objects to be transmitted through the network's downsampling path to the feature layers of the other three scales, thereby enhancing the network's feature fusion capability and improving the accuracy of small object detection. The introduction of additional detection heads serves to broaden the detection range for hazardous items. Consequently, the network can more accurately identify small-sized hazardous items in the image, thereby improving both detection accuracy and range.

4. Experiment

4.1. Experimental Environment and Data Sets

The hardware setup utilized in the experiments features an Intel(R) Core(TM) i5-13400F CPU, 16GB of RAM, and an RTX 3060 GPU with 12GB of VRAM. Operating on Windows 11, Python 3.10, and PyTorch 2.1 deep learning framework, the computational processes are accelerated using CUDA 10.1.

In this study, experiments were conducted utilizing the SIXray dataset, a collaboration between the Institute of Automation, Chinese Academy of Sciences, and the University of Science and Technology of China[14]. The dataset consists of 8,929 Xray images captured in real-world security inspection scenarios, divided into training, validation, and test sets in an 8:1:1 ratio. It

ISSN: 2664-9640

encompasses five categories of hazardous items: wrenches, handguns, knives, pliers, and scissors. The images portray objects with diverse poses, sizes, rotation angles, and levels of occlusion to simulate authentic security inspection scenarios. Detailed annotation information, including object categories, bounding box positions, and image-level labels, is provided within the SIXray dataset.

During the experiment, the training regimen consisted of 300 epochs with a batch size of 16 and an initial learning rate of 0.01. The resolution of input images was uniformly adjusted to 640×640 for consistency. These parameters were chosen to fully exploit the dataset's information and accommodate the network model's input specifications. With this setup, we could precisely evaluate and compare the model's performance on the SIXray dataset.

4.2. Evaluation Index

This paper assesses the algorithm using widely accepted metrics in object detection, such as Precision (P), Recall (R), mean Average Precision (mAP), and Average Precision (AP). These metrics are instrumental in evaluating the performance and efficacy of object detection algorithms. Precision and Recall offer insights into the accuracy and completeness of the detection outcomes, while mAP provides a comprehensive evaluation by considering Precision and Recall across different object categories. AP is employed to measure algorithmic performance across varying object categories.

Precision and Recall are determined using Equations (1) and (2), respectively. True Positive (TP) signifies instances where the prediction is positive and accurately labeled as positive. False Positive (FP) denotes instances where the prediction is negative but incorrectly labeled as positive. True Negative (TN) represents instances where the prediction is positive and correctly labeled as negative. False Negative (FN) indicates instances where the prediction is negative but mistakenly labeled as positive.

$$p = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$AP = \frac{1}{m}\sum_{i}^{m} P_i = \int p(R)dR \tag{3}$$

$$mAP = \frac{1}{c}\sum_{j}^{C} AP_j \tag{4}$$

The calculation equations for AP and mAP are outlined in Equations (3) and (4), respectively. Assuming there are n samples in a particular class of hazardous items, with m being positive samples, each positive sample corresponds to a Recall value R. To compute the AP for a specific class of hazardous items, the maximum Precision value P is determined for each Recall value, and the average of these m P values is obtained. If there are C classes of hazardous items in total, the average AP across these C classes is defined as mAP. Additionally, this paper evaluates the model's size and inference speed using metrics such as the number of parameters (params), model size (size), and Frames Per Second (FPS) for inference speed.

ISSN: 2664-9640

4.3. Experimental Results and Analysis

To validate the effectiveness of the proposed model in detecting hazardous items in X-ray security inspection images, this research conducted comparative experiments on the SIXray dataset with prominent object detection algorithms, including Faster R-CNN, Mask R-CNN[5],SSD512,DETR[32], YOLOv3, YOLOv5s, and YOLOX. The experimental outcomes are summarized in Table 1.Compared to the baseline model YOLOv8n, the proposed model in this paper achieved improvements of 2.1% and 3.5% in mAP50 and mAP50-95, respectively. This suggests that SE-YOLO effectively enhances the model's ability to focus on small-sized objects, resulting in enhanced detection accuracy for Gun, Knife, Wrench Pliers, and Scissors by 0.2%, 3.3%, 2.7%, 1.3%, and 3.1%, respectively. Since the proposed model in this paper is based on YOLOv8n, it outperforms YOLOv5, YOLOv3, and YOLOX in terms of detection effectiveness. The mAP improvements over these algorithms are 6.4%, 8.4%, and 3.3%, respectively. Additionally, compared to other widely used object detection algorithms such as Faster R-CNN, SSD, and DETR, this paper's model demonstrates notable advancements in both accuracy and inference speed.

The comparative analysis between the proposed model and the baseline model in detecting security inspection images is illustrated in Figure 6. In security inspection images, there are numerous subtle edges caused by the overlapping and haphazard placement of hazardous items. As depicted in Figure 6, the original YOLOv8n model exhibits inadequate performance in detecting areas with overlapping and significant occlusion, often resulting in missed detections and false positives. However, the proposed model in this paper integrates fine-grained information extraction from SCDConv and the small object capture capability of the high-resolution detection head, enabling effective detection of heavily occluded and overlapping objects. Additionally, the incorporation of EMMA attention in the neck network promotes multiscale information interaction, thereby improving the detection performance for hazardous items of various sizes.

Table 1. Performance Comparison

Model	Gun (%)	Knife (%)	Wrench(%)	Pliers(%)	Scissors(%)	mAP50 (%)	mAP50-95 (%)
FasterR-CNN	90.1	80.0	79.3	58.3	88.3	84.6	49.3
SSD	88.6	72.1	63.4	76.8	82.7	76.7	45.8
DETR	86.6	69.2	65.1	64.3	82.4	77.3	46.3
YOLOv3	89 7	81 2	78 9	62 1	84 8	83 9	49 0
YOLOv5s	88.3	84.0	82.1	82.0	90.6	85.8	52.9
YOLOX	89.1	86.7	85.7	90.2	93.5	89.0	57.4
YOLOv8n	98.4	86.3	88.2	93.9	84.0	90.2	65.9
OURS	98.6	89.6	90.9	95.2	87.1	92.3	69.5

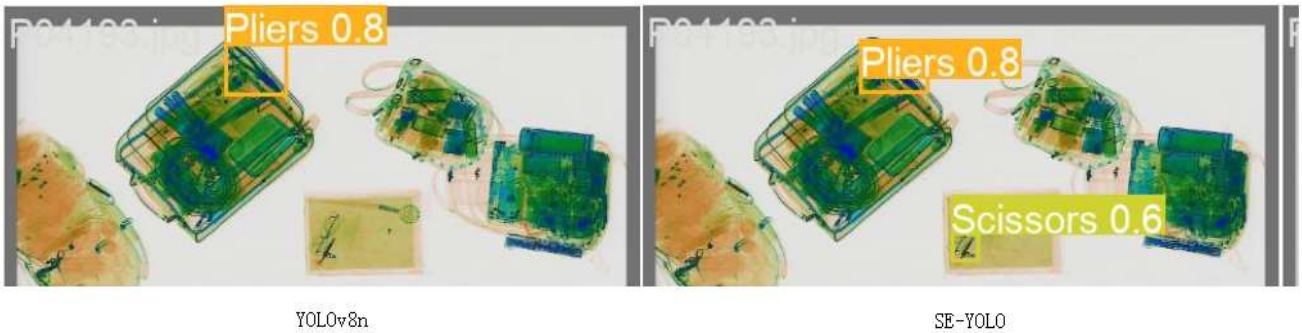

Fig. 6 Compare results

ISSN: 2664-9640

4.4. Ablation Experiment

To validate the effectiveness of the proposed enhancements by each module, ablation experiments were performed on the SIXray dataset to assess the influence of SCDConv, EMMA, and the high-resolution detection head. The results are tabulated in Table 2.

The first row of Table 2 denotes the results obtained by the baseline model. The integration of the EMMA attention mechanism facilitated the interaction of multidimensional information in the feature space, resulting in an expanded receptive field and a 0.6% increase in mAP. The design of the high-resolution detection head enabled the model to detect hazardous items smaller than 8×8 pixels, leading to a 0.9% improvement in mAP. Furthermore, the introduction of SCDConv in the backbone network bolstered the model's fine-grained feature extraction capability. The amalgamation of these enhancement modules enabled the model to effectively handle both large and small objects, achieving a comprehensive detection accuracy of 92.3%, which represents a 2.1% improvement over the baseline model. Thus, the model proposed in this study has demonstrated significant improvements across all aspects, achieving high detection accuracy.

Table 2. Ablation experiment

Model	mAP50 (%)	mAP50-95 (%)
Baseline	90.2	65.9
Baseline+ EMMA	90.8	66.3
Baseline+ EMMA+ Head	91.1	66.6
SE- YOLO	92.3	69.5

5. Conclusion

This study introduces a novel algorithm, SE-YOLO, for detecting hazardous items in security inspection images, with the aim of improving the accuracy of automatic hazardous item identification. The specific enhancements comprise three aspects: firstly, the incorporation of the SCDConv module, which effectively extracts fine-grained features from the images, thereby improving the model's accuracy in identifying smallsized hazardous items. Secondly, the integration of the EMMA attention mechanism enhances the model's feature fusion operation, facilitating effective multidimensional information interaction and augmenting the model's perceptual capability in the feature space. Thirdly, a novel high-resolution small object detection head is devised, capable of generating anchor boxes better suited for small objects, thereby improving the model's ability to capture small objects. Experimental validation on the SIXray dataset demonstrates that the proposed model achieves a detection accuracy of 92.3%. These findings indicate that the proposed enhancement algorithm holds potential in hazardous item detection tasks and offers valuable insights for practical applications.

Nonetheless, this study identifies areas for improvement that demand further exploration and refinement. Although the model has demonstrated improved detection accuracy, there remains potential for optimizing its parameters and computational requirements. Potential avenues for enhancement include fine-tuning the network architecture, improving feature extraction methodologies, and implementing more efficient training strategies.

References

[1] Chensheng Cheng, Can Wang, Dianyu Yang, Xin Wen, Weidong Liu, and Feihu Zhang. Underwater small target detection based on dynamic convolution and attention mechanism. Frontiers in Marine Science, 11:1348883, 2024.

ISSN: 2664-9640

[2] Yishan Dong, Zhaoxin Li, Jingyuan Guo, Tianyu Chen, Shuhua Lu, et al. An improved x-ray contraband detection model for yolov5. Laser & Optoelectronics Progress, 60(4):0415005–0415 005, 2023.

[3] Ross Girshick. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 1440–1448, 2015.

[4] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580–587, 2014.

[5] Kaiming He, Georgia Gkioxari, Piotr Doll´ar, and Ross Girshick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017.

[6] Deao Hu, Mei Yu, Xianyong Wu, Jingbo Hu, Yuyang Sheng, Yanjing Jiang, Chongjing Huang, and Yuelin Zheng. Dgw-yolov8: A small insulator target detection algorithm based on deformable attention backbone and wiou loss function. IET Image Processing, 18(4):1096–1108, 2024.

[7] Yi Huang, JiaYuan Fan, Yong Hu, Jinmeng Guo, and Yongjian Zhu. Tbi-yolov5: A surface defect detection model for crane wire with bottleneck transformer and small target detection layer. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 238(6):2425–2438, 2024.

[8] Mikolaj E Kundegorski, Samet Ak¸cay, Michael Devereux, Andre Mouton, and Toby P Breckon. On using feature descriptors as visual words for object detection within x-ray baggage security screening. 2016.

[9] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In Proceedings of the IEEE/CVF international conference on computer vision, pages 9167–9176, 2019.

[10] Yongjian Li, Huasheng Zhu, Mingzhi He, Shuyin Tang, and Zhanxin Sun. A yolocda x-ray image detection algorithm. In Proceedings of the 2023 6th International Conference on Image and Graphics Processing, pages 168–174, 2023.

[11] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pages 21–37. Springer, 2016.

[12] Chunjie Ma, Li Zhuo, Jiafeng Li, Yutong Zhang, and Jing Zhang. Occluded prohibited object detection in x-ray images with global context-aware multi-scale feature aggregation. Neurocomputing, 519:1–16, 2023.

[13] Ping Ma, Xinyi He, Yiyang Chen, and Yuan Liu. Isod: improved small object detection based on extended scale feature pyramid network. The Visual Computer, pages 1–15, 2024.

[14] 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.

[15] Shuo Miao, Xinwei Li, Yi Yang, Keping Wang, and Kefei Cui. Contraband detection in x-ray images based on improved capsule network. Journal of Henan Polytechnic University: Natural Science, 42(3):129–136, 2023.

[16] Stefan Michel, Saskia M Koller, Jaap C De Ruiter, Robert Moerland, Maarten Hogervorst, and Adrian Schwaninger. Computer-based training increases efficiency in x-ray image interpretation by aviation security screeners. In 2007 41st Annual IEEE international Carnahan conference on security technology, pages 201–206. IEEE, 2007.

[17] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.

ISSN: 2664-9640

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28, 2015.

[19] Bingshan Su, Shiyong An, Xuezhuan Zhao, Jiguang Chen, Xiaoyu Li, and Yuantao He. An aeronautic x-ray image security inspection network for rotation and occlusion. International Journal of Computational Science and Engineering, 26 (3):337–348, 2023.

[20] Li Sun, Zhenghua Cai, Kaibo Liang, Yuzhi Wang, Wang Zeng, and Xueqian Yan. An intelligent system for high-density small target pest identification and infestation level determination based on an improved yolov5 model. Expert Systems with Applications, 239:122190, 2024.

[21] Raja Sunkara and Tie Luo. No more strided convolutions or pooling: A new cnn building block for low-resolution images and small objects. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 443–459. Springer, 2022.

[22] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10781–10790, 2020.

[23] Breckon T P. Turcsany D, Mouton A. Improving feature-based object recognition for x-ray baggage security screening using primed visual words, 2013. Proceedings of the 2013 IEEE International Conference on Industrial Technology,Cape Town,Feb 25-28,2013:1140-1145.

[24] Danijela Vukadinovic, Miguel Ruiz Os´es, and David Anderson. Automated detection of inorganic powders in x-ray images of airport luggage. Journal of transportation security, 16(1):3, 2023.

[25] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, JunWei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pages 390–391, 2020.

[26] Ruxue Wang, Yuliang Shi, and Mingyu Cai. Optimization and research of suspicious object detection algorithm in x-ray image. In 2023 IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), volume 6, pages 1357–1361. IEEE, 2023.

[27] Suxue Wu. Object detection algorithm for contraband in x-ray security inspection images based on improved yolov5. Journal of Modern Computers, 28(20), 2022.

[28] Xuanrui Xiong, Mengting He, Tianyu Li, Guifeng Zheng, Wen Xu, Xiaolin Fan, and Yuan Zhang. Adaptive feature fusion and improved attention mechanism based small object detection for uav target tracking. IEEE Internet of Things Journal, 2024.

[29] Xiaoyu Yu, Wenjun Yuan, and Aili Wang. X-ray security inspection image dangerous goods detection algorithm based on improved yolov4. Electronics, 12(12): 2644, 2023.

[30] Wenming Zhang, Qikai Zhu, Yaqian Li, and Haibin Li. Mam faster r-cnn: Improved faster r-cnn based on malformed attention module for object detection on x-ray security inspection. Digital Signal Processing, 139:104072, 2023.

[31] Zhu J. Zhang N. A study of x-ray machine image local semantic features extraction model based on bag-of-words for airport security. International Journal on Smart Sensing & Intelligent Systems, pages 1–8, 2015.

[32] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159, 2020.