MobileViT Based Lightweight Model for Prohibited Item Detection in X-Ray Images

Peng Sun[1,2], Haigang Zhang[1(✉)], Jinfeng Yang[1], and Dong Wei[2]

[1] Shenzhen Polytechnic University, Shenzhen, China
zhg2018@sina.com
[2] University of Science and Technology Liaoning, Anshan, China

Abstract. The application of computer vision technology to detect prohibited items in X-ray security inspection images holds significant practical research value. We have observed that object detection models built upon the Visual Transformer (ViT) architecture outperform those relying on Convolutional Neural Networks (CNNs) when assessed on publicly available datasets. However, the ViT's attention mechanism, while offering a global response, lacks the CNN model's inductive bias, which can hinder its performance, demanding more samples and learning parameters. This drawback is particularly problematic for time-sensitive processes like security inspections. This research paper aims to develop a lightweight prohibited item detection model grounded in the ViT framework, utilizing MobileViT as the underlying network for feature extraction. To enhance the model's sensitivity to small object features, we have established dense connections among various network layers. This design ensures effective integration of both high- and low-level visual features without increasing computational complexity. Additionally, learnable group convolutions are employed to replace traditional convolutions, further reducing model parameters and computational demands. Simulation experiments conducted on the publicly available SIXray dataset validate the effectiveness of the proposed model in this study. The code is publicly accessible at https://github.com/zhg-SZPT/MVray.

Keywords: Object detection · Lightweight model · Visual Transformer · X-ray security inspection

1 Introduction

X-ray security inspections are ubiquitous in airports, railway stations, large event venues, and similar settings, serving to detect any prohibited items that individuals may be carrying in their luggage. The use of computer vision technology for X-ray image analysis has been drawing increasing attention in both industry and academia. In terms of practical application, the intelligent analysis of

This work was supported by the School-level Project of Shenzhen Polytechinc University (No. 6022310006K), Research Projects of Department of Education of Guangdong Province (No. 2020ZDZX3082, 2023ZDZX1081, 2023KCXTD077).

H. Lu et al. (Eds.): ACPR 2023, LNCS 14407, pp. 45–58, 2023.
https://doi.org/10.1007/978-3-031-47637-2_4

X-ray security inspection images falls within the realm of object detection. Such a model is expected to identify both the type and location of a prohibited item. X-ray security inspection images exhibit unique characteristics that often lead to the failure of traditional object detection algorithms. For one, items in the X-ray images are positioned randomly. The penetrative nature of X-rays results in an overlapping effect among different objects. Additionally, prohibited items come in a variety of types and scales, with smaller ones posing the most significant challenge to the model and often leading to detection failures. In light of these technical complexities inherent in the analysis of X-ray security inspection images, traditional machine learning algorithms often fall short. Conversely, deep learning technologies, notably those that involve Convolutional Neural Networks (CNNs), have shown remarkable results in detecting prohibited items in X-ray images [1].

Research into detecting prohibited items in X-ray security inspection images within the deep learning framework primarily encompasses two aspects. On the one hand, there are efforts aimed at developing targeted detection models, taking into account the unique characteristics of X-ray images. For example, [2] implemented more intricate Deep Neural Network (DNN) structures to alleviate the impact of cluttered backgrounds on model performance. Other research, including those conducted by [3,4], focused on addressing the issue of sample imbalance in security inspection datasets to enhance the detection recall rate for prohibited items. The intelligent analysis of X-ray security inspection images also grapples with the challenge of multi-scale object detection. Certain works have managed to maintain the visual features of small-scale objects within the network's high-level semantics through feature fusion and refinement [4,5]. On the other hand, security inspection operations are subject to certain time constraints, necessitating the consideration of the inference speed of visual detection models. A common contention, although not rigorously proven, suggests that increasing the complexity of DNN structures to enhance model feature extraction capability may conflict with the efficiency of inference. In response, some researchers are dedicated to designing lightweight intelligent analysis models for X-ray security inspection images [6,7]. Notably, we have proposed a lightweight object detection framework based on the YOLOv4 algorithm, incorporating MobileNetV3 [8] as the feature extraction backbone [9]. This approach has demonstrated state-of-the-art performance in prohibited item detection within X-ray images.

This paper focuses on the design of lightweight models for prohibited item detection in X-ray images. An effective and practical approach to enhance model inference efficiency is model compression, encompassing techniques like pruning, quantization, lightweight structural design of complex models, and knowledge distillation [10]. Model compression refers to an array of techniques aiming to diminish the size, complexity, and computational demands of a Deep Neural Network (DNN) model, while striving to preserve its performance or impact it minimally. These techniques include weight pruning, quantization, and network architecture search. Lightweight architecture design targets the creation of neural network structures specifically tailored to have fewer parameters, lower

computational complexity, and a smaller memory footprint, while still delivering competitive performance. Models such as the MobileNet series [8,11,12] and ShuffleNet [13,14] fall within the category of lightweight CNN models. Distinct from the preceding two strategies, knowledge distillation mimics the teacher-student dynamic, wherein a less complex and smaller "student" model is trained to replicate the behavior of a larger and more complex "teacher" model.

Recently, researchers have attempted to apply the Transformer model, which has proven effective in natural language processing, to the field of computer vision. Vision Transformer (ViT) [15] represents a pioneering application of Transformer attention in the realm of vision, which segments an image into flattened 2D patches and subsequently projects each patch into tokens linearly. ViT model leverages self-attention mechanism for global interaction responses among different patch tokens. Unlike the CNN model, the ViT model emphasizes global feature extraction and lacks inductive bias, therefore demanding more data and learning parameters to ensure optimal model performance. For instance, when compared to the classic lightweight model MobileNetV3 [8], the ViT model boasts 16 times more parameters (85M in ViT-B versus 4.9M in MobileNetV3).

We noticed that MobileViT [16], a lightweight ViT model, integrates the feature extraction capabilities of CNN and ViT and shows impressive performance across multiple visual tasks. This paper employs MobileViT with the YOLO detection head to construct a lightweight detection model for prohibited items in X-ray security inspection images. To enhance model performance, two key improvements have been made. Firstly, we modify the connection mode of the feature extraction blocks in MobileViT to dense connections [17]. This change ensures that the model can achieve cross-scale feature fusion without increasing the computational load, and effectively solve the problem of missed detection for smaller-sized prohibited items. Secondly, the Learning Group Convolution (LGC) [17] is applied into the feature extraction backbone to further reduce the network's learning parameters. The proposed model, tested on the SIXray dataset [18], achieves an mAP of 81.21% with 5.64M learning parameters and 2.72G FLOPs. When compared to CSPDarknet [19], the proposed model reduces the amount of learning parameters and FLOPs by 20.96M and 7.54G, respectively.

2 Related Work

2.1 Vision Transformer

Transformer was first applied to natural language processing (NLP) tasks [20], and ViT [15] is a pure transformer directly applies to the sequences of image patches for image classification tasks. The pure Transformer operation lacks the inductive bias ability of CNN, so on medium or small data sets (ImageNet), the performance of ViT is weaker than that of ResNet model. However, when the ViT model is trained on a larger dataset, its performance surpasses the inductive bias ability of CNN [21]. Currently, in order to further improve the performance

of ViT in visual feature extraction, some work combines self-attention and convolution, such as CPVT [22], CvT [23], CeiT [24]. Combining the transformer with convolution can effectively introduce the locality into the conventional transformer. MobileViT [16] is a lightweight version of ViT, which embeds the Transformer block into MobileNet, so that the model has the ability to extract local and global features. MobileViT achieves top-1 accuracy of 78.4% with about 6 million parameters, which is 3.2% more accurate than MobileNetv3 for a similar number of parameters. On the MS-COCO object detection task, MobileViT is 5.7% more accurate than MobileNetv3.

2.2 Prohibited Item Detection in X-Ray Images

From a practical point of view, intelligent analysis of X-ray security inspection images belongs to the task of object detection. In addition to knowing the category of prohibited item, the security inspectors are also very interested in the location information of the prohibited item. Research on prohibited item detection based on computer vision is very common. Mery [25] reviewed the application of computer vision technology in X-ray security inspection image analysis, and pointed out that the object detection technology based on deep learning outperforms traditional algorithms in terms of accuracy and generalization. Deep learning technology in the security field can be divided into three categories according to different needs and application scenarios. They are classification, detection and segmentation [26]. X-ray security package images often suffer from background clutter and target occlusion, compelling researchers to develop more intricate, high-performance network structures [27,28]. However, these complex structures frequently result in slower inference speeds, which could negatively impact security inspection process operations. Consequently, research focused on lightweight object detection models is garnering increased attention within the realm of X-ray security image analysis [9]. The crux of designing a lightweight object detection model lies in avoiding miss-detections, particularly for smaller prohibited items, while simultaneously maintaining a streamlined model structure.

3 MVray Method

In this research, we put forth a novel lightweight object detection model coined as MVray, which is designed explicitly for the analysis of prohibited items in X-ray images. Figure 1 delineates the architectural framework of the MVray model, which employs the MobileViT network [16] as the foundational backbone for feature extraction, and collaboratively interfaces with the YOLO head to execute object detection tasks. The MobileViT network unites the local features extracted by CNN with the global features of the Transformer, thereby amplifying its feature extraction capabilities. While lightweight networks generally falter in detecting smaller objects, to enhance the feature extraction proficiency of our backbone network for X-ray security inspection imagery, we integrate

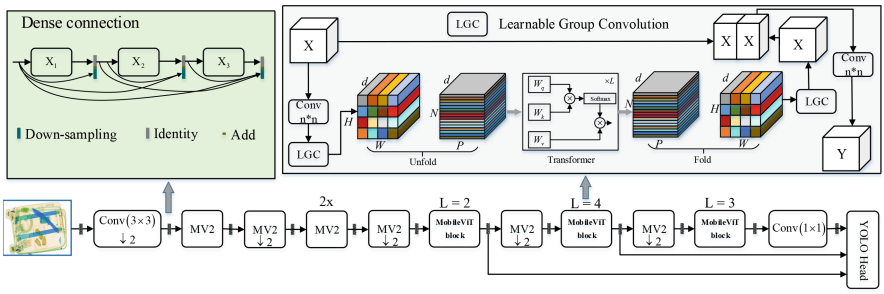

Fig. 1. MVray framework

dense connections within the feature extraction modules [17]. It's notable that these dense connections linearly aggregate features across diverse scales without substantially escalating the computational load of the model. Moreover, to prune redundant features and preserve, or even augment, the lightweight advantage of the backbone network, we employ the *LGC* to supersede traditional convolution operations [17].

3.1 MobileViT Backbone

MobileViT is a potent and streamlined network model, grounded in the MobileNet architecture, which incorporates a Transformer-based self-attention mechanism within CNN feature extraction layers. Traditional ViTs necessitate substantial data and computational power to parallel the performance of CNN models via a multi-head self-attention mechanism, given that Transformer lacks the spatial induction bias innate to CNN. The central philosophy behind MobileViT is to utilize transformers in lieu of convolutions for global representation learning. As delineated in Fig. 1, Transformer blocks are intricately woven into the fabric of the MobileViT network.

As shown in Fig. 1, the Transformer block in MobileViT is a plug and play module. Given an input $X \in R^{C\times H\times W}$, the output Y of Transformer block can be calculated as

$$Y = LGC_{1\times 1}\left[X||Transformer\left(LGC_{1\times 1}\left(X\right)\right)\right] \tag{1}$$

where $Y \in R^{C\times H\times W}$, and C, H and W represent the channels, height, and width of the input respectively. $||$ represents the concatenation operation, $LGC_{1\times 1}(\cdot)$ represents the learnable group convolution operation with 1×1 convolution kernel.

Unlike the traditional Transformer attention mode, the Transformer operation in MobileViT requires fewer learning parameters. The LGC operation projects the input tensor to a high-dimensional space d $(d > C)$. Then Transformer attention is applied to establish the long-range dependencies, which is complementary to the local inductive bias of CNN. We unfold X into N non-overlapping flattened patches $X_U \in R^{d\times P\times N}$, where $P = wh$, and w and h

are the width and height of a patch. So $N = {}^{HW}/_{P}$. Then the self-attention operations are carried among the flattened patches. The Transformer block in MobileViT does not lose the spatial position information of each visual patch, so it is easy to fold X_U with self-attention to CNN-wise features $X_F \in R^{d \times H \times W}$. Concatenation operator is conducted to achieve feature fusion between the ordinary X and X_F, and another LGC implements channel matching to ensure plug-and-play of Transformer block in MobileViT.

3.2 Dense Connection and Learn-Group-Conv

Prohibited items captured in X-ray security inspection images often display multi-scale characteristics. However, lightweight networks often struggle to achieve satisfactory detection performance, especially for smaller objects. This shortcoming can be attributed to the limited local feature field inherent in the conventional CNN convolution kernel's feature extraction process. Moreover, pooling operations within CNN models tend to diminish the responses of smaller object features within higher-level semantics. One effective and straightforward approach to addressing these challenges is to facilitate feature fusion across various feature layers. This ensures that high-level semantic features incorporate elements of low-level visual details. Importantly, the chosen multi-level feature fusion strategy must maintain computational efficiency, aligning with the lightweight modeling principles underscored in this study. In the MVray model, we integrate dense connection operations [17] into the MobileViT backbone. This incorporation enhances the reusability of shallow features at higher layers, incurring minimal additional computational load on the network. However, it's important to note that dense connections could potentially lead to data redundancy through overly dense connection schemes. To mitigate this issue, the MVray model introduces the LGC module to replace the traditional convolution operation. This substitution results in further parameter reduction within the network, achieving a more lightweight architecture while concurrently addressing data redundancy concerns.

Dense Connection: The traditional CNN is dense because of its progressive design layer by layer. Each layer has to copy the features of earlier layers, which causes a large number of data redundancy. Residual structure design is a popular information fusion strategy used to solve the gradient vanishing problem caused by an increase in the number of hidden layers. ResNets [29], as the representative of residual structure, transfer signal from one layer to the next via identity connections. Unlike residual structure design, the dense connection [17] connects all layers directly with each other. Specifically, in dense connection each feature layer obtains additional inputs from all preceding layers, and then transmits the fusion features to all subsequent layers. We use X_p represents the features on the p layer, and the comparison of residual structure and dense connection is as follows.

$$\begin{aligned} &\text{ResnetConnect}: \quad X_p = F_p\left(X_{p-1}\right) + X_{p-1} \\ &\text{DenseConnect}: \quad X_p = F_p\left(\left|\left|\left[X_0, X_1, \cdots, X_{p-1}\right]\right.\right.\right) \end{aligned} \tag{2}$$

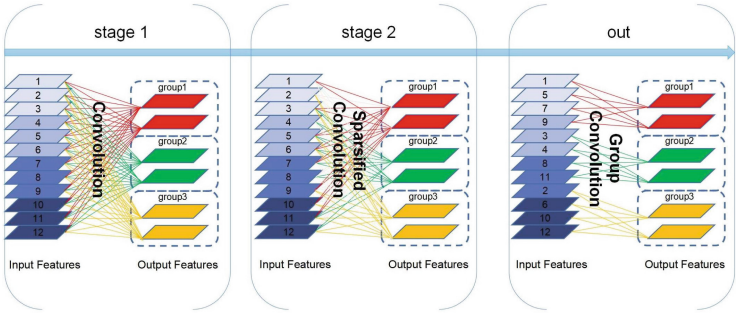

Fig. 2. The diagram of LGC operation with 3 groups.

where $F_p(\cdot)$ represents the composite function composed of convolution operation, batch normalization, and activation function, etc. $||$ is the concatenate operation.

The dense connection strategy integrates concatenation operations to facilitate effective feature fusion, achieving a balance between optimizing feature flow and reducing computational burden. This is accomplished by establishing direct connections between each layer and all of its preceding layers. Such a design minimizes the necessity for feature replication, leading to the accumulation of features from all levels. Consequently, it aids in compensating for potential missing attributes and contributes to improved overall feature representation. Differing from the dense connection mechanism as seen in DenseNet [17], the MVray model employs a distinctive approach to sub-sampling through global average pooling. This enables the model to execute addition-based feature fusion on feature maps of varying resolutions. Importantly, this approach preserves the overarching architectural structure of the model, underscoring its compatibility with the lightweight philosophy emphasized in this research.

Learn-Group-Conv: The *LGC* operation is a robust technique engineered to curtail redundant features in a neural network during the training process, and eliminate the need for pre-established hyperparameters. This operation proves especially beneficial for bolstering the efficiency and precision of models employing dense connections, which could engender feature redundancy. By enabling the model to discern which features are vital and which can be pruned, the *LGC* operation assists in maintaining the model's lightweight nature, whilst still facilitating potent feature extraction. As depicted in Fig. 1, we substitute traditional $1 * 1$ convolution with the *LGC* operation within the Transformer block.

The *LGC* operation is divided into two stages, as shown in Fig. 2, with the standard convolution on the left and the *LGC* on the middle and right. In the first stage, the input features are segregated into groups, each of which is subjected to a sparse regularization operation to filter out insignificant features. This process involves the addition of a penalty term to the loss function, motivating the network to learn a sparse representation of the input features. The penalty

term is designated as the Group-Lasso of the group convolution filter, pushing the filter towards zero weights. In the subsequent stage, following the pruning and fixing of insignificant features, the network undergoes further training. The surviving features are re-grouped, and the group convolution filters are learned via backpropagation. The grouping and filter learning occur in unison, ensuring that each filter can learn the most pertinent input set for the subsequent layers. Equation (3) replaces the traditional $L1$ regularized group-lasso as a criterion for filtering non-critical input features.

$$\sum_{g=1}^{G}\sum_{j=1}^{R}\sqrt{\sum_{i=1}^{O/G}\left(F_{i,j}^{g}\right)^{2}} \tag{3}$$

where G represents the final number of groups, O and R represent the number of output channels and input channels respectively. F is the weights learned by the convolution kernel. Each F^g from F^1 to F^G has size $O/G \times R$. $F_{i,j}^{g}$ corresponds to the weight of the jth input for the ith output within group g. We take the absolute value, that is $F_{i,j}^{g}$ squares and then takes the square root.

Traditionally, model sparsity is achieved through $L1$ regularization. In this paper, we leverage group-lasso, which accentuates the afferent features within the same group, to induce sparsity and ensure that convolution kernels within the same group utilize the same afferent feature subset. During the training process, the input feature subset within each group bearing lower weight is progressively pruned. The significance of the input feature to its corresponding group is determined by aggregating all corresponding weights within the group. To complete the convolution sparsity operation, input features of marginal importance in the convolution kernel are set to zero accordingly.

4 Experiment Results

4.1 Experimental Datasets and Implementation Details

We applied the publicly available SIXray dataset [18] to verify the performance of the proposed MVray algorithm. The SIXray dataset consists of 1,059,231 X-ray images containing 8,929 prohibited items in 5 categories (Gun, Knife, Pliers, Scissor and Wrench). The images in the SIXray dataset are all collected in real life scenarios. We split the training set, the verification set, and the test set in an 8:1:1 ratio.

All experiments are performed on PyTorch under the Ubuntu 16.04 system with NVIDIA A100-SXM4-40G GPU. We added the common feature extraction backbone networks to the comparison of model performance. In order to compare model performance fairly, all the models participating in the comparison experiment adopted the training hyperparameter in Table 1, and carried out training under the condition of zero pre-training weights.

Table 1. Model training parameters

Name	Parameter
input_shape	(320,320)
Optimizer	adam
Momentum	0.937
Lr_decay_type	cos
Init_lr	1e−3
Batch_size	16
Epoch	300
Num_works	4

4.2 Comparison Results

On the SIXray dataset, our proposed MVray model shows satisfactory performance. Table 2 shows the detection performance of MVray model for 5 types of prohibited items. Overall, the MVray model achieves the mAP value of 81.21% on the SIXray dataset. The AP indicators of individual categories of prohibited items also exceed 70%. The MVray model has the best detection performance for "Gun", achieving the 95.89% AP value.

We make comparison between the proposed MVray model with other lightweight feature backbones, such as MobileNet series [11], GhostNet [30], DenseNet [17], CSPDarknet53 [31] and MobileViT [16]. Table 2 presents the comparison results on SIXray dataset. The proposed MVray model achieves the best prohibited item detection performance with 81.21% mAP, which outperforms the DenseNet model by an absolute gain of 1.94% mAP, the GhostNet model by 9.32% mAP, the MobileNetV3 model by 11.02% mAP, the DenseNet model by 1.94%, and the MobileViT model by 3.96%. CSPDarknet53 is the basic feature extraction backbone for YOLO series models, which achieves the best detection performance with 85.42% mAP. However, CSPDarknet53 backbone network has far more learning parameters than MVray model, as shown in Table 3. For the detection of specific categories of prohibited items, MVray model has an overwhelming advantage in detecting "Scissor" and "Wrench" compared with other models. When detecting "Scissor", the MVray model has improved AP metrics by more than 20% points compared to the MobileNet series models. When detecting the "Wrench", the improvement intensity is about 14% points.

The MVray model aims to achieve a balance between object detection accuracy and efficiency. Therefore, model lightweight and inference efficiency are very important indicators. Table 3 shows the volume and computational efficiency of the model, where the Floating Point Operation (FLOPs), model Parameters (Params) and Frames Per Second (FPS) are taken into consideration. In order to ensure the uniqueness of the differences and achieve the comparison effect, YOLOv4 detection head is selected to make object detection. The proposed MVray model has the same number of learning parameters (5.64M) as Mobile-

Table 2. Comparison results among different feature extraction backbones.

Backbone	AP					mAP
	Gun	Knife	Pliers	Scissor	Wrench	
MobileNetV1	93.75%	75.82%	70.20%	52.07%	57.65%	70.90%
MobileNetV2	92.59%	73.64%	70.60%	52.17%	57.82%	69.36%
MobileNetV3	93.23%	76.08%	69.20%	56.32%	56.10%	70.19%
GhostNet	93.82%	75.38%	70.06%	60.43%	59.76%	71.89%
DenseNet	94.91%	76.64%	78.23%	77.37%	69.18%	79.27%
CSPDarknet53	97.23%	77.55%	84.73%	88.59%	79.00%	85.42%
MobileViT	95.30%	75.83%	77.79%	73.36%	63.98%	77.25%
MVray	95.89%	80.01%	79.67%	78.70%	71.78%	81.21%

ViT model, since neither dense connections nor LGC operations add additional parameters. MobileNetV2 has the least number of learning parameters (3.50M), but its detection performance is 12% mAP lower than the MVray model. MVray model contains 2.72G FLOPs, lower than MobileViT by 0.08G FLOPs. Due to the additional calculation of Transformer attention, the FLOPs index of the MVray model is higher than that of the MobileNet series models. Compared with CSPDarknet with 10.26G FLOPs, the Mvray model has relatively obvious advantages. The comparison of "Model Size" shows the similar results with "Params". The FPS indicator can reflect the inference speed of the model. MVray model achieves 38.84 FPS, which shows that the MVray model can fully meet the timeliness requirements of security inspection.

4.3 Ablation Experiment

This paper proposes the MVray model, which employs two additional operations to enhance the performance of prohibited item detection tasks in X-ray security inspection images, superior to the base MobileViT model. The implementation of a dense connection successfully fuses the visual features across various feature extraction layers, enabling the effective transmission of low-level features into high-level semantics. This crucial enhancement proficiently addresses the issue of missed detection of small targets prevalent in lightweight models. Additionally, the LGC technique further diminishes the model's learning parameters and computational load, while simultaneously minimizing feature redundancy brought about by dense connections. We conducted ablation studies to ascertain the contributions of dense connection and LGC operations, the outcomes of which are demonstrated in Table 4. The integration of dense connection and LGC into the MobileViT model significantly boosted its performance. The inclusion of the dense connection allowed the model to achieve the mAP of 78.79%, marking a 1.54% increase from the base MobileViT model. Furthermore, the introduction of LGC elevated the mAP to 77.90%, an improvement of 0.65% over the base

model. When combined, these operations enable the proposed MVray model to reach an impressive mAP of 81.21%, outstripping the MobileViT model by a considerable margin of 3.96% mAP.

In terms of specific prohibited item detection, the incorporation of dense connection and LGC operations significantly enhances the model's performance. Notably, MobileViT achieves the mAP of 63.98% in the detection of the prohibited item of "Wrench". With the introduction of dense connections and LGC operations, the model's performance improved considerably, obtaining 65.58% and 66.65% mAP, respectively. When these two operations were combined, the proposed MVray model displayed a marked improvement. Specifically, in the detection of the "Wrench", the model attains an impressive mAP of 71.78%.

Table 3. Comparison of lightweight indicators.

Backbone	Params	FLOPs	Model Size	FPS	mAP
MobileNetV1	4.23M	1.19G	16.14M	57.28	70.90%
MobileNetV2	3.50M	0.65G	13.37M	46.70	69.36%
MobileNetV3	5.48M	0.46G	20.92M	44.67	70.19%
GhostNet	5.18M	0.30G	19.77M	35.08	71.89%
DenseNet	7.98M	5.88G	30.44M	27.16	79.27%
CSPDarknet	26.6M	10.26G	101.54M	31.76	85.42%
MobileViT	5.64M	2.80G	21.50M	35.98	77.25%
MVray	5.64 M	2.72G	21.50 M	38.84	81.21%

Table 4. Ablation experiment results.

Backbone	AP					mAP
	Gun	Knife	Pliers	Scissor	Wrench	
MobileViT	95.30%	75.83%	77.79%	73.36%	63.98%	77.25%
MobileViT+DC	95.53%	78.65%	76.20%	77.99%	65.58%	78.79%
MobileViT+LGC	95.20%	78.48%	78.09%	71.08%	66.65%	77.90%
MVray	95.89%	80.01%	79.67%	78.70%	71.78%	81.21%

4.4 Visualization

The MVray model proposed herein is capable of identifying the type and location of prohibited items in X-ray security inspection images. This part showcases the MVray model's visual localization performance for detecting the prohibited items. As visualized in Fig. 3 (which can be zoomed in for more precise results), we can draw two primary conclusions. Firstly, compared to other algorithms, the MVray model yields more precise and compact bounding box annotations for prohibited items within X-ray images. Secondly, the MVray model exhibits

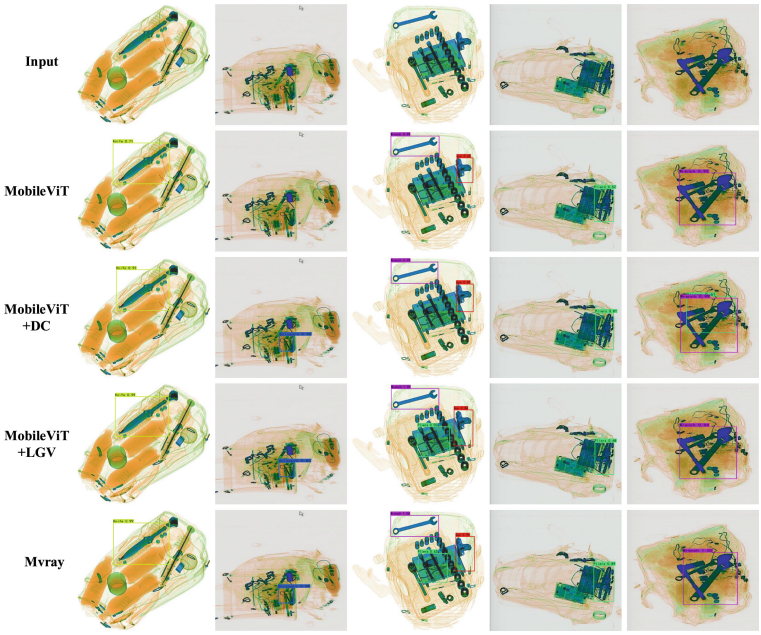

Fig. 3. Visualization results. Zoom in for better results.

a higher degree of confidence in discerning the types of prohibited items. In conclusion, the MVray model's performance in detecting prohibited items in X-ray security inspection images is commendably effective.

5 Conclusion

In this research, we devise a lightweight detection model, named MVray, targeting prohibited item detection in X-ray security inspection images. This model employs the MobileViT paired with a YOLO detection head as its backbone framework for feature extraction. To enhance the detection capabilities for smaller prohibited items, we implement a dense connection operation that effectively amalgamates high-level and low-level features, ensuring the resonance of small target features within high-level semantics. Further, we introduce a learnable group convolution to replace traditional convolution operations, mitigating feature redundancy triggered by dense connections, and thereby solidifying the lightweight characteristic of the model. With only 5.64M learning parameters and 2.72G FLOPs, the proposed MVray model achieves a mean average precision (mAP) of 81.21% on the SIXray dataset.

References

1. Mery, D.: X-ray testing by computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 360–367 (2013)
2. Liang, T., Lv, B., Zhang, N., Yuan, J., Zhang, Y., Gao, X.: Prohibited items detection in x-ray images based on attention mechanism. J. Phys. Conf. Ser. **1986**(1), 012087 (6pp) (2021)
3. An Chang, Yu., Zhang, S.Z., Zhong, L., Zhang, L.: Detecting prohibited objects with physical size constraint from cluttered x-ray baggage images. Knowl.-Based Syst. **237**, 107916 (2022)
4. Tao, R., et al.: Over-sampling de-occlusion attention network for prohibited items detection in noisy x-ray images. arXiv preprint arXiv:2103.00809 (2021)
5. Zhang, Y., Zhang, H., Zhao, T., Yang, J.: Automatic detection of prohibited items with small size in x-ray images. Optoelectron. Lett. **16**(4), 313–317 (2020)
6. Tao, R., et al.: Towards real-world x-ray security inspection: a high-quality benchmark and lateral inhibition module for prohibited items detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10923–10932 (2021)
7. Nguyen, H.D., Cai, R., Zhao, H., Kot, A.C., Wen, B.: Towards more efficient security inspection via deep learning: a task-driven x-ray image cropping scheme. Micromachines **13**(4), 565 (2022)
8. Howard, A., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314–1324 (2019)
9. Ren, Yu., Zhang, H., Sun, H., Ma, G., Ren, J., Yang, J.: LightRay: lightweight network for prohibited items detection in x-ray images during security inspection. Comput. Electr. Eng. **103**, 108283 (2022)
10. Ghosh, S., Srinivasa, S.K.K., Amon, P., Hutter, A., Kaup, A.: Deep network pruning for object detection. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 3915–3919. IEEE (2019)
11. Howard, A.G., et al.: MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
12. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: MobileNetV 2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
13. Zhang, X., Zhou, X., Lin, M., Sun, J.: ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856 (2018)
14. Ma, N., Zhang, X., Zheng, H.-T., Sun, J.: ShuffleNet V2: practical guidelines for efficient CNN architecture design. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 122–138. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_8
15. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
16. Mehta, S., Rastegari, M.: MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer (2022)
17. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
18. Miao, C., et al.: 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, pp. 2119–2128 (2019)

19. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: YOLOv4: optimal speed and accuracy of object detection (2020)
20. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, vol. 1, p. 2 (2019)
21. Han, K., et al.: A survey on vision transformer. IEEE Trans. Pattern Anal. Mach. Intell. **45**(1), 87–110 (2022)
22. Chu, X., et al.: Conditional positional encodings for vision transformers. arXiv preprint arXiv:2102.10882 (2021)
23. Wu, H., et al.: CVT: introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22–31 (2021)
24. Yuan, K., Guo, S., Liu, Z., Zhou, A., Yu, F., Wu, W.: Incorporating convolution designs into visual transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 579–588 (2021)
25. Mery, D., Svec, E., Arias, M., Riffo, V., Saavedra, J.M., Banerjee, S.: Modern computer vision techniques for x-ray testing in baggage inspection. IEEE Trans. Syst. Man Cybernet. Syst. **47**(4), 682–692 (2016)
26. Akcay, S., Breckon, T.: Towards automatic threat detection: a survey of advances of deep learning within x-ray security imaging. Pattern Recogn. **122**, 108245 (2022)
27. Aydin, I., Karakose, M., Akin, E.: A new approach for baggage inspection by using deep convolutional neural networks. In: 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), pp. 1–6. IEEE (2018)
28. Gaus, Y.F.A., Bhowmik, N., Akçay, S., Guillén-Garcia, P.M., Barker, J.W., Breckon, T.P.: Evaluation of a dual convolutional neural network architecture for object-wise anomaly detection in cluttered x-ray security imagery. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2019)
29. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
30. Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C.: GhostNet: more features from cheap operations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1580–1589 (2020)
31. Kim, J.-H., Kim, N., Park, Y.W., Won, C.S.: Object detection and classification based on yolo-v5 with improved maritime dataset. J. Marine Sci. Eng. **10**(3), 377 (2022)