Multi-label X-ray Imagery Classification via Bottom-up Attention and Meta Fusion

Benyi Hu[1], Chi Zhang[1], Le Wang[1], Qilin Zhang[2] and Yuehu Liu[1] *

[1] College of Artificial Intelligence, Xi'an Jiaotong University, Xi'an, Shaanxi, China
[2] HERE technology, Chicago, Illinois, USA

Abstract Automatic security inspection has received increasing interests in recent years. Due to the fixed top-down perspective of X-ray scanning of often tightly packed luggages, such images typically suffer from penetration-induced occlusions, severe object overlapping and violent changes in appearance. For this particular application, few research efforts have been made. To deal with the overlapping in X-ray images classification, we propose a novel Security X-ray Multi-label Classification Network (SXMNet). Our hypothesis is that different overlapping levels and scale variations are the primary challenges in the multi-label classification problem of prohibited items. To address these challenges, we propose to incorporate 1) **spatial attention** to locate prohibited items despite shape, color and texture variations; and 2) **anisotropic fusion** of per-stage predictions to dynamically fuse hierarchical visual information under violent variations. Motivated by these, our SXMNet is boosted by bottom-up attention and neural-guided Meta Fusion. Raw input image is exploited to generate high-quality attention masks in a bottom-up way for pyramid feature refinement. Subsequently, the per-stage predictions according to the refined features are automatically re-weighted and fused via a soft selection guided by neural knowledge. Comprehensive experiments on the Security Inspection X-ray (SIXray) and Occluded Prohibited Items X-ray (OPIXray) datasets demonstrate the superiority of the proposed method.

1 Introduction

With the increasing traffic in transportation hubs such as airports and high speed rail stations, security inspection procedures are becoming the bottleneck of throughput and causes of delays. However, such measures are indispensable due to security concerns [1, 2]. One primary time-consuming security inspection procedure involves X-ray scanning of passenger luggages, which generates pseudo-color images with respect to material properties via a dual energy X-ray scanning process [3]. In this scenario, objects are randomly stacked and heavily overlapped with each other, inducing heavy object occlusions. However, as demonstrated in Fig. 1, the appearance of X-ray imagery is different from regular

* Correspondance: liuyh@mail.xjtu.edu.cn

RGB images in its pseudo colors, textures and its unique partial transparency of materials. With such unique characteristics, dedicated machine learning algorithms and neural networks need to be crafted.

(a) Occlusion (b) Penetration

Figure 1. Illustration of exemplar penetration-induced object overlapping. Occluded items in (a), i.e. the gun (red ellipse) in the center and the knife (blue ellipse) in the left appears in (b) via X-ray penetration, with changes in appearance.

Existing security X-ray scanning requires a trained professional to continuously monitor such images. Such a process is tiresome and difficult to scale up. Therefore, there is an urgent need to improve the automation and accuracy of recognizing prohibited items within X-ray images. Recently, with the prosperity of applying deep learning methods [4], especially the convolutional neural networks, the recognition of prohibited items in X-ray pictures can be regarded as a multi-label classification problem [5].

To recognize prohibited items under penetration-induced overlapping, we assume that low-level information especially colors and edges is the key to distinguish objects in complex and cluttered backgrounds. Moreover, such X-ray images are often consist of objects of dramatically different sizes and existing methods utilize Feature Pyramid Network (FPN) architecture [6] to alleviate scale variation problem, which leverages ConvNets feature hierarchy to build pyramid features. Afterwards, per-level predictions, i.e. outputs made according to features of each level separately, are combined as final classification output through an average fusion. However, these methods are shown in our paper to be suboptimal, possibly due to their application of static fusion, which we speculate is insufficient to account for the challenging penetration-induced object overlapping [7,8] and violent appearance changes. Since such object overlapping phenomena varies universally and diversely among X-ray images and within the same image, it is reasonable to hypothesize that ***features with different spatial attention scores and from different levels contribute unequally in the classification***, which implies the needs for spatial attention and dynamic late fusion.

Based on these assumptions, we propose a framework termed as SXMNet, to solve the challenging penetration-induced overlapping in the X-ray multi-label classification problem. Instead of directly extracting deep features, we construct

a bottom-up spatial attention head, which utilizes the hierarchical character of FPN, to select the foreground region from the complex and cluttered backgrounds. Since larger attention masks mean better locating capabilities [9,10], we build bottom-up pathway to expand the resolution of attention masks, concatenating feature maps with the raw input image in several stages to exploit low-level visual information including edges and colors. Furthermore, neural-guided Meta Fusion is proposed to automatically assign soft weights to the pyramid predictions according to the neural knowledge rather than ad-hoc heuristics like scales [11] or gated fusion [12]. Further experiments show that our hypothesis is supported as our method outperforms the baseline by a large margin in the multi-label classification task on X-ray images.

In summary, our major contributions are three-fold:

1) For locating prohibited items in the penetration-induced overlapping scenarios, we present the bottom-up attention mechanism, which utilizes the raw input image in each stage to infuse low-level visual information such as colors, textures and edge information.

2) We propose a plug-in Meta Fusion module to address the scale variation problem, which can learn from other neural networks and re-weight the multi-stage predictions in a dynamic style.

3) To further evaluate the effects and transferability of the proposed Meta Fusion mechanism, we implement it with several architectures and datasets. Comprehensive experiments on the X-ray datasets prove that our approach achieves superior performance over previous ones. Moreover, we validate that neural knowledge is capable of better generalization performance through neural-guided Meta Fusion.

2 Related Work

2.1 Object Recognition within X-ray Images

Early work with X-ray security imagery primarily utilizes hand-crafted features like Bag-of-Visual-Words (BoVW) together with Supported Vector Machine (SVM) for classification [13, 14]. Recently, object recognition within X-ray images has also witnessed the prominence of powerful features of convolutional neural networks (CNNs) and [15] first introduced the use of CNN to address object classification task by comparing varying CNN architectures to the earlier work extensive BoVW [14]. In the scenario that each image may contain more than one prohibited items, there are two typical types of object recognition methods. The first one worked on the instance level, providing bounding box as well as predicted label for each instance individually [16–19]. The other instead worked on image level which produces a score for each class indicating its presence or absence [20]. [21] augments input images based on generative adversarial networks to improve classification performance. [5] utilizes hierarchical refinement structure to deal with overlapping problem within X-ray images and alleviates data imbalance with a well-designed loss. This paper mainly studies the image-level recognition method.

2.2 Attention Mechanism

Motivated by the human perception process using top information to guide bottom-up feedforward process, attention mechanism has been widely applied to many computer vision tasks such image classification [22, 23], scene segmentation [24, 25] and visual question answering [26]. Squeeze-and-Excitation Networks (SENet) [27] adaptively re-calibrates channel-wise feature responses by explicitly modeling inter-dependencies between channels. Convolutional Block Attention Module (CBAM) [28] sequentially infers attention maps along both channel-wise and spatial-wise. Recently, attention mechanisms have been introduced to deep neural networks for multi-label image classification. It aims to explicitly or implicitly extract multiple visual representations from a single image characterizing different associated labels [29–32]. In the context of object recognition in X-ray images, researchers realized that these images often contain fewer texture information, yet shape information stands out to be more discriminative [5]. Therefore, we design a bottom-up attention to utilize the low-level visual information such as color and texture.

2.3 Fusion Mechanism

Fusion strategy includes early fusion (concatenate multiple features) and late fusion (fuse predictions of different models). [33] adopts gated fusion strategy which derives three inputs from an original hazy image to yield the final dehazed image. [34] utilizes a hard constraint on the matrix rank to preserve the consistency of predictions by various features. The advantage of combining multi-scale input images for multi-label image classification by employing varying fusion approaches has been proved in [35, 36]. In this filed, almost all the methods mentioned above are based on intuition or algebraic knowledge, which may not be the optimal fusion strategy for CNNs. Different from them, our Meta Fusion utilizes neural knowledge to serve as the guidance of the fusion process, supervising the learning of network in a neural-guided manner.

3 Proposed Approach

3.1 Formulation

For object recognition in X-ray images, the primary characteristic is that the possible number and category of prohibited items appearing in the image is uncertain. Therefore, we formulate it as a multi-label classification problem. Suppose there are C classes of possible items in the dataset. For each given image $\mathbf{x}$, our goal is to obtain a C-dimensional vector $\mathbf{y}$ for each $\mathbf{x}$, each dimension $\mathbf{y}_c$ for $c \in \{0, 1, ..., C\}$, is either 0 or 1, with 1 indicating the specified prohibited item is present in this image and 0 otherwise.

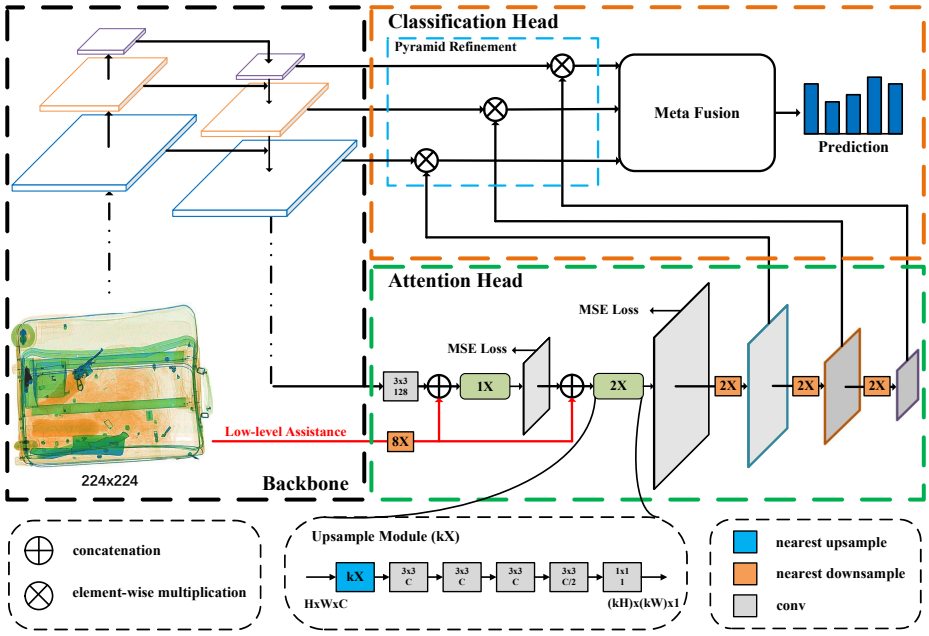


Figure 2. Overview architecture. Backbone extracts pyramid features shared between attention head and classification head. Attention head effectively outputs attention masks which in return improve the performance of classification head.

3.2 Security X-ray Multi-label Classification Network

As shown in Fig. 2, our proposed Security X-ray Multi-label Classification Network (SXMNet) consists of three components: backbone, attention head and multi-label classification head. Backbone outputs shared features among two heads, which is implemented using ResNet50 with the Feature Pyramid Network (FPN) architecture [6]. Attention head cooperates both shared features and raw input image to generate attention masks, which serves as a feature selector to distinguish prohibited items during penetration-induced scenarios. Based on the attention masks, multi-label classification head refines pyramid features and computes final predictions with neural-guided Meta Fusion mechanism.

Pyramid Refinement with Bottom-up Attention The goal is to predict the category of prohibited items in X-ray security imagery. However, it is difficult to capture distinctive visual features within complex and cluttered imagery due to penetration. To solve this problem, we introduce bottom-up attention guidance to select reliable deep descriptors. Specifically, based on the output pyramid feature $\mathbf{p}^l$, where l indicates the pyramid level, attention head predicts the location of all prohibited items using the predicted heat map. From the last feature maps of the backbone, the attention head is constructed by stacking Upsample Modules. Similar to [37], each Upsample Module consists of a bilinear

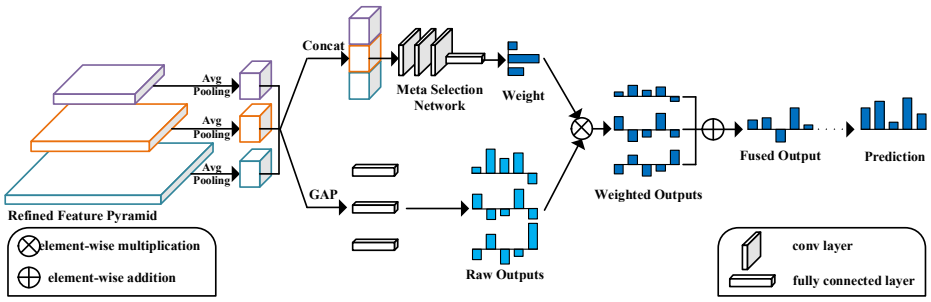


Figure 3. Pipeline of Meta Fusion. meta fusion consists of two branches: predict-branch outputs predictions corresponding to each pyramid feature, and soft-weight-branch output weights indicating the importance of each pyramid feature. GAP means global average pooling.

upsample layer to expand spatial sizes, several dilated convolutional layers to extract deeper features without losing resolutions and a 1x1 convolutional layer as output layer. Notably, the appearance features from raw input image could provide abundant information to support the localization in penetration-induced overlapping scenes, so we concat the feature and the rescale input image together before each upsampling. Based on the outputs of the attention head, multi-scale feature maps are refined in a pyramid refinement way like [38], where pyramid features are multiplied with its corresponding attention mask simultaneously.

Classification with Meta Fusion In feature pyramid based classification networks, features of each level, namely $\mathbf{p}^l$, make their own prediction $\mathbf{y}^l$ independently. Traditionally people merge them into the final prediction by simply average them, as denoted in Eq. (1):

$$\mathbf{y} = \frac{1}{n} \sum_{t=l_{min}}^{l_{max}} \mathbf{y}^t, \tag{1}$$

where l_{min} denotes the minimal pyramid level, l_{max} denotes the maximum and $n = l_{max} - l_{min} + 1$.

However, the contributions from different level features should be different [39], weighting the pyramid predictions can produce performance gains. Especially in X-ray images, late fusion of predictions is essential to address penetration problem as we assumed. In addition, CNN model trained on a given X-ray security image dataset will have a higher degree of generalization if applied to other X-ray datasets [40]. Therefore, we propose to insert a meta-selection network [41] into the architecture, which utilizes the neural knowledge to predict weights for soft prediction selection.

As shown in Fig. 3, we pool each pyramid feature so that each having spatial size of 7x7, which is fed to the meta-selection network after concatenation, outputting a vector of the probability distribution as the weights of soft prediction

selection. Thus, Eq. (1) is augmented into Eq. (2):

$$\mathbf{y} = \sum_{t=l_{min}}^{l_{max}} \mathbf{w}^t \mathbf{y}^t, \tag{2}$$

where $\mathbf{w}^t$ is the output of the meta-selection network.

Details of the meta-selection network is shown in Tab. 1.

Table 1. The architecture of the meta-selection network used in the proposed method.

Layer	Description	activation
input	Refined feature pyramid	None
L-0	Conv, 768x3x3 $\rightarrow$ 256	ReLU
L-1	Conv, 256x3x3 $\rightarrow$ 256	ReLU
L-2	Conv, 256x3x3 $\rightarrow$ 256	ReLU
L-3	Fully connected, 256 $\rightarrow$ 3	Softmax

Loss Function The loss function of the proposed network consists of the attention term, meta selection term and multi-label classification term. For attention term, we utilize Mean Squared Error (MSE) to measure the difference between the ground-truth map and the estimated heatmap that the attention head predicted. For the meta selection term, we use the standard Cross-Entropy (CE). In addition, Binary Cross-Entropy (BCE) loss is utilized to optimize the multi-label classification task.

During the stage of training attention head (**stage I**), the loss is:

$$L_I = L_{att}, \tag{3}$$

where L_{att} is the MSE loss of generating attention mask.

During the stage of training classification head (**stage II**), the total classification loss is given as follow:

$$L_{II} = L_{cls} + \lambda L_{meta}, \tag{4}$$

where L_{cls} is the loss of the multi-label classification task, λ is the hyperparameter that controls the contribution of meta-selection loss, L_{meta}.

Ground-truth Generation To train the SXMNet, we generate both ground-truth heatmaps and ground-truth meta selection labels. For the heatmaps, we generate an inline ellipse of each annotated bounding box, where pixels inside the ellipse is set to 1 and others are 0. Details of the generated ground-truth are shown in Fig. 5. As for meta selection, first each image is forwarded through all levels of feature pyramid by CHR [5]. The ground-truth is a one-hot vector, with 1 indicating this pyramid level yielding the minimal classification loss and 0 otherwise.

Training Strategy Different from self attention mechanism, the training scheme of our model consists of two stages. In the **stage I**, we only train the attention task with instance-level annotated data, which provides bounding box for each prohibited item. So only parameters of backbone and the attention head are updated. After **stage I** training is complete, we continue to train it on a large number of image-level annotated data. The predicted attention maps, along with pyramid feature representations of the input image, can be further utilized to improve the multi-label classification performance.

4 Experiments

4.1 Datasets

To evaluate the effectiveness of our proposed methods, we conduct experiments on two benchmark security inspection datasets, *i.e.*, SIXray [5] and OPIXray [42].

SIXray consists of 1,059,231 X-ray images, in which six classes of 8,929 prohibited items. These images are collected using a Nuctech dual-energy X-ray scanner, where the distribution of the general baggage/parcel items corresponds to stream-of-commerce occurrence. In our experiments, we only use the images containing prohibited items and follow the same division protocol as [5], namely 7496 images for training and 1433 images for testing.

Compared with SIXray, OPIXray is more challenging with lower resolution of prohibited items and higher inter-class similarity of them. It contains a total of 8885 X-ray images of 5 categories of cutters (e.g., Folding Knife, Straight Knife, Scissor, Utility Knife, Multi-tool Knife). The dataset is partitioned into a training set and a testing set, with the former containing 80% of the images (7109) and the latter containing 20% (1776). Since these images contain a lot of pure white backgrounds, we crop each image before training.

4.2 Implementation Details

Network Details: In network implementation, $l \in \{3, 4, 5\}$ and $\lambda = 0.1$.
Training Details: We trained our network with SGD optimizer [43]. For **stage I**, total epochs are 350 and the initial learning rate is 1e-5, which is divided by 10 after every 100 epochs. For **stage II**, total epochs are 150 and the initial learning rate is 1e-1 for fully connected layers and 2e-2 for others, which is divided by 10 after every 30 epoch. The batch size is 64 for all training stages.

4.3 Comparison Methods

Baseline methods: We employ plaining training with ResNet [44] and ResNet-FPN [6] as our baselines.
State-of-the-art methods: For state-of-the-art methods, we choose the recently proposed CHR [5] which achieves good multi-label classification accuracy on SIXray datasets.

Table 2. Multi-label classification performance (AP, %).

Dataset	Category	ResNet50 [44]	ResNet50-FPN [6]	CHR [5]	Ours
SIXray	Gun	98.70	98.23	98.29	**98.84**
	Knife	92.59	93.45	94.87	**95.22**
	Pliers	96.41	96.66	96.40	**98.33**
	Scissors	91.09	92.30	91.75	**96.11**
	Wrench	86.74	88.66	88.51	**95.38**
	mean	93.36	93.86	93.96	**96.78**
OPIXray	Folding	92.93	93.92	94.62	**96.11**
	Straight	65.40	64.76	67.42	**75.37**
	Scissor	99.05	99.18	98.93	**99.34**
	Utility	78.25	78.83	80.39	**84.32**
	Multi-tool	96.15	96.20	97.28	**97.69**
	mean	86.35	86.58	87.73	**90.83**

4.4 Overall Performance

Experimental Results on SIXray Tab 2 shows the results on the SIXray datasets. Since CHR only gives the results on both positive and negative images, we re-train it on the only positive images with the source codes provided by the authors [5]. As shown in the table, we demonstrate that our method achieves the best results across all categories, when comparing the baseline methods and previous state-of-the art. Our method achieves larger performance enhancement, i.e. **2.82%** in terms of mean AP.

Experimental Results on OPIXray Similar to SIXray, our performance still surpasses other method by a large margin in terms of mAP. We achieve **3.10%** improvements on OPIXray, from Tab 2.

4.5 Ablation Studies

Effects of Each Component To validate the effects of each component, we conduct ablations on the attention mechanism and meta fusion mechanism. As show in Tab. 3, the attention mechanism brings the largest improvements, i.e. **2.05%** in terms of mean AP, indicating that spatial attention plays an important role in handling heavily clusttered in X-ray images. The power of effectively localization in return verifies our assumption: effective spatial attention mechanism could improve recognition performance in penetration-induced over-lapping scenarios. As we assumed, different levels of penetration require anisotropic fusion of predictions and our meta fusion is designed to soft-weight predictions during training. As shown in Tab. 3, Meta Fusion can result in a slight improvements compared with our attention mechanism, i.e. **0.4%**. With both attention mechanism and meta fusion mechanism, we achieve the best performance on SIXray, showing that spatial attention and anisotropic fusion are jointly contributing to the multi-label classification task within X-ray images.

Table 3. Ablation Results of Each Component (AP, %).

	Backbone	FPN	Attention	Meta Fusion	mAP
Baseline	ResNet-50				93.36
Ours	ResNet-50	√			93.86
		√	√		95.91
		√		√	94.26
		√	√	√	**96.78**

Table 4. Ablation studies for the effects and transferability of Meta Fusion.

Dataset	SIXray	OPIXray
Gated Fusion [12]	96.15	90.67
Our MF-I	96.33	90.36
Our MF-N	**96.78**	**90.83**

MF-I denotes meta fusion with label provided by intuition. MF-N denotes meta fusion where label is provided by CHR [5].

Effects and Transferability of Meta Fusion For better understanding of our proposed Meta Fusion, we conduct experiments on different fusion strategies. We compare three fusion methods including gated fusion, Meta Fusion by Intuition (MF-I) and Meta Fusion by Neural (MF-N). MF-I denotes using intuitive label according to the proportion of the smallest bounding box of prohibited items. Specifically, for proportion under 2%, we set p_3 level as the ground truth, between 2% and 20% are set to p_4 level, and others are p_5 level. MF-N means using the label generated by CHR [5], setting the pyramid level with smallest classification loss as the ground-truth. As shown in Tab. 3, meta fusion achieves better performance compared with gated fusion. In addition, optimization with label provided by neural knowledge (see +MF-N) allows better utilization of pyramid features than label given by intuition (see +MF-I).

Furthermore, to evaluate the transferability of Meta Fusion, we train our network on OPIXray with label generated by CHR, which is trained only on SIXray. The categories of prohibited items in SIXray is different from those in OPIXray. As shown in the right column of Tab. 3, Meta Fusion still performs better than naive gated fusion, proving the neural knowledge will be capable of better generalization performance through Meta Fusion. Note that the performances of MF-I on OPIXray is slightly reduced. We speculate that the reason is that prohibited items are much smaller in size than those in SIXray, so the prior knowledge mismatches OPIXray situations.

To facilitate the understanding of our proposed Meta Fusion mechanism, we visualize the distribution of the Meta Fusion label during training and fusion weights during inference, as indicated by the histogram in Fig. 4. The left panel of Fig. 4 shows a histogram of labels provided by prior Knowledge (MF-I) and provided by minimal classification loss according to CHR (MF-N). The figure suggests that the distribution of labels provided by intuition and CHR differs a lot, we conjecture that the proportion of smallest prohibited item always be in

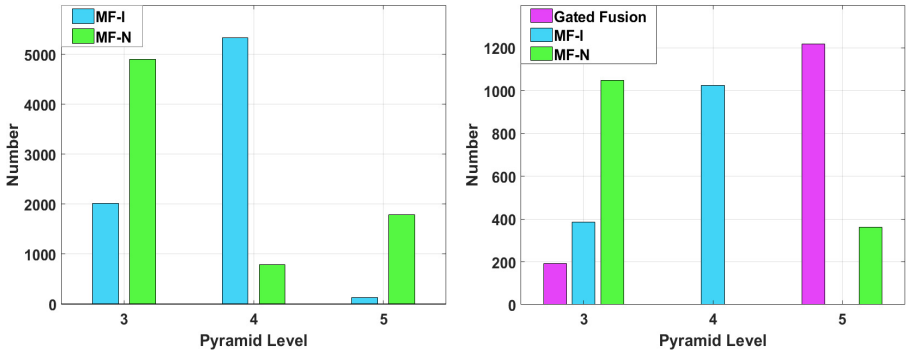


Figure 4. Histogram of the selected pyramid level on SIXray. (left) shows the histogram of training label provided by prior Knowledge (MF-I) and label provided by minimal classification loss according to CHR (MF-N); (right) shows the histogram of the pyramid level with the max weight predicted by gated fusion, MF-I, and MF-N during inference.

a moderate level, so most of the intuitive labels indicate the p_4 as the optimal level, while CHR labels prefer to p_3.

The right panel of the of Fig. 4 shows the distribution of the fusion weights for the test set during inference. We just plot the pyramid layers corresponding to the maximum selection weight for simplicity. As the figure shows, the predicted fusion weights are affected by its training label since the histograms are similar between them. What is more, whether the neural knowledge the network utilizes comes from itself (Gated Fusion) or other networks (MF-N), they perform in a similar manner: both concentrating on the most top (p_5) or most down (p_3) features, neglecting the middle pyramid feature. This phenomenon may suggest that the p_3 and p_5 have the most distinct representations in FPN architecture.

Table 5. Validation of training strategy(AP, %). Our strategy(2nd row) is compared with self attention mechanism(1st row).

Strategy	SIXray (only classification)	SIXray (attention & classification)	mean
Self	√		94.11
Ours		√	**96.78**

Effects of Training Bottom-up Attention To determine the cause of the performance improvement an increase parameters caused by the attention head or the selection capability of the attention head, we conduct experiments on different ways of training the attention head. For fair comparison, we keep the

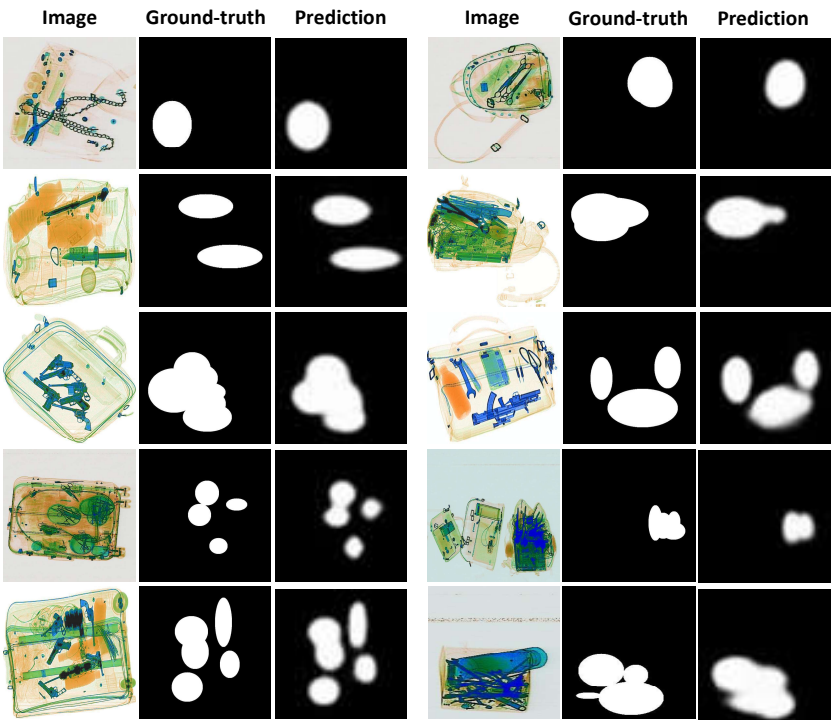


Figure 5. Qualitative results of attention masks on test set. The results show that our low-level assistance attention is effective in complex and cluttered background due to penetration-induced overlapping problem. (Best viewed in color and zoomed in.)

architecture of the network unchanged and only change task used in **stage I**. As shown in Tab. 5, Self means self attention, which skips the **stage I** and initializes the network with parameters trained on ImageNet [45], and Ours indicates our training strategy. Experimental results demonstrate that our training strategy outperforms the self-attention strategy, indicating the capability of locating prohibited items can be beneficial.

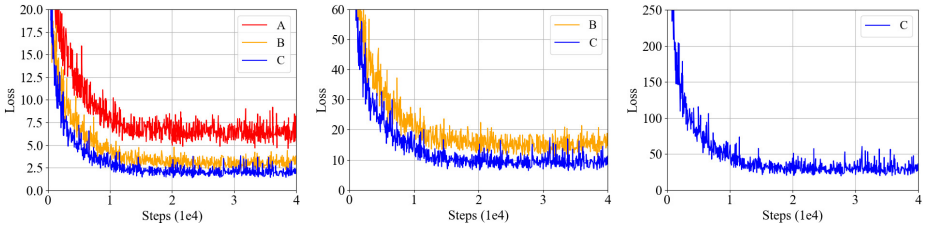
Fig. 5 shows some qualitative results of predicted attention masks on the test set, indicating that our low-level assistance help to discover prohibited items under complex and cluttered backgrounds.

Table 6. Quantitative results of different output attention resolution.

Resolution	SIXray
28	96.12
56	**96.78**
112	95.47

Table 7. Results of whether concatenating the raw input image while constructing attention head.

	Raw Input Image?	SIXray
Attention		95.72
Attention	√	**96.78**



(a) Resolution-28 (b) Resolution-56 (c) Resolution-128

Figure 6. Attention training loss of each resolution on SIXray. We construct attention head with different outputting resolution masks based on none upsample module (A), 1 upsample modules (B) and 2 upsample modules (C), respectively. It is obviously that network has lower training loss in the former stage with larger masks follows afterwards.

Effects of Resolution of Attention Mask To demonstrate the impact of the resolution of attention mask, we conduct several experiments with different resolutions of attention masks. As shown in Fig. 6, with more upsampling modules, namely higher outputting attention resolution, the loss of each stage is much smaller, showing large attention mask help precisely locate prohibited items. Based on **stage I** model, we continue training the multi-label classification task. From Tab. 6, large resolution with better locating ability could further boost the multi-label classification performance. For the case with resolution 112, we observed a significant drop in performance and our preliminary assessment is due to overfitting, possibly due to limited granularity in ground-truth generation.

Effects of Low-level Assistance To validate the effects of low-level assistance while building the attention head, we conduct experiments on whether adding the raw input image to construct the attention head. As Tab. 7 shows, concatenating raw input image as the low-level assistance can bring **1.06%** improvements, indicating that low-level visual information, such as edges and colors, is important in locating objects under heavily cluttered backgrounds.

Table 8. Performance comparison between Meta Fusion-integrated (MF-I, MF-N) network and baselines for two multi-label classification approaches.

Method	SIXray	OPIXray
Res50-FPN [6]	93.86	86.58
Res50-FPN + MF-I	**94.65**	87.33
Res50-FPN + MF-N	94.24	**87.64**
CHR [5]	93.96	87.33
CHR + MF-I	93.87	**88.29**
CHR + MF-N	93.77	87.84

4.6 Comparing with Different Baselines

To further evaluate the effectiveness of Meta Fusion and verify Meta Fusion can be applied to various networks in the prediction fusion stage, we conduct experiments on two approaches with multi-prediction architecture, i.e., Res50-FPN [6] and CHR [5]. The results are shown in Tab. 8.

As we can see from Tab. 8, the performance of Meta Fusion-integrated networks are improved by **1.06%** and **0.96%** compared with Res50-FPN, and CHR respectively on OPIXray. Meanwhile, **0.79%** improvement is gained compared with Rer50-FPN on SIXray, which indicates that our module can be inserted as a plug-and-play module into networks with predictions fusion stage and receive a better performance. Note that the performance of MF-integrated CHR on SIXray is sightly below the baseline, and we speculate that the labels are given by CHR trained on SIXray itself, which limits further performance boost.

5 CONCLUSION

In this paper, we investigate the prohibited items discovery problem in X-ray scanning images. We propose a novel SXMNet to deal with the penetration-induced overlapping with both spatial attention and dynamic fusion. For selecting reliable foregrounds, the raw input image is utilized as low-level assistance to construct attention head in bottom-up pathway. Subsequently, per-stage predictions of refined pyramid features are fused adaptively in a weighted manner, where the weights are dynamically predicted by the proposed neural-guided Meta Fusion scheme. Experimental results demonstrate that presented framework outperforms the baselines and previous state-of-art by a large margin.

Acknowledgement. This work was supported by the National Key R&D Program of China (Grant 2018AAA0102504) and the National Natural Science Foundation of China (Grant 61973245). We thank Xiaojun Lv of Institute of Computing Technologies China Academy of Railway Sciences Corporation Limited for his great contribution to this work.

References

1. 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 Transactions on Systems, Man, and Cybernetics: Systems **47** (2016) 682–692
2. Akcay, S., Breckon, T.P.: Towards automatic threat detection: A survey of advances of deep learning within x-ray security imaging. CoRR **abs/2001.01293** (2020)
3. Mouton, A., Breckon, T.P.: A review of automated image understanding within 3d baggage computed tomography security screening. Journal of X Ray Science and Technology **23** (2015) 531–555
4. LeCun, Y., Bengio, Y., Hinton, G.E.: Deep learning. Nature **521** (2015) 436–444

5. Miao, C., Xie, L., Wan, F., Su, C., Liu, H., Jiao, J., Ye, Q.: Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 2119–2128
6. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 2117–2125
7. Akcay, S., Breckon, T.P.: An evaluation of region based object detection strategies within x-ray baggage security imagery. In: 2017 IEEE International Conference on Image Processing (ICIP), IEEE (2017) 1337–1341
8. Mery, D., Arteta, C.: Automatic defect recognition in x-ray testing using computer vision. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE (2017) 1026–1035
9. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. (2016) 4724–4732
10. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 5693–5703
11. Wang, J., Chen, K., Yang, S., Loy, C.C., Lin, D.: Region proposal by guided anchoring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 2965–2974
12. Cheng, Y., Rui, C., Li, Z., Xin, Z., Huang, K.: Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
13. Mery, D., Svec, E., Arias, M.: Object recognition in baggage inspection using adaptive sparse representations of x-ray images. In: Image and Video Technology - 7th Pacific-Rim Symposium, PSIVT 2015, Auckland, New Zealand, November 25-27, 2015, Revised Selected Papers. Volume 9431 of Lecture Notes in Computer Science., Springer (2015) 709–720
14. Kundegorski, M.E., Akcay, S., Devereux, M., Mouton, A., Breckon, T.P.: On using feature descriptors as visual words for object detection within x-ray baggage security screening. In: 7th International Conference on Imaging for Crime Detection and Prevention, ICDP 2016, Madrid, Spain, November 23-25, 2016, IET / IEEE (2016) 1–6
15. Akcay, S., Kundegorski, M.E., Devereux, M., Breckon, T.P.: Transfer learning using convolutional neural networks for object classification within x-ray baggage security imagery. In: 2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016, IEEE (2016) 1057–1061
16. Steitz, J.O., Saeedan, F., Roth, S.: Multi-view x-ray R-CNN. In: Pattern Recognition - 40th German Conference, GCPR 2018, Stuttgart, Germany, October 9-12, 2018, Proceedings. Volume 11269 of Lecture Notes in Computer Science., Springer (2018) 153–168
17. Akcay, S., Kundegorski, M.E., Willcocks, C.G., Breckon, T.P.: 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** (2018) 2203–2215
18. Bastan, M.: Multi-view object detection in dual-energy x-ray images. Mach. Vis. Appl. **26** (2015) 1045–1060

19. Cao, S., Liu, Y., Song, W., Cui, Z., Lv, X., Wan, J.: Toward human-in-the-loop prohibited item detection in x-ray baggage images. In: 2019 Chinese Automation Congress (CAC). (2019) 4360–4364
20. Wang, Q., Jia, N., Breckon, T.P.: A baseline for multi-label image classification using an ensemble of deep convolutional neural networks. In: 2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, September 22-25, 2019, IEEE (2019) 644–648
21. Yang, J., Zhao, Z., Zhang, H., Shi, Y.: Data augmentation for x-ray prohibited item images using generative adversarial networks. IEEE Access **7** (2019) 28894–28902
22. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society (2017) 6450–6458
23. Chen, X., Wu, J., Lin, L., Liang, D., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y., Tong, R.: A dual-attention dilated residual network for liver lesion classification and localization on CT images. In: 2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, September 22-25, 2019, IEEE (2019) 235–239
24. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE (2019) 3146–3154
25. Li, K., Wu, Z., Peng, K., Ernst, J., Fu, Y.: Tell me where to look: Guided attention inference network. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, IEEE Computer Society (2018) 9215–9223
26. Peng, L., Yang, Y., Wang, Z., Wu, X., Huang, Z.: Cra-net: Composed relation attention network for visual question answering. In: Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019, ACM (2019) 1202–1210
27. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, IEEE Computer Society (2018) 7132–7141
28. Woo, S., Park, J., Lee, J., Kweon, I.S.: CBAM: convolutional block attention module. In: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII. Volume 11211., Springer (2018) 3–19
29. Wang, Z., Chen, T., Li, G., Xu, R., Lin, L.: Multi-label image recognition by recurrently discovering attentional regions. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society (2017) 464–472
30. Zhu, F., Li, H., Ouyang, W., Yu, N., Wang, X.: Learning spatial regularization with image-level supervisions for multi-label image classification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society (2017) 2027–2036
31. You, R., Guo, Z., Cui, L., Long, X., Bao, Y., Wen, S.: Cross-modality attention with semantic graph embedding for multi-label classification. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press (2020) 12709–12716

32. Guo, H., Zheng, K., Fan, X., Yu, H., Wang, S.: Visual attention consistency under image transforms for multi-label image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE (2019) 729–739
33. Ren, W., Ma, L., Zhang, J., Pan, J., Cao, X., Liu, W., Yang, M.: Gated fusion network for single image dehazing. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, IEEE Computer Society (2018) 3253–3261
34. Dong, X., Yan, Y., Tan, M., Yang, Y., Tsang, I.W.: Late fusion via subspace search with consistency preservation. IEEE Trans. Image Process. **28** (2019) 518–528
35. Wang, M., Luo, C., Hong, R., Tang, J., Feng, J.: Beyond object proposals: Random crop pooling for multi-label image recognition. IEEE Trans. Image Process. **25** (2016) 5678–5688
36. Durand, T., Mordan, T., Thome, N., Cord, M.: WILDCAT: weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society (2017) 5957–5966
37. Li, C., Du, D., Zhang, L., Luo, T., Wu, Y., Tian, Q., Wen, L., Lyu, S.: Data priming network for automatic check-out. In: Proceedings of the 27th ACM International Conference on Multimedia. (2019) 2152–2160
38. Zhong, Z., Zhang, C., Liu, Y., Wu, Y.: Viaseg: Visual information assisted lightweight point cloud segmentation. In: 2019 IEEE International Conference on Image Processing (ICIP), IEEE (2019) 1500–1504
39. Zhu, C., He, Y., Savvides, M.: Feature selective anchor-free module for single-shot object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 840–849
40. Gaus, Y.F.A., Bhowmik, N., Akcay, S., Breckon, T.P.: Evaluating the transferability and adversarial discrimination of convolutional neural networks for threat object detection and classification within x-ray security imagery. In Wani, M.A., Khoshgoftaar, T.M., Wang, D., Wang, H., Seliya, N., eds.: 18th IEEE International Conference On Machine Learning And Applications, ICMLA 2019, Boca Raton, FL, USA, December 16-19, 2019, IEEE (2019) 420–425
41. Zhu, C., Chen, F., Shen, Z., Savvides, M.: Soft anchor-point object detection. arXiv preprint arXiv:1911.12448 (2019)
42. Wei, Y., Tao, R., Wu, Z., Ma, Y., Zhang, L., Liu, X.: Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module. CoRR **abs/2004.08656** (2020)
43. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010. Springer (2010) 177–186
44. 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. (2016) 770–778
45. Deng, J., Dong, W., Socher, R., Li, L.J., Li, F.F.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA. (2009)