Occluded Prohibited Items Detection: An X-ray Security Inspection Benchmark and De-occlusion Attention Module

Yanlu Wei[1*], Renshuai Tao[1*], Zhangjie Wu[1], Yuqing Ma[1], Libo Zhang[2], Xianglong Liu[1,3†]
[1] State Key Lab of Software Development Environment, Beihang University
[2] Institute of Software Chinese Academy of Sciences
[3] Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University
{weiyanlu, rstao, zhangjiewu, mayuqing, xlliu}@buaa.edu.cn, libo@iscas.ac.cn

ABSTRACT

Security inspection often deals with a piece of baggage or suitcase where objects are heavily overlapped with each other, resulting in an unsatisfactory performance for prohibited items detection in X-ray images. In the literature, there have been rare studies and datasets touching this important topic. In this work, we contribute the first high-quality object detection dataset for security inspection, named Occluded Prohibited Items X-ray (OPIXray) image benchmark. OPIXray focused on the widely-occurred prohibited item "cutter", annotated manually by professional inspectors from the international airport. The test set is further divided into three occlusion levels to better understand the performance of detectors. Furthermore, to deal with the occlusion in X-ray images detection, we propose the De-occlusion Attention Module (DOAM), a plug-and-play module that can be easily inserted into and thus promote most popular detectors. Despite the heavy occlusion in X-ray imaging, shape appearance of objects can be preserved well, and meanwhile different materials visually appear with different colors and textures. Motivated by these observations, our DOAM simultaneously leverages the different appearance information of the prohibited item to generate the attention map, which helps refine feature maps for the general detectors. We comprehensively evaluate our module on the OPIXray dataset, and demonstrate that our module can consistently improve the performance of the state-of-the-art detection methods such as SSD, FCOS, etc, and significantly outperforms several widely-used attention mechanisms. In particular, the advantages of DOAM are more significant in the scenarios with higher levels of occlusion, which demonstrates its potential application in real-world inspections. The OPIXray benchmark and our model are released at https://github.com/OPIXray-author/OPIXray.

CCS CONCEPTS

• **Computing methodologies** → **Object detection**.

KEYWORDS

object detection; security inspection; X-ray images; occlusion

ACM Reference Format:
Yanlu Wei, Renshuai Tao, Zhangjie Wu, Yuqing Ma, Libo Zhang, Xianglong Liu. 2020. Occluded Prohibited Items Detection: An X-ray Security Inspection Benchmark and De-occlusion Attention Module. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413828

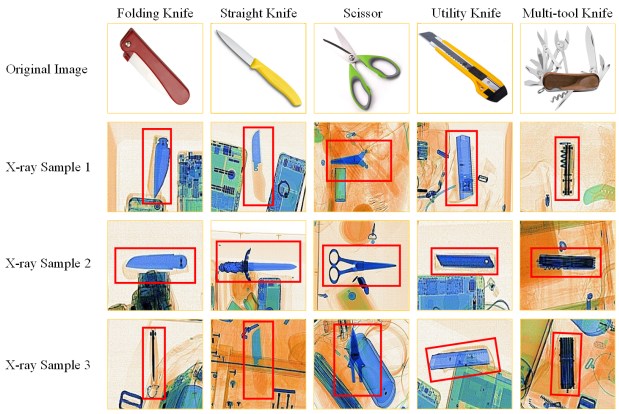


Figure 1: Samples of the five categories of cutters and corresponding X-ray images.

1 INTRODUCTION

With the increasing crowd density in public transportation hubs, security inspection has become more and more important in protecting public safety. Security inspection usually adopts X-ray scanners to find whether there is any prohibited item in passenger luggage. In this scenario, objects in the luggage are randomly stacked and heavily overlapped with each other, leading to heavy object occlusion. As a result, after a long time localizing prohibited items in large amounts of complex X-ray images without distraction, security inspectors struggle to accurately detect all the prohibited items, which may cause severe danger to the public. And changing shifts frequently will cost a large number of human resources, which is not advisable.

*indicates equal contribution.
†Corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10.
https://doi.org/10.1145/3394171.3413828

Therefore, a rapid, accurate and automatic approach to assist inspectors to detect prohibited items in X-ray scanned images is desired eagerly. As the technology of deep learning [9] especially the convolutional neural network develops [3, 22], the recognition of occluded prohibited items from X-ray pictures can be regarded as an object detection problem of computer vision, which has been widely studied in the literature.

There are several works trying to solve occlusion problems in different scenarios, such as person re-identification [24, 29, 30], face recognition [5, 17, 25]. The object occlusion in Person Re-identification or Face Recognition belongs to intra-class occlusion, and every occluded object has a corresponding annotation. Therefore, a loss function can be designed by annotation information to decrease the impact of occlusion. However, object occlusion in X-ray images for security inspection often exists between prohibited items and safety items, which belongs to inter-class occlusion. For the prohibited items detection task, what we obtained are the annotations of prohibited items, so methods of these senses can not be used for comparison. To the best of our knowledge, up to now, no dataset targeting occluded prohibited items detection in X-ray images has been proposed by researchers even there are also two released X-ray benchmark, namely GDXray [12] and SIXray [13]. However, GDXray [12] contains images which are grayscale, while another dataset SIXray [13] only contains less than 1% images having annotated prohibited items. And both GDXray [12] and SIXray [13] are used for classification task. As a result, both of the two datasets are inconsistent with our task that detecting occluded prohibited items.

To torch this important topic, we contribute the first high-quality object detection dataset for security inspection, named Occluded Prohibited Items X-ray (OPIXray) image benchmark. Considering that cutter is the most common tool passengers carry, we choose it as the prohibited item to detect. OPIXray contains 8885 X-ray images of 5 categories of cutters (illustrated in Fig. 1). Each sample has at least one prohibited item, while some samples have more. All samples are annotated manually by the professional inspectors from the international airport and the standard of annotating is based on the standard of training security inspectors. Our dataset brings meaningful challenges to this topic in two main folds. First, OPIXray mimics a similar testing environment to the real-world scenario, where items randomly overlapped with each other, leading to object occlusion challenge. Second, cutters of different categories usually share the similar shape appearance, *e.g.*, folding knives and multi-functional knives, bringing difficulties to discriminate.

Furthermore, to deal with the occlusion in X-ray images, we propose the De-occlusion Attention Module(DOAM), a plug-and-play module that can be easily inserted into most popular detectors. As we have observed, X-ray imaging preserves the shape appearance in the heavy occluded part and assigns various colors to different materials in the visual part. Inspired by the fact we observed, our module simultaneously lays particular emphasis on edge information and material information of the prohibited item by utilizing two sub-modules, namely, Edge Guidance (EG) and Material Awareness (MA). Then, our module leverages the two information above to generate an attention distribution map as a high-quality mask for each input sample to generate high-quality feature maps, serving identifiable information for the general detectors.

The main contributions of this work are as follows:

- We provide the first benchmark for occluded prohibited items detection in X-ray images for security inspection. The OPIXray dataset we contributed is high-quality because all prohibited items are manually annotated by professional security inspectors from the international airport.
- We present the De-occlusion Attention Module (DOAM), simultaneously laying particular emphasis on edge information and material information of the prohibited item, inspired by the X-ray imaging principle.
- DOAM can be easily inserted as a plug-and-play module into various detectors, including SSD [10], YOLOv3 [15] and FCOS [20], etc., which means our module can be widely applied.
- We evaluate our method on the OPIXray dataset and compare it to various baselines, including popular detection approaches and widely-used attention mechanisms. These results show that DOAM can not only consistently improve the performance of the state-of-the-art detection methods but also significantly outperform several widely-used attention mechanisms.

2 RELATED WORK

2.1 X-ray Images and Benchmarks

X-ray offers powerful ability in many tasks such as medical imaging analysis [1, 6, 11] and security inspection [8, 13]. However, the visibility of the object information contained in X-ray images suffers a lot because of object occlusion.

Several studies in the literature have attempted to address this challenging problem. Unfortunately, due to the particularity of security inspection, very few X-ray datasets have been published. A released benchmark, GDXray[12] contains 19407 images, partial of which contains three categories of prohibited items including gun, shuriken and razor blade. However, GDXray only contains gray-scale images in a very simple background, which is far away from real-world scenario. Recently, SIXray[13] is a large-scale X-ray dataset which is about 100 times larger than the GDXray dataset[12]. SIXray consists of 1059231 X-ray images, but the positive samples are less than 1% to mimic a similar testing environment to the real-world scenario where inspectors often aim at recognizing prohibited items appearing in a very low frequency. Different from ours, SIXray is a dataset for the task of classification, focusing on the problem of data imbalance.

2.2 Attention Mechanism

Attention can be interpreted as a means of biasing the allocation of available computational resources towards the most informative components of a signal, which has been widely studied in many tasks, like image retrieval [18, 26], visual question answering [14, 28]. It captures long-range contextual information and has been widely applied in various tasks such as machine translation [21], image captioning [2], scene segmentation [4] and object recognition [19]. The work [23] is related to self-attention module, mainly exploring the effectiveness of non-local operation in space-time dimensions for videos and images. [4] proposed a dual attention network (DANet) for scene segmentation by capturing contextual

dependence based on the self-attention mechanism. Squeeze-and-Excitation Networks (SENet) [7] terms the Squeeze-and-Excitation block (SE), that adaptively re-calibrates channel-wise feature responses by explicitly modeling inter-dependencies between channels.

Table 1: The category distribution of the OPIXray dataset. Due to that some images contain more than one prohibited item, the sum of all items in the different categories is greater than the total number of images.

OPIXray	Categories					Total
	Folding	Straight	Scissor	Utility	Multi-tool	
Training	1589	809	1494	1635	1612	7109
Testing	404	235	369	343	430	1776
Total	1993	1044	1863	1978	2042	8885

3 THE OPIXRAY DATASET

The performance of deep learning models largely depends on the quality of the dataset. Only with high quality dataset can the detection ability of a model be evaluated reasonably. Thus, a professional dataset with high-quality annotations is necessary for training models and performing evaluations. In this work, we build the first dataset specially designed for occluded prohibited items detection in security inspection.

3.1 Data properties

Data Acquisition: The backgrounds of all samples are scanned by the security inspection machine and the prohibited items are synthesized into these backgrounds by the professional software. In the international airport, these synthesized images are used to train security inspectors to recognize prohibited items, which is exactly what we want to be executed automatically. And each prohibited item is annotated manually by professional inspectors from the international airport, which localized by a box-level annotation with a bounding box. These X-ray images still retain the specific property that X-ray imaging preserves the shape appearance in the heavy occluded part and assigns various colors to different materials, mainly reflected in the visual part.

Data Structure: The OPIXray dataset contains a total of 8885 X-ray images of 5 categories of cutters, namely, Folding Knife, Straight Knife, Scissor, Utility Knife, Multi-tool Knife. A statistics of category distribution is shown in Tab. 1. All images are stored in JPG format with the resolution of 1225*954. 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), where the ratio is about 4 : 1. The statistics of category distribution of training set and testing set are also shown in Tab. 1. Note that there are 35 images of the dataset, each of which contains more than one prohibited item, by 30 in the training set and 5 in the test set.

Data Occlusion Levels: In order to study the impact brought by object occlusion levels, we divide the testing set into three subsets and name them Occlusion Level 1 (OL1), Occlusion Level 2 (OL2) and Occlusion Level 3 (OL3), where the number indicates occlusion level of prohibited items in images. As illustrated in Fig. 2, there is no or slight occlusion on prohibited items in OL1 and partial occlusion in OL2. To maximally evaluate the ability of models to deal with object occlusion, we construct OL3 by choosing images where severe or full occlusion exists in. The category distribution of the three subsets with different occlusion levels are shown in Tab. 2.

Table 2: The category distribution of different occlusion levels in the testing set.

Testing set	Categories					Total
	Folding	Straight	Scissor	Unility	Multi-tool	
OL1	206	88	160	214	255	922
OL2	148	84	126	88	105	548
OL3	50	63	83	41	70	306
Total	404	235	369	343	430	1776

3.2 Dataset Analysis

Data Authenticity: The OPIXray dataset mostly mimics a similar environment to the real-world scenario. **First**, the occlusion of prohibited items is inspired that items within personal luggage are usually stacked randomly and overlapped with each other, which we describe in detail in this work. **Second**, the statics of category distribution is inconsistent obviously. The number of folding knife and multi-tool knife are higher than straight knife because the former two categories are more common for passengers to bring. And the number of OL3 is significantly less than OL1 because cutters are usually small and move freely in luggage, as a result, cutters are seldom fully occluded in the real scenario.

Data Application: OPIXray dataset has two major application scenarios. **First**, the dataset can evaluate the ability of a model to detect prohibited items in X-ray images. A better model can achieve better performance no matter which occluded levels. As we can see from Fig. 3, there is a significant decline in the performance of

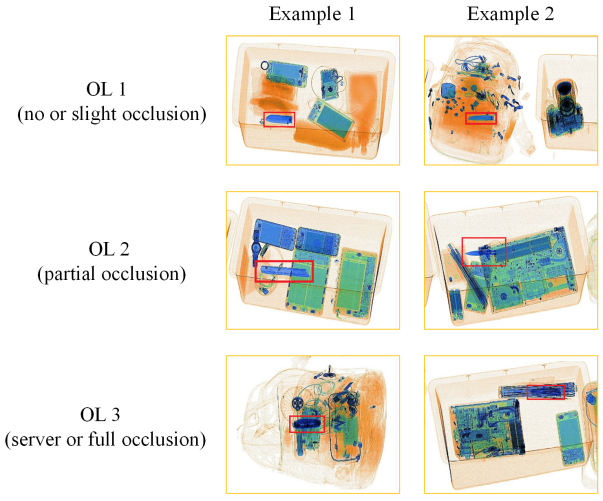
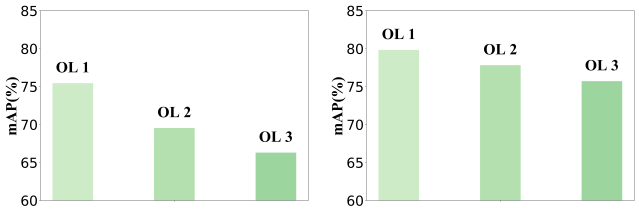


Figure 2: Samples of different occlusion levels.

famous detection approaches *e.g.*, SSD [10] and YOLOv3 [15], with the occlusion level increasing. **Second**, the dataset can evaluate the ability of a model of solving object occlusion problem, by comparing the improvement than other methods in different occlusion level settings. The improvement amount of an approach increases with the occlusion level increases, which illustrates the effectiveness of this approach to the object occlusion problems.



(a) SSD (b) YOLOv3

Figure 3: The performance of SSD and YOLOv3 under three different object occlusion levels.

4 DE-OCCLUSION ATTENTION MODULE

We propose the De-occlusion Attention Module(DOAM), simultaneously laying particular emphasis on edge information and material information of the prohibited item by utilizing two sub-modules, namely, Edge Guidance (EG) and Material Awareness (MA). Then, our module leverages the two information above to generate an attention distribution map as a high-quality mask for each input sample to generate high-quality feature maps, serving identifiable information for general detectors. Without losing of generality, we apply the DOAM to the widely-used SSD [10] and demonstrate our design from the following aspects: 1) how DOAM work briefly (4.1); 2) how to impel the edge information to guide the model (4.2); 3) how to aggregate the region information to express the material information (4.3); 4) how to leverage the two source of information and generate the attention map (4.4); 5) how to compare our module with the base detector and other counterparts (4.5).

4.1 Network Architecture

Fig. 4 illustrates the architecture of the SSD detector with the proposed DOAM. On the top of the SSD, DOAM leverages edge and material information generated by two parallel branches, namely EG and MA, to generate an attention distribution map, providing enhanced features for further accurate detection.

Specifically, suppose there are n training images in the dataset $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$. Each input image $\mathbf{x} \in X$ is fed into EM and MA to obtain $\mathbf{F}_E$ and $\mathbf{F}_M$, laying particular emphasis on edge information of the occluded part and material information of the visual part, respectively. In EG, we first extract the edge map through an edge detection operation and then generate the edge guidance information $\mathbf{F}_E$ which emphasizes the complete edge information of the prohibited item, especially in occluded part. In MA, we take the concatenation of $\mathbf{x}$ and edge guidance $\mathbf{F}_E$ as input and denote it as $\mathbf{P}$, and extract a temporary feature map $\mathbf{F}_{tmp1}$ (Note that $\mathbf{F}_{tmp1}$, $\mathbf{F}_{tmp2}$ and $\mathbf{F}_{tmp3}$ are intermediate states of the refined feature $\mathbf{F}_M$ during the refining process). To further aggregate the region information to emphasis the material characteristics of the input image, we design a Region Information Aggregation (RIA) operation where different pooling kernels are utilized to aggregate multi-scale region-wise features which will be selected by a gated CNN to further adaptively generate the material awareness information $\mathbf{F}_M$. $\mathbf{F}_M$ remains and emphasizes information of identifiable properties of the visible part. Finally, We fuse the edge guidance information $\mathbf{F}_E$ and material awareness information $\mathbf{F}_R$ to obtain the attention distribution map $\mathbf{S}$. With the help of $\mathbf{S}$, we can obtain the enhanced features $\mathbf{F}$ of the input image for further accurate detection. The entire process of DOAM is illustrated in detail in Algorithm 1.

Algorithm 1 The Operation Process of DOAM.

Input: an X-ray image $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$;
Generate the horizontal edge image $\mathbf{E}^h$ and the vertical edge image $\mathbf{E}^v$ by the *Sobel* operator.
Generate the edge image $\mathbf{E}$ by synthesizing $\mathbf{E}^h$ and $\mathbf{E}^v$.
for N_1 steps **do**
    Refine the feature map $\mathbf{F}_E$ by operating $\mathbf{E}$ through $f_{\mathbf{e}}(\cdot)$.
end for
Generate the concatenated image $\mathbf{P}$ by concatenating $\mathbf{x}$ and $\mathbf{E}$.
for N_2 steps **do**
    Refine the feature map $\mathbf{F}_{tmp1}$ by operating $\mathbf{P}$ through $f_{\mathbf{r}}(\cdot)$.
end for
for $k \in \{\mathbf{k}_1, \ldots, \mathbf{k}_n\}$ **do**
    Generate refined feature map $\mathbf{F}^k_{tmp2}$ by operating $\mathbf{F}_{tmp1}$ through Eq. (5).
    Generate refined feature map $\mathbf{F}^k_{tmp3}$ by concatenating $\mathbf{F}_{tmp1}$ and $\mathbf{F}^k_{tmp2}$.
    Update the feature map set $S = S \cup \mathbf{F}^k_{tmp3}$.
end for
Choose the appropriate feature map $\mathbf{F}_M$ from S by drawing the gated convolutional network.
Generate the fused feature map $\mathbf{F}_{fus}$ by operating $\mathbf{F}_E$ and $\mathbf{F}_M$.
Generate the attention map $\mathbf{S} = \sigma(\mathbf{F}_{fus})$.
Generate the final feature map $\mathbf{F}$ by performing a matrix multiplication between $\mathbf{S}$ and $\mathbf{P}$.
Output: the refined feature map $\mathbf{F} \in \mathbb{R}^{C_h \times H \times W}$.

4.2 Edge Guidance (EG)

For each input image $\mathbf{x} \in X$, we utilize the convolutional neural network with the horizontal and vertical kernel denoted as s_h, s_v of the *Sobel* operator, to respectively compute the edge images $\mathbf{E}^h$ and $\mathbf{E}^v$ in horizontal and vertical directions. We further generate the edge image $\mathbf{E}$ of the input image $\mathbf{x}$ by synthesizing the above two results $\mathbf{E}^h$ and $\mathbf{E}^v$. To lead EG to only magnify edge information of the prohibited items, we use $\mathbf{N}_1$ network blocks (Here, we define $\mathbf{N}_1$ as the Module Operation Intensity of EG, which represents that the performance of the module changes with the value of $\mathbf{N}_1$.), in which each block consists of a convolutional layer with a 3×3 kernel size, a batch normalization layer, and a relu layer, to extract the feature map $\mathbf{F}_E$. The operations can be formulated as follows:

$$f_{\mathbf{e}}(\mathbf{a}) = relu\left(\mathbf{W}_e \cdot \mathbf{a} + \mathbf{b}_e\right). \quad (1)$$

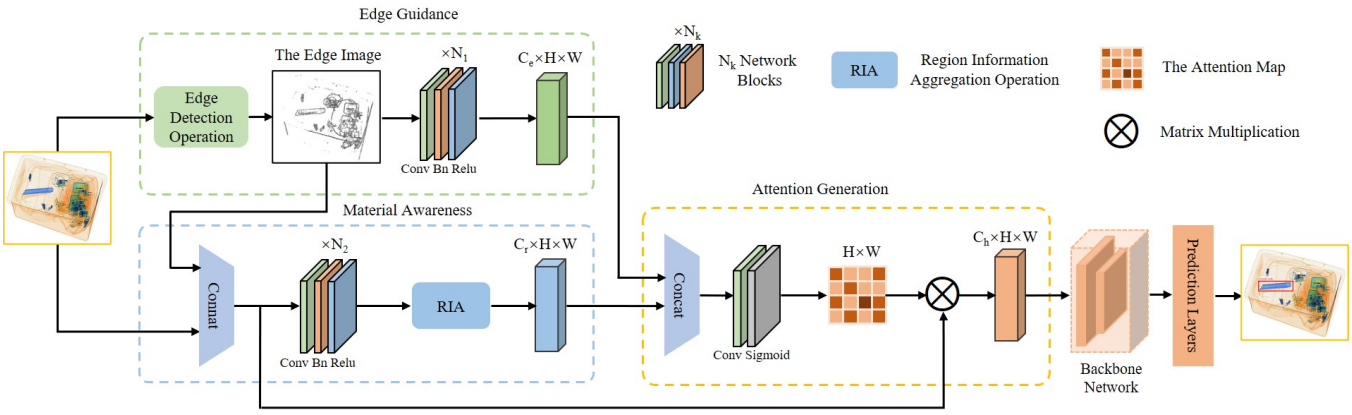


Figure 4: DOAM integrated with a general backbone network architecture. As illustrated, two feature maps are generated by Edge Guidance (EG) and Material Awareness (MA) and fused to generate the attention map in Attention Generation. Further, the attention map is applied to the input image to generate refined feature maps we desire. Finally, the refined feature map can be utilized by the SSD network.

$$\mathbf{F}_E = \{f_{\mathbf{e}}(\mathbf{E})\}_{N_1}. \tag{2}$$

where $\{\cdot\}_{N_1}$ means that the operation is repeated N_1 times, $\mathbf{W}_e$, $\mathbf{b}_e$ are parameters of the convolutional layer. After extracting the feature map $\mathbf{F}_E$ as shown in Eq. 2, we adaptively attend to the edge guidance information of the prohibited item within the feature map $\mathbf{F}_E$ by optimization.

4.3 Material Awareness (MA)

Material information is mainly reflected in color and texture. For color information, each position of the image has the ability to represent. However, when it comes to the texture information, each position needs to combine its surroundings to represent. Inspired by the fact that the aggregation of regional information can represent both color and texture, we define that the region information after aggregated is the representation of the material information. In order to construct relations between each position of the concatenated image (the input image $\mathbf{x}$ and its edge image $\mathbf{E}$ in EG) and a certain region around the point, we utilize $\mathbf{N}_2$ network blocks (As we state $\mathbf{N}_1$ in EG, $\mathbf{N}_2$ is the Module Operation Intensity of MA.), in which each block consists a convolutional layer with the kernel size is 3×3, a batch normalization layer, a relu layer, to extract a temporary feature map $\mathbf{F}_{tmp1}$ from the concatenated image as follows:

$$f_{\mathbf{r}}(\mathbf{a}) = relu\left(\mathbf{W}_r \cdot \mathbf{a} + \mathbf{b}_r\right). \tag{3}$$

$$\mathbf{F}_{tmp1} = \{f_{\mathbf{r}}(\mathbf{x}||\mathbf{E})\}_{N_2}. \tag{4}$$

where $||$ represents concatenating operation. We further generate the refined feature map $\mathbf{F}_M$ of MA by refining $\mathbf{F}_{tmp1}$ through the Region Information Aggregation (RIA) operation, central of MA.

Fig. 5 illustrates the detailed process of RIA operation. For the input feature map $\mathbf{F}_{tmp1}$ and a parameter k, RIA operation aggregates the information of a certain size of $k \times k$ of region around it by average pooling and extending to generate another temperate feature map $\mathbf{F}^k_{tmp2}$. The average pooling and extending operations can be formulated together as follows:

$$\mathbf{F}_{tmp2}{}^k_{ij} = \frac{\sum_{m=i-(i \bmod k)}^{i-(i \bmod k)+k} \sum_{n=j-(j \bmod k)}^{j-(j \bmod k)+k} \mathbf{F}_{tmp1}{}_{mn}}{k^2}. \tag{5}$$

where $\mathbf{F}_{tmp2}{}^k_{ij}$ represents the feature of the i-th row and j-th column of feature map $\mathbf{F}_{tmp2}$ when the kernel size for the average pooling layer is k.

We further concatenate the two feature maps ($\mathbf{F}_{tmp1}$ and $\mathbf{F}_{tmp2}$) in the dimension of channel to generate a new feature map $\mathbf{F}_{tmp3}$, where the dimensions are $2C_r \times H \times W$. Then every point of the new feature map has the ability to perceive the region of size $k \times k$ around it, which means that the relations have been constructed. Due to different sizes of region information to aggregate (different values of k), the module generates a feature map set $S = \left\{\mathbf{F}^{k_1}_{tmp3}, \cdots, \mathbf{F}^{k_n}_{tmp3}\right\}$.

In order to enable RIA operation to perform well on various scales of prohibited items, it is necessary to design a mechanism to adaptively choose an optimal value for k. We exploit the gated convolutional neural network [27] $\mathbb{G}$ with 3×3 kernels into RIA, to select the proper feature map $\mathbf{F}_M$ from the feature map set S as output. The operations are formulated as follows:

$$\mathbf{F}_M = \mathbb{G}(S). \tag{6}$$

where $S = \left\{\mathbf{F}^{k_1}_{tmp3}, \cdots, \mathbf{F}^{k_n}_{tmp3}\right\}$.

4.4 Attention Generation

As is illustrated in Algorithm 1, for the result feature maps $\mathbf{F}_E$ and $\mathbf{F}_M$ outputted by the EG and the MA respectively, where $\mathbf{F}_E \in \mathbb{R}^{C_e \times H \times W}$, $\mathbf{F}_M \in \mathbb{R}^{C_r \times H \times W}$, we concatenate them for information fusion. And further we feed the concatenated feature into a convolutional layer, where the kernel size is 1×1, to generate the feature map $\mathbf{F}_{fus} \in \mathbb{R}^{(C_e+C_r) \times H \times W}$, which have confused the edge information and the reginal information, both strengthened. The operation can be formulated as follows:

$$\mathbf{F}_{fus} = \mathbf{W}_m\left(\mathbf{F}_E||\mathbf{F}_M\right) + \mathbf{b}_m. \tag{7}$$

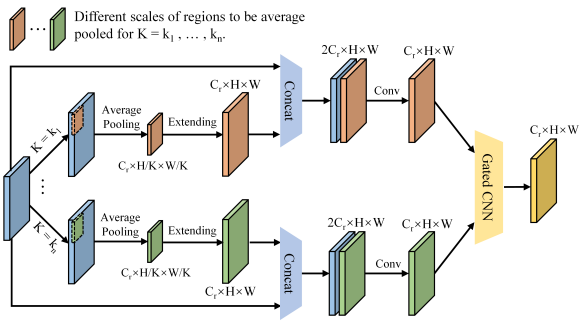


Figure 5: The Operation Process of RIA.

where || represents the operation of concatenating, and $\mathbf{W}_m$, $\mathbf{b}_m$ are parameters of the convolutional layer. Then we utilize the feature map $\mathbf{F}_{fus}$ as the input of a sigmoid function to generate the attention map $\mathbf{S}$ as follows:

$$\mathbf{S} = \sigma(\mathbf{F}_{fus}) = \frac{1}{1 + e^{-\mathbf{F}_{fus}}}. \tag{8}$$

where $\mathbf{S} \in \mathbb{R}^{H \times W}$. Finally, we calculate the inner product of the attention map $\mathbf{S}$ and the concatenated image $\mathbf{P}$ as follows:

$$\mathbf{F}_j = \sum_{i=1}^{H \times W} \mathbf{S}_{ji} \mathbf{P}_i. \tag{9}$$

where $\mathbf{F} \in \mathbb{R}^{C_h \times H \times W}$, and it is the final refined feature map we desire to serve to detectors, of which the information highly contributes to the detection of the prohibited item are emphasised.

4.5 Module Complexity Analysis

In this section, we analyze the model complexity with or without DOAM in SSD [10] and compare the complexity including the total number of parameters, model size and computation cost, with other attention mechanisms.

Table 3 reports that DOAM only brings a slight increase in computational cost (7.14% in GFLOPs), compared to the SSD [10] without any attention mechanisms. Additionally, we compare the complexity between DOAM and three variants of attention mechanisms, including SE [7], Non-local [23] and DA [4]. The three attention mechanisms focus on channel information, spatial information and combination of the two kinds of information, respectively. As we can see from table 3, compared to the single SSD [10], **First**, for total number of parameters, SE [7], Non-local [23] and DA [4] respectively bring 32.23%, 27.69% and 88.43% increases, while the increase our module brings is almost negligible. **Second**, for model size, SE [7], Non-local [23] and DA [4] respectively bring 31.86%, 27.32% and 88.01% increases, while the increase our module brings is almost negligible. However, for computational cost, SE [7], Non-local [23] and DA [4] respectively bring 3.15%, 6.54% and 23.07% increases, while our module brings 7.14%.

In conclusion, for total number of parameters and model size, DOAM is much more computation efficient than the three famous attention mechanisms. For computational cost, DOAM is slightly more computationally expensive. We conjecture that it is mainly because that different values of parameter k cause repetitive computation in RIA.

Table 3: Complexity comparison of different models. PARAMs, SIZE and GFLOPs represent the total number of parameters, the Model Size and the Giga Floating Point operations, respectively.

Method	PARAMs	SIZE(MB)	GFLOPs
SSD [10]	24.2×10^6	92.6	30.6522
SSD+SE [7]	32.0×10^6	122.1	**31.6169**
SSD+Non-local [23]	30.9×10^6	117.9	32.6577
SSD+DA [4]	45.6×10^6	174.1	37.7231
SSD+DOAM(ours)	$\mathbf{24.3 \times 10^6}$	**92.7**	32.8435

Table 4: Performance comparison between DOAM and other different attention mechanisms on object categories. FO, ST, SC, UT and MU represent Folding Knife, Straight Knife, Scissor, Utility Knife and Multi-tool Knife, respectively.

Method	mAP	Categories				
		FO	ST	SC	UT	MU
SSD [10]	70.89	76.91	35.02	93.41	65.87	83.27
SSD+SE [7]	71.85	77.17	38.29	92.03	66.10	**85.67**
SSD+Non-local [23]	71.41	77.55	36.38	**95.26**	64.86	82.98
SSD+DA [4]	71.96	79.68	37.69	93.38	64.14	84.90
SSD+DOAM(ours)	**74.01**	**81.37**	**41.50**	95.12	**68.21**	83.83

5 EXPERIMENTS

In this section, we carry on extensive experiments to evaluate the DOAM we proposed. In our work, the main task is to detect occluded prohibited items in X-ray images in security inspection scenario. To the best of our knowledge, no dataset targeting this task has been proposed in the literature, so we only adopt the OPIXray dataset in the experiments. **First**, we verify that DOAM outperforms all the attention mechanisms mentioned above, over different categories and different occlusion levels. **Second**, we perform ablation experiments to thoroughly evaluate the effectiveness of DOAM. **Third**, we demonstrate the general applicability of DOAM across different architectures and the effectiveness after DOAM-integrated. **Finally**, we apply the Grad-CAM [16] to visualize the attention mechanism of DOAM.

Evaluation strategy: All experiments are carried on the OPIXray dataset. In experiments of comparing with different attention mechanisms over different occlusion levels, every model is trained by training set data in Tab. 1 and tested on OL1, OL2 and OL3 in Tab. 2 respectively. In any other experiments, every model is trained by training set data and tested by the testing set data in Tab. 1.

Baseline Detail: In experiments of comparing with different attention mechanisms, we respectively plug DOAM and each of the attention modules into SSD [10] and report the performances of SSD [10] and these integrated networks. In our experiments, these attention modules are added to the backbone (VGG16) of

SSD. More specifically, they are inserted behind the max pooling layer where the feature map is scaled to half. In ablation study, we plug each sub-module of DOAM into SSD [10] separately and report the performances of SSD [10] and these integrated networks. In experiments of comparing with different detection approaches, we plug DOAM into a number of popular detection networks and report the model performance with or without DOAM in every detection network.

Parameter setting: In all experiments following, all models are optimized by the SGD optimizer and the learning rate is set to 0.0001. The batch size is set to 24 and the momentum and weight decay are set to 0.9 and 0.0005 respectively. We evaluate the mean Average Precision (mAP) of the object detection to measure the performance of the model and the IOU threshold is set to 0.5. We further select the best performance model to calculate the AP of each category to observe the performance improvement in different categories. Furthermore, in order to avoid the influence of image data modification on edge image generation, we do not use any data augmentation methods to expand the data or modify the pixel value of the original image, which helps us to better analyze the impact of edge information.

5.1 Comparing with Different Attention Mechanisms

We compare three variants of attention mechanisms above, including SE [7], Non-local [23] and DA [4]. Tab. 4 and 5 reports the performances of all models.

Object Categories: As we observed from Tab. 4, the DOAM-integrated model outperforms SSD [10] by 3.12%. Besides, DOAM outperforms SE [7], Non-local [23] and DA [4], by 2.16%, 2.60%, 2.05%, respectively. Moreover, Tab. 4 shows the improvement of DOAM is mainly reflected in Straight Knife, Folding Knife and Utility Knife, all of which are with the high level occlusion. Especially for Straight Knife, which is the category with highest level occlusion, DOAM outperforms SSD [10] by an impressive amount of 6.48% and Non-local [23] by 5.12%. For Scissor, the lightest occlusion category, the performance of DOAM is only improved by 1.71% compared to SSD [10] and similar to Non-local [23]. It is obvious that DOAM surpasses these current popular attention mechanisms over different categories.

Object Occlusion Levels: The experimental results are shown in Tab. 5. Further, Fig. 6 is drawn according to Tab. 5 to illustrate the effectiveness of DOAM to occluded object detection in X-ray images more clearly. In Fig. 6, we can clearly obtain a conclusion that DOAM can achieve a higher performance than the baseline and other attention mechanisms with the X-ray images suffer a higher level of occlusion. It verifies the effectiveness of DOAM that it has a significant effect on the performance of detecting occluded prohibited items in X-ray images. (Note that in OL3, the performance of "SSD+Non-local" is lower than "SSD". Due to the attention mechanism of Non-local is to capture spatial information by constructing the relations between regions, we conjecture that this type of relation reduces effect when the noises of image increase.)

Table 5: Performance comparison between DOAM and other different attention mechanisms on object occlusion levels.

Method	OL 1	OL 2	OL 3
SSD [10]	75.45	69.54	66.30
SSD+SE [7]	76.02	70.11	67.53
SSD+Non-local [23]	75.99	70.17	65.87
SSD+DA [4]	77.41	69.68	66.93
SSD+DOAM(ours)	**77.87**	**72.45**	**70.78**

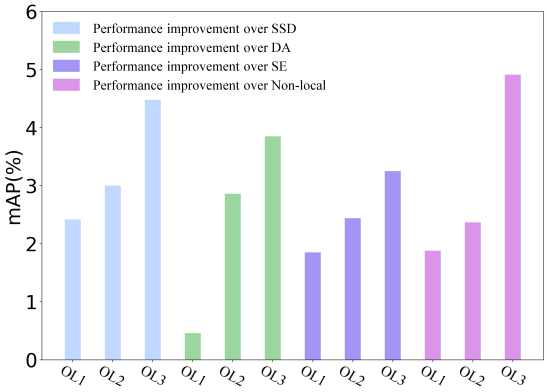


Figure 6: The amount changes of performance improvement of DOAM over different models with occlusion level increasing.

5.2 Ablation Study

Tab. 6 shows that EG improved the performance by 0.43% compared with the method of simply concatenating the input image and the corresponding edge image without any other operations of EG. We conjecture that it is mainly because the EG has the ability to focus adaptively on the prohibited items we desire to detect by specifically increasing the weight of edge information through the optimization of a loss function, while simply concatenating operates all the objects in the image equally whether the object is we desire to detect or not for feature fusion.

Besides, model integrating both EG and MA achieves better performance by 0.37% than integrating EG alone, which verifies the effectiveness of MA. Note that we observe prohibited item size is about 10×10 averagely, so we choose 10×10 as the region scale to perceive for each position of the feature map.

We choose three different scales of the regions (5×5, 10×10, 15×15 respectively), and draw the gated convolutional neural network [27] $\mathbb{G}$ into MA, to adaptively select the best feature map which generated by operation of average pooling with appropriate pooling size. The experimental results show that after drawing $\mathbb{G}$, the performance improves by 0.9%.

Table 6: Ablation studies of DOAM. "C" represents simply concatenate operation, "DOAM-MA" represents DOAM without Material Awareness module, "DOAM-$\mathbb{G}$" represents DOAM without the Gate Convolutional Neural Network.

Method	mAP	Category				
		FO	ST	SC	UT	MU
SSD [10]	70.89	76.91	35.02	93.41	65.87	83.27
SSD+C	72.32	79.00	36.46	94.13	68.85	83.18
SSD+(DOAM-MA)	72.75	80.26	35.54	94.81	67.96	**85.19**
SSD+(DOAM-$\mathbb{G}$)	73.12	79.94	38.58	93.39	**69.40**	84.28
SSD+DOAM(ours)	**74.01**	**81.37**	**41.50**	**95.12**	68.21	83.83

5.3 Comparing with Different Detection Approaches

To further evaluate the effectiveness of DOAM and verify DOAM can be applied to various detection networks, we conduct experiments on the famous detection approaches, SSD [10], YOLOv3 [15] and FCOS [20]. The results are shown in Tab. 7.

Table 7: Performance comparison between DOAM-integrated network and baselines for three famous detection approaches.

Method	mAP	Category				
		FO	ST	SC	UT	MU
SSD [10]	70.89	76.91	35.02	93.41	65.87	83.27
SSD+DOAM(ours)	**74.01**	**81.37**	**41.50**	**95.12**	**68.21**	**83.83**
YOLOv3 [15]	78.21	**92.53**	36.02	**97.34**	70.81	94.37
YOLOv3+DOAM(ours)	**79.25**	90.23	**41.73**	96.96	**72.12**	**95.23**
FCOS [20]	82.02	86.41	68.47	90.22	78.39	86.60
FCOS+DOAM(ours)	**82.41**	**86.71**	**68.58**	**90.23**	**78.84**	**87.67**

As we can see from Tab. 7, the performance of DOAM-integrated networks are improved by 3.12%, 1.04% and 0.39% compared with SSD [10], YOLOv3 [15] and FCOS [20] respectively, which verify that our module can be inserted as a plug-and-play module into most detection networks and receive a better performance. Note that the performances on Folding Knife and Scissor after DOAM-integrated are slightly reduced. We speculate that the reason is that these images of the two categories in the dataset are occluded not seriously. When the occlusion level increases, the attention mechanism pays more attention to the objects occluded highly like straight knives while less attention to the objects occluded slightly, which results in the slight performance degradation for Folding Knife and Scissor.

5.4 Attention Visualization Analysis

In this section, we visualize the attention map generated in DOAM to observe the effects of DOAM. The attention distribution can be visualized in Fig. 7. In rows 1 and 3, we select 10 input X-ray images (each category has two images) and show their corresponding attention visualizations in rows 2 and 4. We observe that DOAM could capture edge and region information accurately. For example, in column 4, a red box is marked on a utility knife of the X-ray image (in row 1), and the boundaries of the utility knife are very clear in the attention visualization (in row 2). Moreover, in the first column, a red box is marked on a folding knife and the corresponding attention map (in row 2) highlights most of the areas where the folding knife lies on. In short, these visualizations further demonstrate the effectiveness of capturing edge and region information for improving feature representation in occluded prohibited items detection.

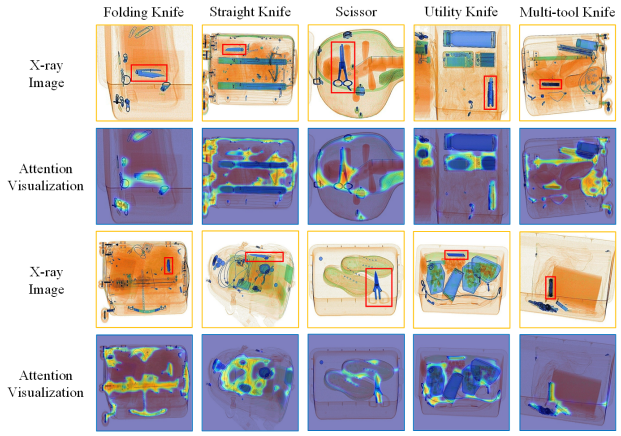


Figure 7: Attention visualization results.

6 CONCLUSION

In this paper, we investigate occluded prohibited items detection in X-ray scanned images, which is a promising application in industry yet remains fewer studied in computer vision. To facilitate research in this field, we contribute the first high-quality object detection dataset for security inspection, named Occluded Prohibited Items X-ray (OPIXray) image benchmark. OPIXray focused on the widely-occurred prohibited item "cutter", annotated manually by professional inspectors from the international airport. To deal with the occlusion in X-ray images detection, we propose the De-occlusion Attention Module (DOAM), a plug-and-play module that can be easily inserted into and thus promote most popular detectors. We comprehensively evaluate our module on the OPIXray dataset, and demonstrate that our module can consistently improve the performance of the state-of-the-art detection methods such as SSD, FCOS, etc, and significantly outperforms several widely-used attention mechanisms. In particular, the advantages of DOAM are more significant in the scenarios with higher levels of occlusion, which demonstrates its potential application in real-world inspections.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (61872021), Beijing Nova Program of Science and Technology (Z191100001119050), State Key Lab of Software Development Environment (SKLSDE-2020ZX-06) and Fundamental Research Funds for Central Universities (YWF-20-BJ-J-646).

REFERENCES

[1] Arjun Chaudhary, Abhishek Hazra, and Prakash Chaudhary. 2019. Diagnosis of Chest Diseases in X-Ray images using Deep Convolutional Neural Network. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 1–6.

[2] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5659–5667.

[3] Zhi-Qi Cheng, Jun-Xiu Li, Qi Dai, Xiao Wu, Jun-Yan He, and Alexander G Hauptmann. 2019. Improving the learning of multi-column convolutional neural network for crowd counting. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1897–1906.

[4] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. 2019. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3146–3154.

[5] Shiming Ge, Jia Li, Qiting Ye, and Zhao Luo. 2017. Detecting masked faces in the wild with lle-cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2682–2690.

[6] Shuai Guo, Songyuan Tang, Jianjun Zhu, Jingfan Fan, Danni Ai, Hong Song, Ping Liang, and Jian Yang. 2019. Improved U-Net for Guidewire Tip Segmentation in X-ray Fluoroscopy Images. In *Proceedings of the 2019 3rd International Conference on Advances in Image Processing*. 55–59.

[7] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[8] Shengling Huang, Xin Wang, Yifan Chen, Jie Xu, Tian Tang, and Baozhong Mu. 2019. Modeling and quantitative analysis of X-ray transmission and backscatter imaging aimed at security inspection. *Optics express* 27, 2 (2019), 337–349.

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[10] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

[11] Jianjie Lu and Kai-yu Tong. 2019. Towards to Reasonable Decision Basis in Automatic Bone X-Ray Image Classification: A Weakly-Supervised Approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9985–9986.

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

[13] Caijing Miao, Lingxi Xie, Fang Wan, Chi Su, Hongye Liu, Jianbin Jiao, and Qixiang Ye. 2019. 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*. 2119–2128.

[14] Liang Peng, Yang Yang, Zheng Wang, Xiao Wu, and Zi Huang. 2019. CRA-Net: Composed Relation Attention Network for Visual Question Answering. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1202–1210.

[15] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[16] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[17] Lingxue Song, Dihong Gong, Zhifeng Li, Changsong Liu, and Wei Liu. 2019. Occlusion Robust Face Recognition Based on Mask Learning With Pairwise Differential Siamese Network. In *Proceedings of the IEEE International Conference on Computer Vision*. 773–782.

[18] Xie Sun, Lu Jin, and Zechao Li. 2019. Attention-Aware Feature Pyramid Ordinal Hashing for Image Retrieval. In *Proceedings of the ACM Multimedia Asia on ZZZ*. 1–6.

[19] Jinhui Tang, Lu Jin, Zechao Li, and Shenghua Gao. 2015. RGB-D object recognition via incorporating latent data structure and prior knowledge. *IEEE Transactions on Multimedia* 17, 11 (2015), 1899–1908.

[20] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 9627–9636.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[22] Pichao Wang, Zhaoyang Li, Yonghong Hou, and Wanqing Li. 2016. Action recognition based on joint trajectory maps using convolutional neural networks. In *Proceedings of the 24th ACM international conference on Multimedia*. 102–106.

[23] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7794–7803.

[24] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. 2018. Repulsion loss: Detecting pedestrians in a crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7774–7783.

[25] Jian Yang, Lei Luo, Jianjun Qian, Ying Tai, Fanlong Zhang, and Yong Xu. 2016. Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. *IEEE transactions on pattern analysis and machine intelligence* 39, 1 (2016), 156–171.

[26] Xingxu Yao, Dongyu She, Sicheng Zhao, Jie Liang, Yu-Kun Lai, and Jufeng Yang. 2019. Attention-aware polarity sensitive embedding for affective image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*. 1140–1150.

[27] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2019. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*. 4471–4480.

[28] Zheng-Jun Zha, Jiawei Liu, Tianhao Yang, and Yongdong Zhang. 2019. Spatiotemporal-Textual Co-Attention Network for Video Question Answering. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 2s (2019), 1–18.

[29] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. 2018. Occlusion-aware R-CNN: detecting pedestrians in a crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 637–653.

[30] Chunluan Zhou and Junsong Yuan. 2018. Bi-box regression for pedestrian detection and occlusion estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 135–151.