Proceedings of 10th IEEE International Conference
on CYBER Technology in Automation, Control, and Intelligent Systems
October 10 - 13, 2020, Xi'an, China

Automatic Annotation Approach for Prohibited Item in X-ray Image based on PANet

Bowen Ma, Tong Jia*, Songsheng Wu

Abstract— **In this paper, we propose an approach based on Polygon-Attention Network (PANet) for automatic annotation of prohibited item instances in X-ray images, aiming at accelerating annotation process for new datasets. X-ray image is fully special, mainly because it has a large amount of overlapping phenomenon, resulting in blurred boundaries of prohibited items. To solve this problem, we add an adaptive multi-level attention module to a generic encoder-decoder annotation architecture, which enables the model to adaptively fuse the required middle layer features in real-time according to characteristics of object boundary during each inference. To evaluate the proposed approach, we present a high-quality X-ray segmentation dataset named Prohibited Item X-ray (PIXray). Experimental results demonstrate that, through our approach, the annotation process speeds up by a factor of 2.4 in all classes in PIXray, and achieves an accord of 92.3% with ground-truth in IoU.**

I. Introduction

Recently, in order to prevent the occurrence of terrorism and security incidents, security inspections have become particularly important. Among all kinds of inspection equipment, a low-cost, non-touch, and imageable X-ray security inspection machine is the most common one. Recent years, deep learning develops rapidly, especially convolutional neural networks. Its application in the field of smart security inspection is triggering the upsurge of research and development of new technologies, e.g., prohibited item automatic detection [1], prohibited item instance segmentation [2], and threat image projection. But unfortunately, there are only two X-ray datasets which are published for research purposes, GDXray [3] and SIXray [4], and both of them have some limitations. GDXray dataset only contains about 1,000 X-ray images, in which three major categories of items are prohibited, including shuriken, gun and razor blade, and few background clutters and overlap are provided in the images, so GDXray cannot mimic real-world scenarios. Unlike GDXray, SIXray dataset contains a large number of X-ray images collected from subway stations. Six categories are included in this dataset, gun, knife, wrench, pliers, scissors and hammer respectively. But the author only provided weak annotations such as image tags or bounding boxes. Problems like data-scarce, few-variety, and weak-annotation have restricted research in this field.

However, manually labeling a dataset of such a large scale and high quality is a both time-consuming and expensive task, especially labeling very detailed annotations such as object segmentation masks [5]. Moreover, due to the overlapping phenomenon, the current annotation methods are not for X-ray images. Overlapping phenomenon occurs when personal luggage is inspected in X-ray. Because of the penetration property of X-ray, we can see both the front objects and the occluded objects in the image. It raises several new challenges to image annotation, e.g., the images can be heavily cluttered, and prohibited items can appear in any angle of view in any postures, sizes, and shapes, resulting in blurred boundaries. In this paper, we solve the above problem by adding an adaptive multi-level attention module to a generic encoder-decoder annotation architecture, which enables the model to adaptively fuse the required middle layer features in real time according to characteristics of object boundary during each inference. Experimental results verify the effectiveness of our approach.

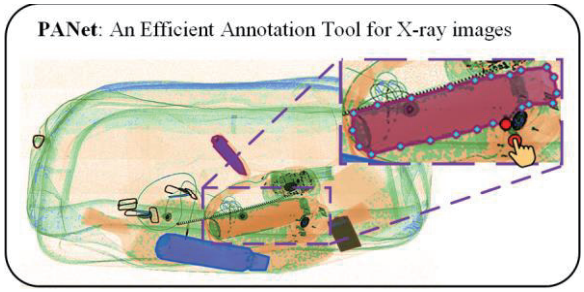

Fig. 1. Given a bounding box around the prohibited item of interest, our model can predict the polygon outlining of the prohibited item instance inside the box automatically. And our approach allows the user to correct the deviant vertices at the end.

There are two major contributions of this work. (1) We propose an efficient approach based on Polygon-Attention Network (PANet) for automatic annotation of prohibited items in X-ray images, as shown in Fig. 1. (2) We present a high-quality X-ray segmentation dataset named Prohibited Item X-ray (PIXray) for our annotation research and other challenging work in this field.

II. PANet Model

In this section, we will introduce PANet model, which is used to label prohibited item instances with closed polygons. The whole PANet consists of two parts, namely, a generic encoder-decoder architecture and a core adaptive multi-level attention module. We follow the idea of Polygon-RNN [5], and use a generic encoder-decoder architecture to predict

All authors are with the College of Information Science and Engineering, Northeastern University, Shenyang, China.
*Contacting Author: Tong Jia is with the College of Information Science and Engineering, Northeastern University, Shenyang, China. (phone: 13332421509; email: jiatong@ise.neu.edu.cn).

This research is supported by the Notional Natural Science Foundation of China (Grant No.U1613214) and the National Key Research and Development Program of China (No.2018YFB14041).

978-1-7281-9010-5/20/$31.00 © 2020 IEEE

Authorized licensed use limited to: HANGZHOU DIANZI UNIVERSITY. Downloaded on March 31,2022 at 03:34:20 UTC from IEEE Xplore. Restrictions apply.

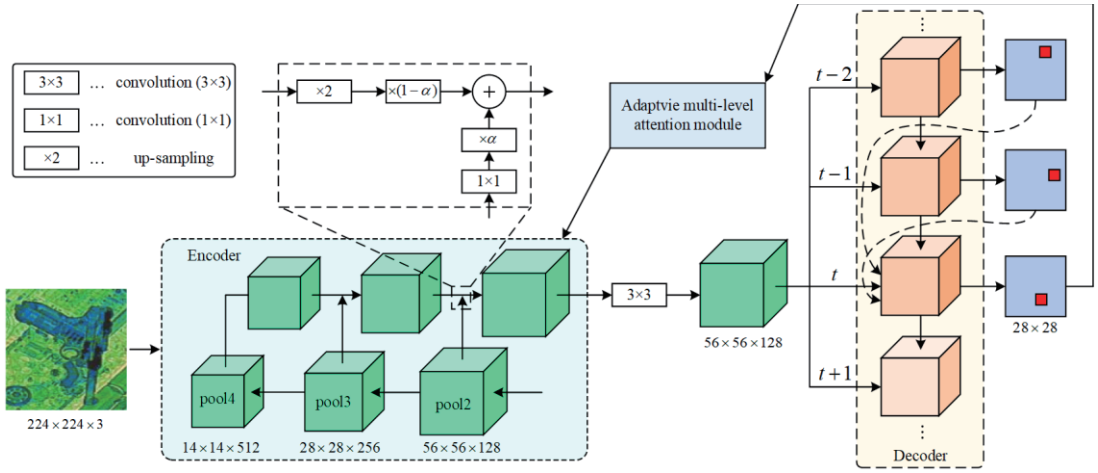

Fig. 2. The overview of our PANet model.

polygons. On this basis, we further add an adaptive multi-level attention module to the encoder-decoder architecture to eliminate the influence of overlapping phenomenon, as shown in Fig. 2. Specifically, we parameterize a polygon as a sequence of 2D vertices, in which two consecutive vertices form an edge. It is worth noting that a closed polygon is a cyclic sequence. Thus, we can obtain several equivalent parameterizations by defining any of the vertices as the first vertex, as well as selecting different directions. In this paper, the polygon is fixed to follow the clockwise direction, while the first vertex is arbitrary and automatically selected by the model.

We assume the annotator provides an initial bounding box around one prohibited item of interest each time, and our model will automatically crop out the image patch inside the box as input, where the encoder is used to extract image features, including low-level spatial visual features including corners, edges, et al, and high-level semantic features like object- or category-level evidence. The decoder is used to predict one polygon vertex at a time step, and the adaptive multi-level attention module is used to automatically fuse the features required at the next time step according to the vertex predicted at the current time step. Output vertices are represented as locations in a gird. At the end, our model allows the users to correct deviant vertex to produce as accurate annotation results as desired. Fig. 2 is the overview of the PANet model.

We start by introducing the generic encoder-decoder architecture, and then detail the core adaptive multi-level attention module.

A. Generic Encoder-Decoder Architecture

Most networks fuse features only by simple concatenate operations, but due to the overlapping phenomenon, this fusion way dose not perform well on X-ray images. In order to alleviate this issue, we construct a special feature pyramid, which also adopt a bottom-up pathway, a top-down pathway, and lateral connections like Feature Pyramid Networks (FPN) [6], but the difference is that when hallucinating higher resolution features from higher pyramid levels by up-sampling, it needs to be point-wise multiplied by a coefficient 1-α after up-sampling. The encoder architecture is shown in Fig. 2, and the adaptive coefficient α will be detailed in Section. B.

We adopt the VGG-16 architecture [7] as our encoder and modify it for our task. We retain the output of *pool2*, *pool3*, and *pool4*, where the former helps the model generate sharp and detailed boundaries, while the latter enables the model to "see" the object.

Following [5], we also employ a Convolutional LSTM (ConvLSTM) [8] as our decoder. ConvLSTMs operate in 2D, so it can not only predict the vertices of a polygon outlining the prohibited item continuously, but also retains the spatial information received from CNN. A simplest form of single-layer ConvLSTM equations are defined as follows:

$$\begin{pmatrix} i_t \\ f_t \\ g_t \\ o_t \end{pmatrix} = W_x * X_t + W_h * H_{t-1} + b$$

$$C_t = \sigma(f_t) \odot C_{t-1} + \sigma(i_t) \odot \tanh(g_t) \qquad (1)$$

$$H_t = \sigma(o_t) \odot \tanh(C_t)$$

where X_t denotes the input, H denotes the hidden state while C is the cell state. i, f, o are the input, forget, and output gate respectively. W_x denotes the input-to-state convolution kernel and W_h the hidden-to-state convolution kernel. $\odot$ denotes an element-wise product and $*$ is a convolution. σ represents the sigmoid function.

In particular, we use a single-layer ConvLSTM with kernel size of 3×3 and 32 channels as the decoder to model polygon, outputting one vertex at each time step. Fig. 2 depicts its architecture. We turn the vertex prediction into classification and thus introduce a fully connected layer. Specifically, at time step *t*, the output is represented as one-hot encoding of a $D \times D+1$ grid. The first $D \times D$ dimensions indicate the possible 2D positions of the vertex, while the last dimension represents the end-of-sequence is token, in other words, the polygon is closed.

Our decoder gets four kinds of information as input, namely, the CNN feature representation of the image, the

Authorized licensed use limited to: HANGZHOU DIANZI UNIVERSITY. Downloaded on March 31,2022 at 03:34:20 UTC from IEEE Xplore. Restrictions apply.

first vertex y_0 and the one-hot encodings of the two previous vertices y_{t-1} and y_{t-2}.

Given two consecutive previous vertices and an implicit direction, the next vertex of a polygon is uniquely defined, except for the first vertex. We use two virtual vertices to initialize the first vertex to solve this problem and reduce the complexity of the model, and the positions (coordinates) of both virtual vertices are set to (0, 0).

B. Adaptive Multi-Level Attention Module

In order to get high quality annotations in overlapping image data, we propose an adaptive multi-level attention module. Its function is to enable the model to adaptively fuse the middle layer features required at the next time step according to the current predicted vertex information. In other words, when the decoder predicts the edge details of a polygon, it more needs the low-level spatial visual features extracted from the encoder and meanwhile, the decoder needs the high-level semantic features more.

Specifically, we fuse the required features according to the number of vertices contained within a certain range. The more vertices in the range, the more complex and blurred the boundaries are, and the more low-level features are required. Similarly, the fewer vertices in the range, the simpler and clearer the boundaries are, and the more high-level features are required. Thus, we design an adaptive coefficient α to quantify the above metric, and $\alpha \in (0,1)$, the formula of α is defined as:

$$\alpha = \frac{N}{S} \tag{2}$$

where N denotes the number of vertices contained within the range of x pixels around the current predicted vertex position, and S denotes the area contained in the above range. x is a variable. In our experiment, we used $x = 8$.

As shown in Fig. 2. At time step *t*, the model first calculates the adaptive coefficient α based on the previous predicted vertex y_{t-1}, and passes it to the encoder. Then the encoder fuses the middle layer features. During the features fusion, the low-resolution features first up-sample by a factor of 2 and multiply by the coefficient $1-\alpha$ to determine how many high-level features are retained, and then fuse the corresponding feature maps from bottom-up pathway (which reduces the channel dimensions by 1×1 convolutional layer and multiplies coefficient α) by summation, and repeat this iteration until the features reach the required resolution. Finally, the fused features are passed to the decoder for next vertex prediction.

III. PIXRAY DATASET

To evaluate the proposed approach, we present an X-ray segmentation dataset named PIXray. The dataset consists of 2,623 X-ray images by manually labeling, in which 10 classes of 7,257 prohibited items have pixel-level ground-truth. Prohibited item categories include gun, knife, wrench, pliers, scissors, lighter, battery, bat, razor, and pressure vessel (PV). The detailed class distribution of the PIXray is shown in Table 1. Specifically, the ground-truth of each prohibited item is composed of the coordinates of vertices, and each image has a JSON file to save the ground-truth.

Table 1. The class distribution of the PIXray dataset.

The PIXray Dataset (2,623)				
Gun	Knife	Wrench	Pliers	Scissors
457	974	175	136	48
Lighter	Battery	Bat	Razor	PV
840	1,423	692	1,135	1,377

IV. EXPERIMENTAL RESULTS

In this section, we provide an extensive evaluation of our PANet model. Note that we assume a ground-truth box is provided by user around one prohibited item of interest in all our experiments. We first compare the quality of the generated polygons on our PIXray with several strong pixel-wise methods. Then, we evaluate the time required to label prohibited items using our approach. Finally, we test how many human clicks required to correct our results.

A. PIXray dataset

As is described above, the PIXray dataset consists of 2,623 X-ray images. We randomly select 2,100 X-ray images as our training set, 173 images as the validation, and the images remained as our test set.

B. Quantitative Results

To evaluate quality, we use the intersection over union (IoU) metric, calculated on a per-instance basis, and averaging across all instances. The results are shown in Table 2. We use the recently proposed DeepMask [9], SharpMask [10], and Polygon-RNN [5] as the state-of-the-art baselines. Note that the first two models are pixel-wise methods and thus the annotator cannot easily correct the deviant vertices at the end. To be fair, we only use our automatic mode (without using any fine-tuning) to compare with their models. From Table 2., we know that our approach outperforms the baselines in 9 out of 10 categories, as well as in the average across all classes. Our approach performs particularly well in gun, lighter, and razor, outperforming Polygon-RNN by 11.59%, 7.36%, and 13.19%, respectively.

The main advantage of our approach is that it allows the user to easily correct the deviant vertices at the end. Thus, we compare the average time of per instance required as well as average IoU of per instance achieved for all classes on the PIXray dataset by manual, automatic, and semi-automatic annotating (with manual fine-tuning). The results shown in Table 3. demonstrate that the annotations are accelerated

Table 2. Performance (IoU in %) on all the PIXray classes without any fine-tunning.

Model	Gun	Knife	Wrench	Pliers	Scissors	Lighter	Battery	Bat	Razor	PV	Mean
DeepMask	56.92	63.57	49.32	38.60	57.38	74.19	77.53	59.58	61.14	64.18	60.24
SharpMask	59.93	68.63	52.84	44.37	62.96	78.15	81.72	66.25	64.45	67.38	64.67
Polygon-RNN	61.37	71.01	54.79	47.67	68.65	78.82	82.48	**65.04**	66.52	69.73	66.61
Ours	**72.96**	**75.12**	**55.84**	**50.72**	**70.51**	**86.18**	**87.75**	63.93	**79.71**	**74.39**	**71.71**

Authorized licensed use limited to: HANGZHOU DIANZI UNIVERSITY. Downloaded on March 31,2022 at 03:34:20 UTC from IEEE Xplore. Restrictions apply.

Table 3. Performance on PIXray dataset with different annotating method.

Dataset	PIXray dataset		
Annotating method	Manual	Annotation mode	Semi-annotation mode
Time (s)	29.3	3.2	12.4
IoU (%)	100	71.7	92.3

Table 4. The number of clicks required in semi-automatic mode.

Dataset	PIXray dataset				
Classes	Gun	Knife	Wrench	Pliers	Scissors
Clicks	8.8	7.2	15.4	16.8	14.3
IoU (%)	91.2	92.7	90.2	90.6	91.6
Classes	Lighter	Battery	Bat	Razor	PV
Clicks	1.7	1.4	11.6	5.9	7.6
IoU (%)	95.0	96.8	90.5	92.8	91.6

by a factor of 2.4 through our approach, while achieving 92.3% accord in IoU with ground-truth.

In Table 4., we further report the IoU achieved as well as the average number of clicks required for each category. We can observe that by using our approach, only 9.1 clicks are required on average to obtain 92.3% IoU accord.

C. *Qualitative Results*

Fig. 3 shows example predictions obtained in automatic mode on PIXray dataset. We remind the reader that these results are obtained by exploiting GT bounding boxes without fine-tuning. Our model has the ability to segment prohibited items with a variety of shapes and sizes correctly. But the model still performs poorly on the parts with serious overlapping phenomenon. Further improving accuracy is subject of ongoing work.

V. Conclusion

In this paper, we proposed an approach based on PANet model to facilitate annotation of prohibited item instances. Our PANet model obtains annotation by predicting a polygon outlining of an object, and allows users to correct the deviant vertices at the end. There is a main advantage in our approach. It can adaptively fuse the required middle layer features in real-time according to characteristics of object boundaries during each inference. To evaluate our model, we present a high-quality X-ray segmentation dataset named PIXray. Experimental results demonstrate that, through our approach, the annotation process speeds up by a factor of 2.4 in all classes of PIXray, and this method achieves an accord of 92.3% with ground-truth in IoU.

Our future research will focus on two directions. (1) We look forward to designing a physical model based on the principle of X-ray image generation to completely eliminate the influence of overlapping phenomenon. (2) We will look for the relation between overlapping data and natural images (e.g., object occlusion) in order to extend these methods to a wider range.

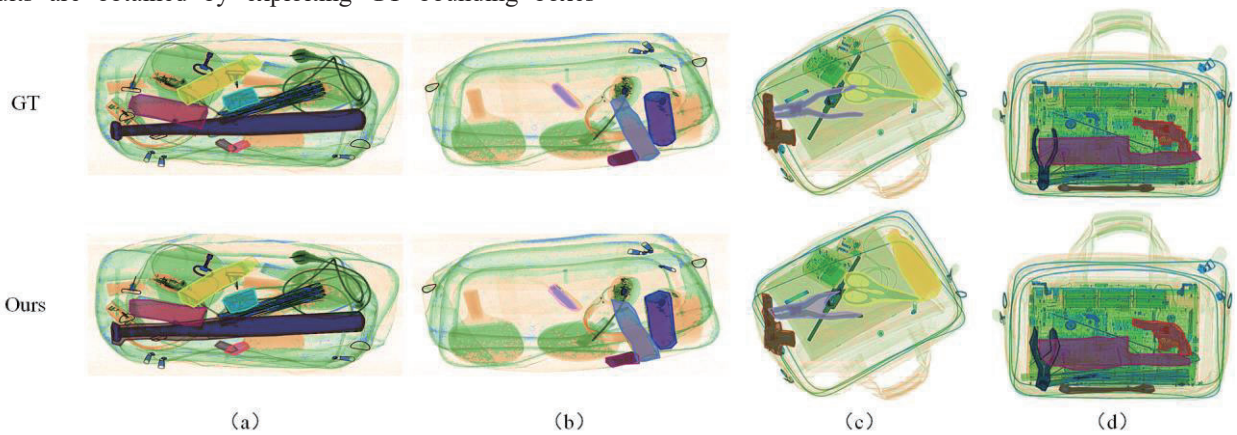

Fig. 3. Qualitative results on all the PIXray classes without any manual fine-tunning.

References

[1] D. Mery, E. Svec, M. Arias, V. Riffo, J. M. Saavedra and S. Banerjee, "Modern Computer Vision Techniques for X-Ray Testing in Baggage Inspection," in *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 4, pp. 682-692, April 2017, doi: 10.1109/TSMC.2016.2628381.

[2] H Wang, L Yang, Y Yu, "Contraband segmentation of compton back-scattering images based on CNN." Acta Electroniac Sinica, 2011, 39(3): 549-554.

[3] D. Mery, V. Riffo, U Zscherpel, *et al*., "GDXray: The Database of X-ray Images for Nondestructive Testing." *Journal of Nondestructive Evaluation* 34.4 (2015).

[4] C. Miao *et al*., "SIXray: A Large-Scale Security Inspection X-Ray Benchmark for Prohibited Item Discovery in Overlapping Images," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 2114-2123, doi: 10.1109/CVPR.2019.00222.

[5] L. Castrejón, K. Kundu, R. Urtasun and S. Fidler, "Annotating Object Instances with a Polygon-RNN," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 4485-4493, doi: 10.1109/CVPR.2017.477.

[6] T. Y. Lin, P. Dollar, R. Girshick, *et al*., "Feature Pyramid Network for Object Detection," *computer vision and pattern recognition* (2017): 936-944.

[7] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition." *computer vision and pattern recognition* (2014).

[8] X. Shi, Z. Chen, H. Wang, *et al*., "Convolutional LSTM Network: a machine learning approach for precipitation nowcasting." *neural information processing systems* (2015): 802-810.

[9] P. O. Pinheiro, R. Collobert, P. Dollar, *et al*., "Learning to segment object candidates."*neural information processing systems* (2015): 1990-1998.

[10] P. O. Pinheiro, T. Y. Lin, R. Collobert and P. Dollár, "Learning to refine object segments." In *European conference on computer vision* (pp. 75-91). Springer, Cham.

Authorized licensed use limited to: HANGZHOU DIANZI UNIVERSITY. Downloaded on March 31,2022 at 03:34:20 UTC from IEEE Xplore. Restrictions apply.