ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0056

Weakly Supervised Object Detection Based on Domain Transfer and Continuation Multiple Instance Learning

Xinpei Tian

School of Control and Computer Engineering, North China Electric Power University, Baoding, 071003, China

Abstract

This paper focuses on the problems of high object labeling cost during the detection process of strongly supervised object detection and lower accuracy of the detection results of weakly supervised object detection. This paper proposes an object detection model based on weakly supervised learning, which combines domain transfer technology and the Continuation Multiple Instance Learning algorithm. First, the CycleGAN domain transfer technology is used to convert the source domain image dataset into comic, watercolor, and clipart three types of dataset images similar to the object domain image. Then, the feature extraction of object domain image is accomplished by C-MIL algorithm, and obtains the instance label estimation through the two modules of instance selection and instance partition; finally, using the obtained domain transfer image and pseudo-label annotation image to fine-tune the pre-trained object detector in turn. The experimental results show that, compared with the detection results of the weakly supervised object detection of domain transfer, the improved model has a higher average accuracy of the detection results of the three datasets.

Keywords

Weakly-supervised learning; C-MIL; Domain transfer; Pseudo-label annotation.

1. INTRODUCTION

Object detection has attracted extensive attention and research from many scholars since the 20th century. The purpose of object detection is to determine whether the object is included in the image and find its location [1]. In common methods of object detection, the bounding box label of the object and the corresponding class annotations are required to optimize and train the parameters of the neural network, and gradually develop into the main model architecture in the aspect of object detection, and the accuracy of the detection results has been greatly improved. However, the accuracy of the object detection result of a strongly supervised learning will be affected to a higher degree by the accuracy of the bounding box labeling of the detected object, and the result of the bounding box annotations of the object is easily affected by subjective judgments, and it is extremely time-consuming and expensive to obtain a large number of accurate bounding box annotations in strongly supervised learning, so the applicability of strongly supervised learning is also limited.

In order to solve the above problems, a series of weakly supervised learning object detection methods that only need to provide the image-level labels of the object are proposed in the research process of object detection [2]. In weakly supervised learning, object detection can be achieved by only indicating whether a certain type of object is included in the image of the datasets, which can solve the problems of difficult and expensive cost of object detection labeling in strongly supervised learning, and it is easier to obtain labels compared with strongly supervised learning. A weakly supervised deep detection network architecture (WSDDN) is

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0056

presented in literature [3], which is superior to other weakly supervised object detection system models on the PASCAL VOC dataset. The literature [4] proposes a Sequential Label Propagation and Enhancement Networks model (Label-PEnet), which realizes object detection and instance segmentation in images by gradually converting image-level labels to pixel-level labels. A method of generating proposal clusters is proposed in the literature [5], and then the refined instance classifier is learned through iterative process to realize weakly supervised object detection of image-level annotations. In order to solve the problem that the model only focuses on the most distinguishing parts of the object in the image rather than the entire object, the literature [6] proposes an end-to-end trainable network model, an object detection module and a perceptual triplet loss are added to enhance the ability of weakly supervised learning to detect foreground objects. The literature [7] mainly transfers the source domain data in the source domain into images with instance labels through the domain transfer technology, that is the namely image transformation technology, or generates corresponding pseudo-label annotations to the object domain data through making pseudo-label. In the reference [8], a Continuation Multiple Instance Learning model is proposed to a continuation optimization multiple instance functions by smoothing loss functions to alleviate the non-convexity problem of MIL. However, most of the labeling annotations of weakly supervised learning are incomplete, inaccurate or inexact, so the accuracy of object detection algorithm based on weakly supervised learning still has a great room for improvement.

In response to the above problems, this paper proposes a method that combines domain transfer technology and continuation multiple instance learning algorithm to achieve object detection based on weakly supervised learning. First of all, it is necessary to use CycleGAN technology [9] to convert the source domain image into image data similar to the object domain image and with corresponding instance-level labels. Then, the pseudo-label annotations of the object domain image with high accuracy are obtained through the instance subset partition and instance selection. Finally, combining the results of the two techniques to fine-tune the pre-trained object detector to further improve the accuracy of weakly supervised learning object detection.

2. RELATED TECHNOLOGY

2.1. CycleGAN Technology

CycleGAN technology is a ring network model structure which is about deep learning extended by two mirror-symmetric generation against network GAN. Assuming that there are two sample spaces of X and Y, the purpose of domain transfer is to learn the mapping from X to Y. However, in order to avoid all X being mapped to the same Y, the two GANs in the CycleGAN model share two generators and each comes with a discriminator. As for the generator and discriminator in the CycleGAN model structure, the function of the generator is to convert the image x in the sample space X into the image y in the sample space Y, and the function of the discriminator is to distinguish whether the image generated by the generator is a real image.

Input the Clipart1k dataset image in the CycleGAN model as an example to obtain the network model architecture from source domain A to the object domain B after transformation, as shown in Figure 1. In the model, the source domain image data input_A is first input from the source domain A, and the input image is transferred to the first generator(Generator A2B) of the model, the corresponding transformation of the given source domain image input_A to the object domain image Generated_B from object domain B is completed through the Generator A2B. The transferred object domain image is then transmitted to the second generator(Generator B2A) of the model, through this generator the object domain image can be reconverted to image result Cyclic_A which is consistent with the input image input_A. The two discriminators

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0056

Discriminator A and Discriminator B corresponding to the generators in the architecture can respectively judge the authenticity of the images generated by the two generators.

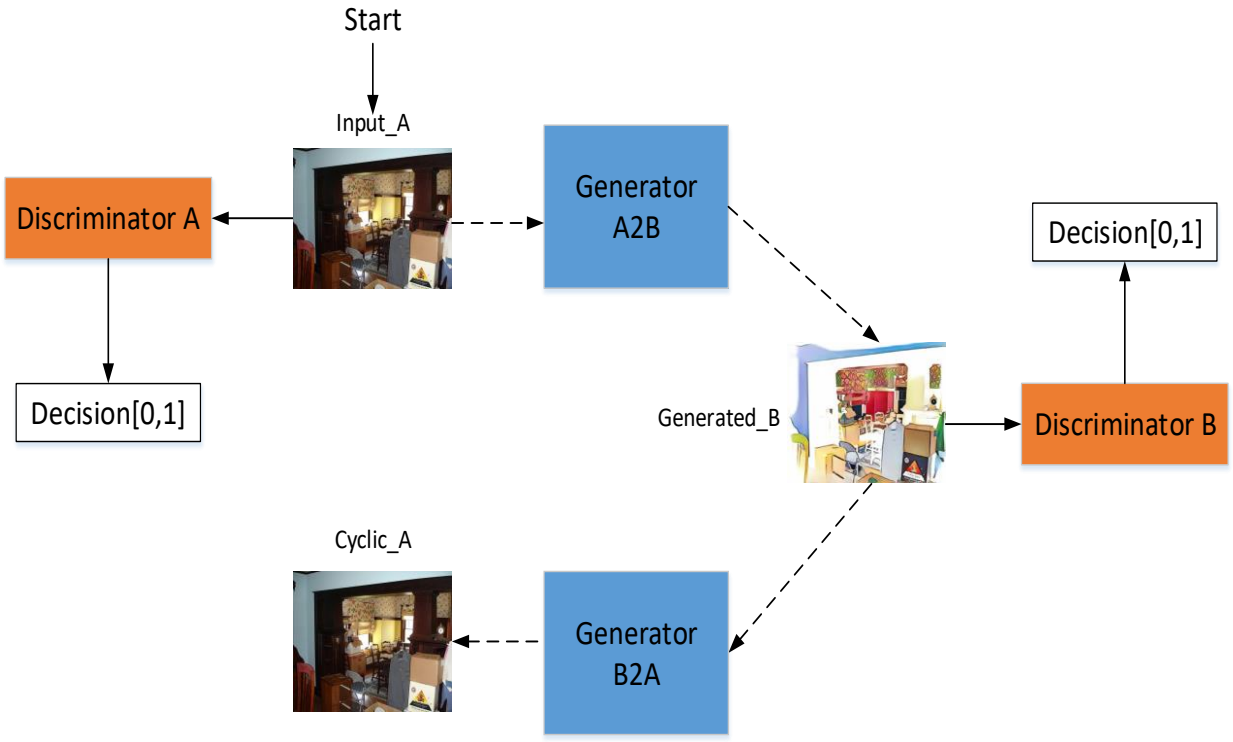

Figure 1. CycleGAN structure

The loss of the discriminator DY(Discriminator A) for the mapping of X->Y in the model as

$$L_{GAN}(G, D_Y, X, Y) = \mathrm{E}_{y \sim p_{data}(y)}\left[\log D_Y(y)\right] + \mathrm{E}_{x \sim p_{data}(x)}\left[\log\left(1 - D_Y(G(x))\right)\right] \tag{1}$$

And the loss of discriminator DX(Discriminator B) to the mapping Y->X is similar to formula (1).

The cyclic loss function between two generators G(Generator A2B) and F(Generator B2A) in the model is defined as

$$L_{cyc}(G, F) = \mathrm{E}_{x \sim p_{data}(x)}\left[\left\|F(G(x)) - x\right\|_1\right] + \mathrm{E}_{y \sim p_{data}(y)}\left[\left\|G(F(y)) - y\right\|_1\right] \tag{2}$$

this cyclic loss function is actually the L1 loss function.

Total loss of the CycleGAN model defined by Eq.3.

$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + \lambda L_{cyc}(G, F) \tag{3}$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0056

Which is the sum of the respective mapping losses of two discriminators and the cyclic losses between two generators.

2.2. Multiple Instance Learning

In the Multiple Instance Learning algorithm (MIL) [10], a bag in the training data is used as the object to make labels only divided into two categories which are positive and negative, instead of selecting a sample in the data as the object. One or more samples of data are combined together called a bag, and each bag has its own corresponding label. When the label corresponding to a bag are negative, then all the sample data in this bag is negative; when the label corresponding to a bag is positive, then the label of at least one sample data in this bag is positive. In this process, a classifier can be obtained through training and learning. This classifier can classify the newly input samples and give positive and negative labels of the sample data. Different from the traditional deep learning network model in which all instances and samples are one-to-one correspondence, in the MIL model, a sample can contain multiple instances. That is, there can be one-to-many correspondence between instances and samples. The literature [11] presents a Coupled Multiple Instance Detection Network model (C-MIDN), in which the location information of a pair of Multiple Instance Detection networks are further coupled to locate multiple objects and obtain a tighter bounding box. Their attempts to introduce a discriminative probability model with EM inference algorithm for the missing labels in multiple-instance and multiple-label learning in this paper [12], which improved the performance of missing label object detection.

Continuation multiple instance learning is a new learning algorithm developed from MIL. The difference from the traditional MIL algorithm is mainly in two parts. One is in the process of training, which is different from MIL regards one or more sample data as a bag, C-MIL regards each image as a bag, and each proposal in the image as an instance. Another one is that in the training process, the traditional MIL algorithm only selects to activates the most characteristic discriminative part of the object to learning, while C-MIL selects and learns a subset of instances, and activates all instances in the subset during the back-propagation procedures. There is a spatial correlation or class correlation between all the instances included in all the instance subsets in C-MIL.

3. METHODOLOGY

Figure 2 shows the weakly supervised object detection model that combines domain transfer technology and C-MIL proposed in this paper. It is mainly composed of three parts: the domain transfer training network, the C-MIL algorithm training network and fine-tuning object detector. Firstly, the object detector model is obtained by pre-training the source domain data through the pre-trained model. Then, the desired domain transfer image is obtained through domain transfer model transformation, and the object detector model is fine-tuned for the first time. Finally, C-MIL algorithm is used to obtain the pseudo-label annotations of object domain image through feature extraction and instance selection and partition, then further fine-tuning the object detector model is completed.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0056

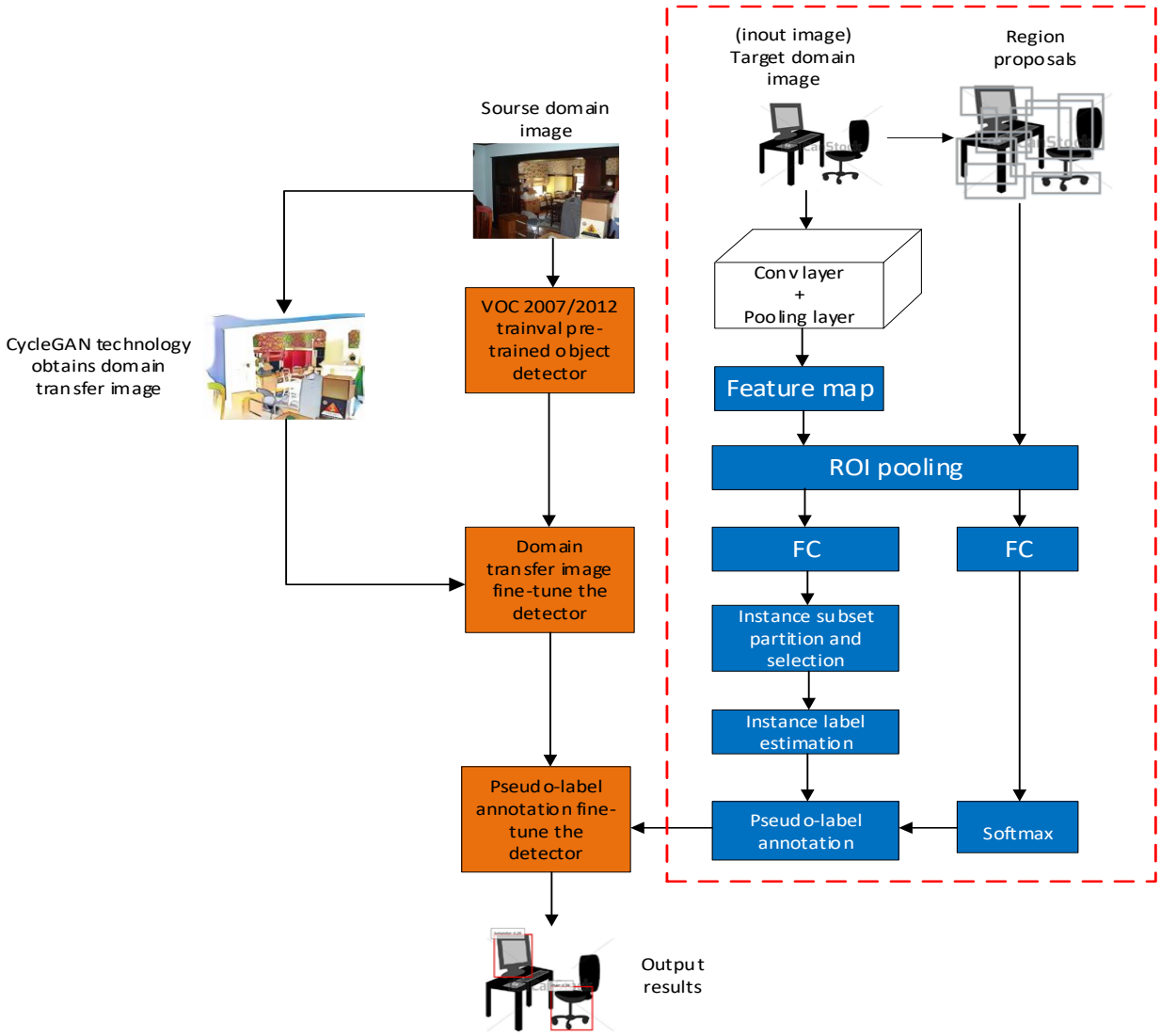

Figure 2. Weakly supervised object detection model

3.1. Domain Transfer Technology Training

Before fine-tuning the object detector by the results of domain transfer training, the VOC2007/2012-trainval pre-trained model provided by ChainerCV should be used to pre-train the image data of the source domain, so as to obtain an object detector with better local optimal solution as the centre detector of the whole experiment fine-tuning. Next, based on the Pascal voc 2007 dataset as the experiment source image, the domain transfer technology is used to complete the image conversion by capturing and processing the color difference and texture difference between the source domain image and the object domain image, thus, a domain transfer image dataset with instance-level annotations is generated that is similar to the object domain image. This paper uses CycleGAN technology, the main function of which is to use unpaired instances to learn the mapping function between the source domain and the object domain.

Table 1 shows examples of domain transfer images generated by the CycleGAN technology model that are similar to comic dataset images.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0056

Table 1. Detection results for the Comic2k dataset

Example	PASCAL VOC	Comic images
Example 1		
Example 2		
Example 3		

3.2. Pseudo-label Annotation

The content in the red virtual box in Figure 2 is the part of the object detection model that makes pseudo-label annotations to the image. The algorithm is applied to complete the pseudo-label annotations of object domain image data based on the pre-trained VGG16 convolutional neural network model.

When completing the feature extraction of the input object domain image, use the selective search method to extract region proposals from each image. The image is sequentially extracted through the convolutional layer and pooling layer of the backbone network for feature extraction, and the region proposals of the image are mapped to the last layer of the convolutional feature map of the backbone network, then each region proposal will be passed through the ROI pooling layer to generate a feature map with a fixed size, and feature expression for each proposal instance is generated. After further feature extraction through two fully connected layers of backbone network, is performed to obtain the feature expression of object domain image.The new instance subsets are established by continuation instance selection module and subsets of continuation instances are divided and selected according to continuation parameters.

In the process of subset selection of instances, there are two extreme phenomena. When the parameter is the largest the loss functions of the MIL instance selector in the instance subset partition and selection as

$$\mathrm{F}_f\left(B_i,\omega_f\right)=\max\left(0,1-y_i\max_j f\left(B_{i,j},\omega_f\right)\right) \tag{4}$$

When the parameter is the smallest,the loss function of the C-MIL instance selector is defined as

$$\mathrm{F}_f\left(B_i,\mathrm{B}_{i,J(\lambda)},\omega_f\right)=\max\left(0,1-y_i\max_{J(\lambda)} f\left(B_{i,J(\lambda)},\omega_f\right)\right) \tag{5}$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0056

$$f\left(B_{i,J(\lambda)},\omega_f\right)=\frac{1}{\left|B_{i,J(\lambda)}\right|}\sum_j f\left(B_{i,j},\omega_f\right) \tag{6}$$

Instance classification is performed on the selected continuation instance subsets, and the continuation instance label estimation of the instance subsets is obtained. Therefore, the pseudo-label bounding box with different scores of each instance contained in the object domain image instance subsets is obtained.

3.3. Instance Bounding Box Selection

After performing a series of continuation instance label estimation on a continuation instance subset, each instance subset will obtain a label estimation bounding box with a corresponding confidence score for each instance contained there.

The appropriate object bounding box confidence threshold is set to eliminate other candidate boxes with high confidence score in the instance subset, and the label estimation bounding box with high confidence score is obtained after screening, so as to be the pseudo-label annotation of the object domain image needed for further fine-tuning the object detection results during the experiment.

3.4. Output Test Results

The training results of the domain transfer image obtained by the domain transfer technology conversion are fine-tuned to the object detector obtained by the previous pre-trained. Using the continuation instance labels of the sample instances obtained through C-MIL, select the bounding box with the highest score in each instance subset as the pseudo-label of the instance class label. The final result is to further fine-tuning the object detector after the fine-tuning result of the domain transfer. Then the bounding box of object detection with higher detection accuracy is obtained, and the bounding box of object detection with higher confidence score was taken as the final detection result average precision of this experiment and output.

4. EXPERIMENT AND RESULT ANALYSIS

The operating system selected during the experiment in this paper is Ubuntu 16.04LTS, NVIDIA GeForce GTX 1080Ti was selected as the GPU, and the deep learning framework is Chainer.

4.1. Experiment Datasets

In this paper, the PASCAL VOC [13] dataset is used as the source domain image data, and the Comic2kare used as the object domain image data which is provided by Inoue [7]. The image dataset of the object domain also follows the PASCAL VOC dataset format and consists of JPEGImages, ImageSets/Main and Annotations. Among them, JPEGImages are the image of the object domain, ImageSets/Main is the ID list of the image and they are divided into two subsets of training set and test set according to train and test, Annotations are composed of pseudo-label annotations of the image set obtained through the continuation instance subset selection and the instance label estimation of the continuation multiple Instance Learning algorithm during the experiment.

All the comic images included in Comic2k datasets was collected from a large number of noise-labeled images in BAM [14], in the end, each image was selected 2000 images and used in this paper as the experiment datasets. These selecting Comic2k instances contained 6389 respectively. Comic2k datasets are divided into a training set and a test set according to the ratio of 1:1,and Comic2k include only 6 object domain classes, is a fraction of the 20 classes in the source domain dataset, and a subset of the classes in the source domain dataset.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0056

4.2. Experiment Results

In this experiment, mean Average Precision (mAP[%]) was adopted as the main evaluation index of experiment results, in which AP represents the accuracy of the detection results of each class in the dataset, while mAP represents the average detection result accuracy of all classes. Using the domain transfer algorithm and the continuation multiple instance learning algorithm as the benchmark method, the accuracy of the weakly supervised object detection model after the algorithm improvement is compared. This experiment mainly uses the detection results of WSDDN, CLNet [15], ADDA [16] and domain transfer algorithm (DT+PL) to complete the comparative analysis of experiment results. Among them, Baseline represents the accuracy of the detection results obtained by using the SSD300 algorithm to directly perform object detection experiments on the object domain image dataset. CLNet represents the object detection accuracy of ContextLocNet, a method based on weakly supervised learning object detection. WSDDN is the accuracy of the object detection result for the weakly supervised learning deep detection network. And ADDA is the accuracy of the experimental results of the object detection by unsupervised domain adaptation method.

Table 2 shows an example of the comparison of object detection results of Comic2k comic image dataset. According to Table 1, it can be seen that compared with the Cross-domian algorithm (DT+PL), the average accuracy of this algorithm in this paper for comic object detection has increased from 37.2% to 39.4%, which is an increase of 2.2%. Among them, the accuracy of the object detection results of the car and cat classes has been greatly improved, increasing by 4.7% and 3.1% respectively.

Table 2. Detection results for the Comic2k dataset

method	bike	bird	car	cat	dog	person	mAP
WSDDN	1.5	0.1	11.9	6.9	1.4	12.1	5.6
CLNet	0.0	0.0	2.0	4.7	1.2	14.9	3.8
ADDA	39.5	9.8	17.2	12.7	20.4	43.3	23.8
DT+PL	55.2	18.5	38.2	22.9	34.1	54.5	37.2
Our	57.2	20.3	42.9	26.0	33.7	56.4	39.4

Figure 3 shows an example graph of image detection results of Comic2K object domain dataset after algorithm model detection. According to the comparison between the average accuracy of object detection experiments under Comic2K data set and the detection results of object detection models with different weakly supervised learning algorithms, it can be concluded that the weakly supervised learning object detection model proposed in this paper has achieved a higher object detection accuracy. The model in this paper is more accurate for weakly supervised learning object detection with only image-level labels or incomplete, inaccurate and inaccurate labels.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0056

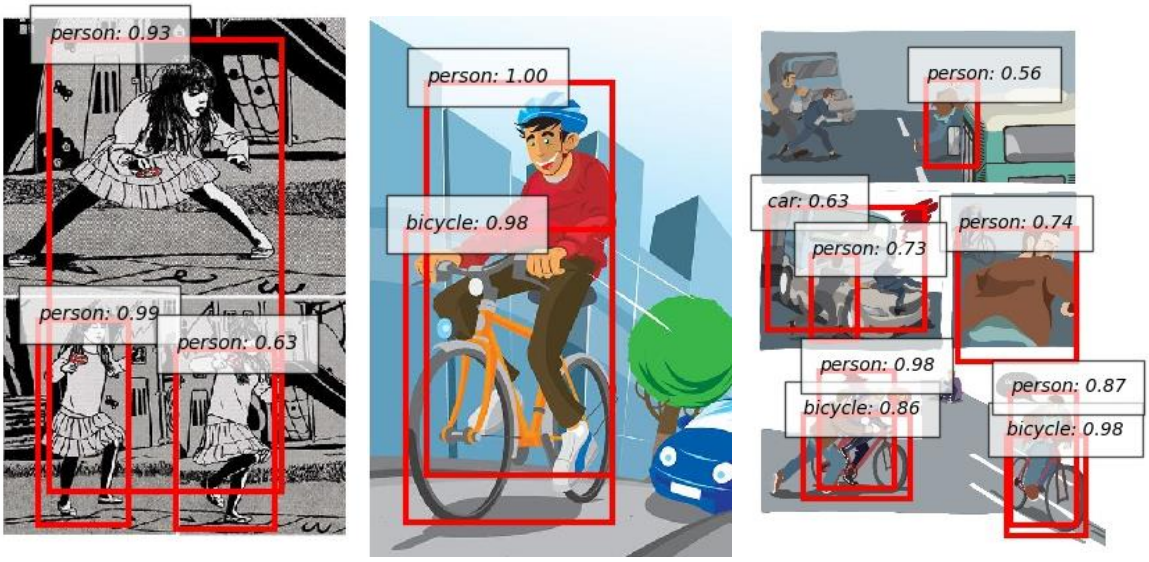

Figure 3. Comic image sample image of object detection results

5. CONCLUSION

In this paper, an object detection model based on weakly supervised learning is proposed, which can combine domain transfer model with continuation multiple instance learning algorithm, aiming at the problems of high cost to completing object labeling and low accuracy of detection results in weakly supervised learning object detection. Use the domain transfer image obtained by the domain transfer technology and the pseudo-label annotation image obtained by C-MIL to complete the two fine-tuning of the pre-trained object detector model, and the object detection bounding box with a higher confidence score was selected as the detection result output at last. The final detection results in this paper show that the object detection model based on weakly supervised learning has improved the accuracy of the object detection results after corresponding improvements.

REFERENCES

[1] Zhou Xiaolong, Chen Xiaojia, Chen Shengyong, LeiBangjun. Overview of object detection algorithms under weak supervised learning [J / OL]. Computerscience: 1-13.

[2] Xie Xiaowei, Shi Jianfang.Research on Multi-Target Image Detection of Weakly Supervised Convolutional Neural Networks [J] .Journal of Electronic Measurement and Instrument, 2019,33 (06): 31-37.

[3] Bilen H, Vedaldi A. Weakly Supervised Deep Detection Networks[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2016:2846-2854.

[4] W. Ge, W. Huang, S. Guo and M. Scott, "Label-PEnet: Sequential Label Propagation and Enhancement Networks for Weakly Supervised Instance Segmentation," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 3344-3353, doi: 10.1109/ICCV.2019.00344.

[5] P. Tang et al., "PCL: Proposal Cluster Learning for Weakly Supervised Object Detection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 1, pp. 176-191, 1 Jan. 2020, doi: 10.1109/TPAMI.2018.2876304.

[6] Y. Chen and W. H. Hsu, "Saliency Aware: Weakly Supervised Object Localization," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 1907-1911, doi: 10.1109/ICASSP.2019.8682756.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0056

[7] N. Inoue, R. Furuta, T. Yamasaki and K. Aizawa, "Cross-Domain Weakly-Supervised Object Detection Through Progressive Domain Adaptation," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 5001-5009.

[8] Fang Wan; Chang Liu; Wei Ke; Xiangyang Ji; Jianbin Jiao; Qixiang Ye. C-MIL: Continuation Multiple Instance Learning for Weakly Supervised Object Detection [arXiv]. 11 April 2019.

[9] J. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 2242-2251, doi: 10.1109/ICCV.2017.244.

[10] Li Yang, Wang Pu, Liu Yang, Liu Guojun, Wang Chunyu, Liu Xiaoyan, Guo Maozu. Weak surveillance real-time target detection based on saliency map [J / OL]. Acta automatica Sinica: 1-13

[11] G. Yan et al., "C-MIDN: Coupled Multiple Instance Detection Network With Segmentation Guidance for Weakly Supervised Object Detection," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 9833-9842, doi: 10.1109/ICCV.2019.00993.

[12] T. Nguyen and R. Raich, "Incomplete Label Multiple Instance Multiple Label Learning," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2020.3017456.

[13] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge[J]. International journal of computer vision, 2010, 88(2): 303-338.

[14] M. J. Wilber, C. Fang, H. Jin, A. Hertzmann, J. Collomosse and S. Belongie, "BAM! The Behance Artistic Media Dataset for Recognition Beyond Photography," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 1211-1220, doi: 10.1109/ICCV.2017.136.

[15] Kantorov V, Oquab M, Cho M, et al. Contextlocnet: Context-aware deep network models for weakly supervised localization[C]//European Conference on Computer Vision. Springer, Cham, 2016: 350-365.

[16] E. Tzeng, J. Hoffman, K. Saenko and T. Darrell, "Adversarial Discriminative Domain Adaptation," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 2962-2971, doi: 10.1109/CVPR.2017.316.