AI-based detection system of resident's behaviors in automatic trash sorting booths: a background computing-based solution

Gong Baojun[1], Zhang Wei[2], Shi Zhebin[2], Huang Qiucheng[2], Zhang Dongping[3]
[1]Zhejiang Xiaoniu Xunbao Environmental Technology Co., Ltd., Hangzhou, China
[2]Zhejiang Sci-Tech University, Hangzhou, China
[3]China Jiliang University, Hangzhou, China

2022 China Automation Congress (CAC) | 978-1-6654-6533-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/CAC57257.2022.10055608

In order to track the serious problem of garbage siege, four-categories garbage sorting systems have been increasingly installed and employed in urban communities. However, manual supervision of trash delivery behaviors is costly, inefficient, and unsustainable from the long term. AI-based automated garbage sorting booths based on computer vision technology become an attractive solution. Under the four categories sorting system, the sorting and treatment of perishable or organic waste belong to the most challenging part. For the delivery of perishable waste in automatic garbage sorting booths, the key of AI-based solution is to accurately identify residents' abnormal delivery behavior using object detection models. The composition of organic waste is extremely complicated, making it difficult to employ the object detection model in real applications. Therefore, this paper proposes a solution to use the YOLO model to identify non-organic objects such as garbage bags in the perishable barrels. The proposed solution in this study adopts the backend computing architecture and employs the multi-scale YOLOv4 model as the core, which can effectively identify the abnormal delivery of perishable waste. This system has been applied in dozens of intelligent waste delivery booths in Yiwu City, and has achieved satisfactory results.

Index Terms—**AI-based garbage sorting; object detection; YOLOv4 model; Darknet53.**

I. Introduction

At present, the four-category garbage classification system is widely adopted in urban communities in China. Established four-categories garbage sorting system divides domestic trash into organic or perishable garbage, recyclables, other garbage and hazardous materials. Due to the habits of Chinese residents, perishables usually contain a considerable amount of kitchen waste that contain high percent of water. This indicate that the subsequent treatment methods for perishables are significantly different from those of recyclables and other wastes, and should be treated separately. If non-perishable garbage such as plastic bags are delivered to the perishable trash buckets, it will cause trouble to the subsequent garbage disposal. In this case, abnormal trash delivery behaviors of residents can be effectively detected during the delivery process, thus generating reliable evidence for community administrative.

In order to cooperate with the four-category garbage classification system, a variety of intelligent garbage delivery booths have been designed and implemented in city communities. Some solutions have re-designed the booths as fully enclosed, which are different from the previous semi-open delivery booths. This enclosed booths usually indicate high maintenance costs in daily operation. Another problem that hardly be ignored is the convenience of garbage transfer. Old-fashioned trash buckets that are convenient to use have strong environmental adaptability to the daily operation as well as subsequent transportation of garbage. Therefore, another solution is designed with high similarity with the existing semi-open garbage delivery booths. By adding cameras and data transmission equipment, the semi-open intelligent delivery booth with AI models is trained to monitor the abnormal delivery behavior of residents[1], [2], [3], especially the delivery of non-organic objects to perishable buckets. To achieve accurate and reliable detection, the way to deploy AI models should be taken into consideration[4], [5].

With the rapid development of high-performance hardware such as GPU and the emergence of well-labeled image datasets, convolutional neural networks(CNN) models have developed rapidly[6], [7]. Szegedy et al. proposed GoogLeNet, which increased the depth of the network to 22 layers[8], [9]. Edge computing-based solutions need the support of intelligent terminals, which is capable of performing real-time object detection on the terminal side. The edge computing system sends the detection results instead of video streams back to the data center. The advantages of this solution lie in the reduced processing in the data center and improved efficiency. There is no need to configure high-speed VPNs dedicated line to transmit data. The disadvantage is that the processing power and space of the intelligence devices are limited. This indicates that the employed AI model should be relatively lightweight, and the accuracy in detecting abnormal objects is difficult to match with the traditional YOLO model. For the detection task of unapproved delivery behaviors s, a certain degree of missed detection is allowed, and the real-time requirements are not high. Therefore, the solution based on background computing is adopted in this paper.

The back-end computing based solution firstly needs to transmit the video stream back, and then perform frame-by-frame processing to accomplish object detection on photos at a specific time interval. When new violation events are detected, the event will be backtracked, and the corresponding videos and pictures will be saved as evidence of unapproved delivery. In this way, high-speed VPN devices can be configured to guarantee the stable transmission of videos. Furthermore, AI servers in the data center are employed to handle hundreds of intelligent grocery delivery booths. The advantage is that

Corresponding author: Zhang Wei (email: zhangweicse@foxmail.com).

it can use a simple camera to reliably identify illegal objects without relying on terminal smart devices. One disadvantage of backend computing-based solution is the relatively high computational burden, especially when the number of booths is large. It should be noted that daily operating costs of backend based solution is relatively low, considering the quantity of abnormal grocery delivery events.

A. *Object detection in garbage sorting booths*

You only look once (YOLO) model is regarded as a state-of-the-art, real-time object detection system. Prior detection systems repurpose classifiers to perform detection. The YOLO network divides the image into regions and predicts bounding boxes and probabilities for each region. These bounding boxes are weighted by the predicted probabilities. YOLO models have several advantages over classifier-based systems.For instance, YOLOv3 looks at the whole image at test time so its predictions are informed by global context in the image. It also makes predictions with a single network evaluation unlike systems like R-CNN which require thousands for a single image. In YOLOv3, the confidence and coordinates are predicted separately and a residual network called DarkNet-53 is employed to extract features.

1) *YOLOv4 models*

YOLOv4 network architecture consists of three sections[10], i.e. backbone, neck and detection head. CSP-Darknet53 that contain large receptive field and high input resolution is used as the backbone for YOLOv4 models. Larger receptive field contribute to an improved view of entire objects in an image as well as a better understanding about the contexts. In addition, high input resolution will be useful in detecting small sized objects. Hence, Darknet53 is regarded as a suitable backbone in recognizing multiple objects with different sizes in a single image. Network structure of the YOLOv4 model is illustrated in Fig.1.

In Fig.1, backbone network takes the form of Darknet53. The neck section consists of various bottom-up and top-down aggregation paths. This section aims to increase the receptive field in the network and separates out the most significant context features, with limited reduction of the network operation speed. Spatial pyramid pooling(SPP) blocks have been added as neck section over the Darknet53 backbone, while path aggregation network(PANet) is used as the method of parameter aggregation from different backbone levels. SPP block fuse multiscale fusion with pooling. In the sparse prediction part, detection head processes the aggregated features from the neck section and predicts the bounding boxes, classification scores.

It should be noted that pretraining mechanism play a crucial role in training a useful and robust YOLO model[11].

2) *Detection of abnormal objects in buckets for perishables*

For perishable trash buckets, the image extraction section consists of hourglass network, three types of heatmaps, embedding vector features, offset extraction network and other modules. Among these components, the hourglass networks are employed to extract the corner and center points of the objects, while the heat map module obtains the location maps of center points. Positions of the corner or center points in the heat maps are extracted by the offset extraction network. In this study, the embedding part aims to extract the feature vectors that match the corner points. These vectors corresponding to the corner points are computed by the embedded vector module. In order to explore the matching relationship of corner points, the similarity matrices are constructed by using the vectors. With the positions of center points, incorrect target frames are filtered to obtain the final frames of perishable trash buckets.

In Figure2, feature maps are extracted from images of the perishable buckets by the backbone network. These feature maps play the role of input to the branch networks of five scales. Candidate frames obtained by each branch network is mapped to the original image and then subjected to non-maximum suppression. If the candidate frames retained after suppression belong to a large target, then all convolution modules in the feature map in the large target branch will be trained. Similar processing procedures are performed for minor and medium targets respectively. At the end of the backbone network, spatial pyramid pooling is used to extract multi-scale local features.

3) *Detection of non-scheduled delivery behaviors*

During the non-scheduled garbage delivery time, garbage detection is performed on the behavior surveillance video installed on the garbage booths. When new garbage bag is detected around the garbage booth, historical monitoring events will be retrieved to detect pedestrians and garbage bags. When the bounding boxes of person and garbage bags become overlapped, it is considered that the residents' garbage bags appear near the garbage booth, and the similarity between the garbage bag in the pedestrian's hand and the garbage bag in the booth will be matched. If a new garbage delivery event is detected, it indicates that the garbage bag in the previous garbage booth was thrown by the pedestrian.

II. Experimental outcomes and analysis

The labeled training set contain 8 valid categories: *tong, shou, youtong, suliaozhipin, zhizhipin, mianzhipin, lvzhipin*, and *bumingwu* while the AI system designs 17 categories. This suggest that several categories are absent in the current system. Among valid categories, 'tong' refers to perishable barrels, blue, only green perishable barrels are marked, and gray trash cans are ignored; *'youtong'* refers to the barrels used for residents' delivery, and *'suliaozhipin'* refers to plastic products, which are the main illegal delivery items, and plastic bags are common; *'zhizhipin'* corresponds to paper products. Considering the significance of violations, the training sample set mainly selects items with color and large size such as express boxes; *'lvzhipin'* is a metal tool used by staff to sort violations, including tongs and hooks; *'bumingwu'* represents the small animals that move around the perishable barrels, mainly cats. The color and texture of cats are closer to the violations under the perishable camera, and there are fewer corresponding labeled samples.

For typical scenarios, detected outcomes of plastics, hands, papers, cat were demonstrated in Fig. 3.

It can be observed from Fig. 3 that hands and sleeves of residents around the zone of green buckets can be detected

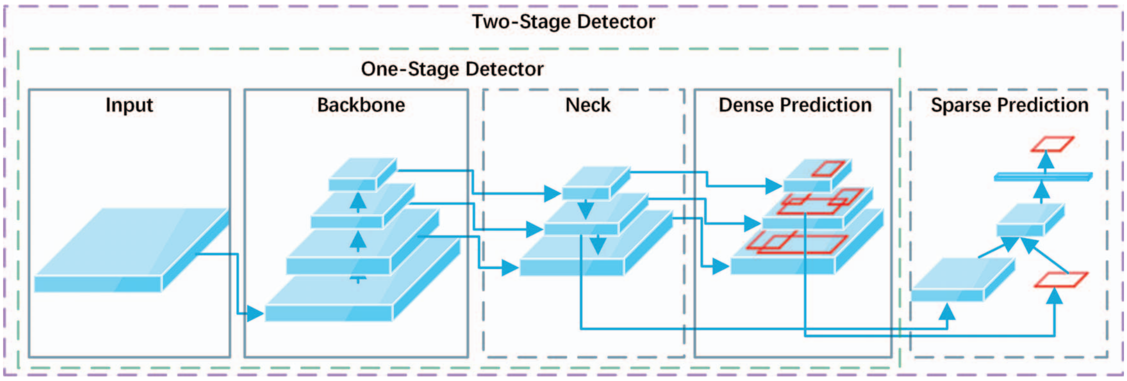


Fig. 1. Architecture of the YOLOv4 model. Backbone, neck, dense prediction are essential parts in one-stage detector.

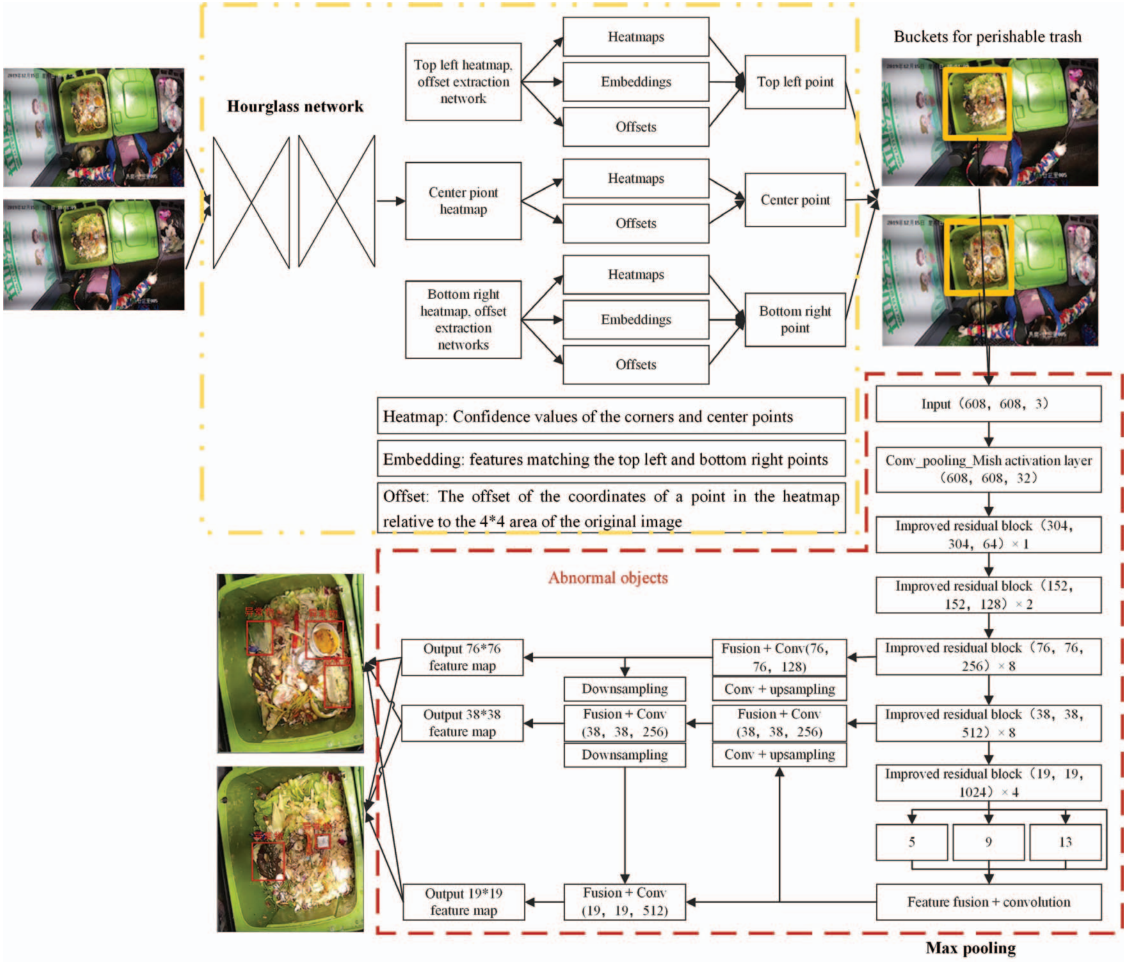
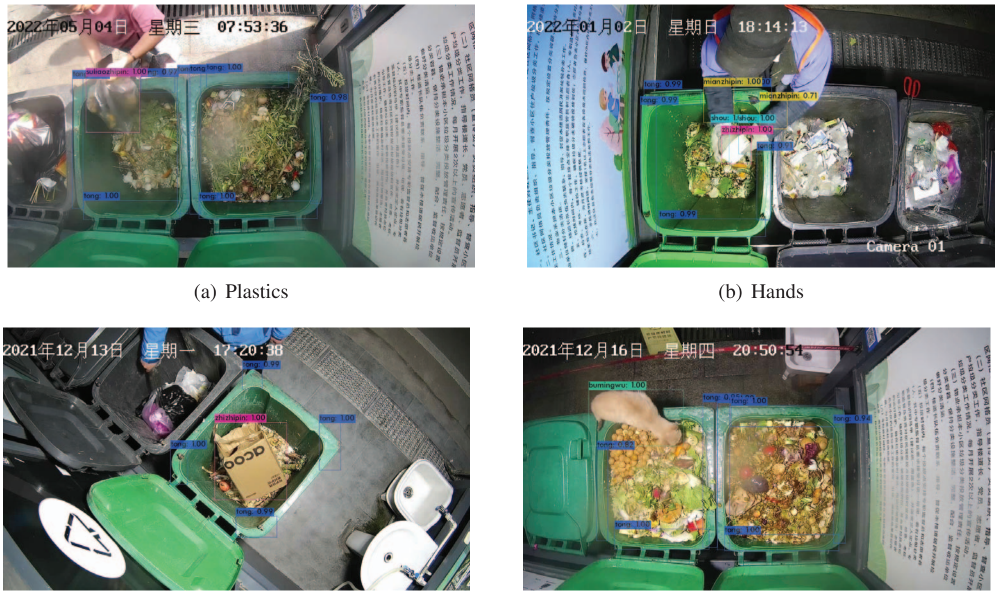

Hourglass network
Top left heatmap, offset extraction network
Center piont heatmap
Bottom right heatmap, offset extraction networks
Heatmaps
Embeddings
Offsets
Heatmaps
Offsets
Heatmaps
Embeddings
Offsets
Top left point
Center point
Bottom right point
Buckets for perishable trash
Heatmap: Confidence values of the corners and center points
Embedding: features matching the top left and bottom right points
Offset: The offset of the coordinates of a point in the heatmap relative to the 4*4 area of the original image
Input（608，608，3）
Conv_pooling_Mish activation layer（608，608，32）
Improved residual block（304，304，64）×1
Improved residual block（152，152，128）×2
Improved residual block（76，76，256）×8
Improved residual block（38，38，512）×8
Improved residual block（19，19，1024）×4
5
9
13
Feature fusion + convolution
Max pooling
Abnormal objects
Fusion + Conv(76，76，128)
Conv + upsampling
Fusion + Conv（38，38，256)
Conv + upsampling
Downsampling
Fusion + Conv（38，38，256)
Downsampling
Fusion + Conv（19，19，512)
Output 76*76 feature map
Output 38*38 feature map
Output 19*19 feature map

Fig. 2. Detection of abnormal objects in perishable trash buckets. Multi-scale objects in green buckets including plastic and paper products will be detected as the triggering condition of video recording.

’shou’ and ’mianzhipin’ respectively. These two categories indicate human activities are still on the way. Under such circumstance, abnormal events will not triggered since human activities are unfinished.

In the back-end computing solution, detection accuracy metrics of YOLOv4 models have been computed using an independent testing set which contain 100 samples that cover multiple booths. With collected samples, multiple versions of YOLOv4 have been trained and employed in the server. Evaluation metrics of baseline YOLOv2 and YOLOv4 models have been computed. These metrics include IOU, recall, AP for specific category as well as the average mAP indexes are listed in Table. II.

In this AI system, hands, buckets and plastics correspond to


2022年05月04日 星期三 07:53:36
2022年01月02日 星期日 18:14:13
Camera 01
2021年12月13日 星期一 17:20:38
2021年12月16日 星期四 20:50:54

(a) Plastics

(b) Hands

(c) Papers

(d) Cat

Fig. 3. Illustration of four typical objects that appear in the view of behavioral cameras. Among the names of detected objects, 'tong' denote the organic buckets that are colored with green, while 'suliaozhipin' represent the plastics waste.

Metrics Model	Size	IOU	Recall	AP for hands	AP for buckets	AP for plastics	mAP
Baseline	460	42.31%	45.11%	0.4003	0.4303	0.1624	0.0677
Phase I	678	46.28%	54.01%	0.5124	0.5138	0.1602	0.1114
Phase II	878	47.05%	54.09%	0.6168	0.5220	0.1970	0.1365
Phase III	1136	49.5%	56.56%	0.6175	0.5020	0.1966	0.1454
Phase IV	1328	51.86%	62.69%	0.5975	0.5154	0.1604	0.1645

the categories of 'shou', 'tong', and 'suliaozhipin' respectively. The reason of low mAP value is due to several categories are absent in current AI system, which correspond to zero AP value. From AP value for plastics, the YOLOv4 in model Phase III obtain the supreme index of 0.1966.

In Table. II, increasing sample size contribute to improved evaluation metrics including AP and Recall metrics. In addition, non-organic waste 'plastics' have multiple physical size and colors, leading to relative low detection accuracy. Similar situation exist for the non-organic category of paper products named 'zhizhipin'. YOLO model in this study mainly focus on brown cardboard box with high salience.

One interesting phenomenon is that non-organic waste in green buckets can be detecting, while the YOLO model seems ignore these waste in black buckets. This indicates the trained detection model focus on waste in and around the green-colored buckets which are designed for perishable waste only.

III. Conclusion

In this study, the multi-scale YOLO model is effective in identify the abnormal objects including plastic and papers in the perishable bucket area, thereby realizing the supervision of residents' garbage delivery behavior. The back-end computing solution used in this study identifies abnormal delivery events in perishable buckets by transmitting video streams and analyzing images. The AI system analyzes the delivery behavior of residents by judging the new incidents of illegal objects, and retrospect the video to save it as evidence.

The YOLOv4 based AI system aims to find abnormal objects in perishable barrels. The performance is affected by a variety of factors in practical applications, which may cause false triggering in certain circumstances. Changes in lighting conditions, activities of people and small animals, strong light reflection, and the existence of shadows disturb the AI recognition system to varying degrees. In order to improve the performance and robustness of AI systems, it is necessary to improve the detection accuracy.

In the winter scene, the texture and luster of gloves and sleeves contrasted with the perishable barrel background during the activities of the staff and residents in the guard box, so that the AI model judged them as illegal objects such as plastic products, which frequently triggered the reporting of violation events.The problem that kitchen waste and sleeves are frequently mis-detected as plastics become prominent under this circumstance. In this study, the hands of residents in the perishable barrel area, the sleeves of the staff, and the small animals moving around the perishable barrels were identified, and used as filter conditions to reduce false positives. After training with the YOLOv4 model, it can effectively identify categories such as hands and sleeves wearing gloves, and effectively filter the adverse interference of light source changes and personnel activities, so that the AI recognition system can adapt to summer and winter occasions.

The YOLOv4 model deployed in the Linux C environment adopts the .weights format, with a size of about 245MB. In this study, it takes up about 2400MB of memory to load into the GPU, which has certain requirements for hardware computing power. This indicate that models of this scale cannot be deployed on mobile terminals, such as cameras. If the front-end intelligence (edge computing) deployment method is adopted, it is necessary to perform research on light-weight models. Under the premise of not significantly reducing the model recognition performance, control the model scale and reduce the dependence on computing power. Relevant research on lightweight detection models, including the deployment and application of EfficientNet and MobileNet models in intelligent garbage sorting booths, will help to achieve rapid mobile detection and reduce the computational pressure of background processing.

ACKNOWLEDGMENT

The authors declare no conflict of interest. This study is funded by Zhejiang Provincial Key RD Program with project number of 2021C03192.

REFERENCES

[1] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, “Learning transferable architectures for scalable image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8697–8710, 2018.

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[4] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[5] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[6] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, *et al.*, “Recent advances in convolutional neural networks,” *Pattern recognition*, vol. 77, pp. 354–377, 2018.

[7] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, “A survey of convolutional neural networks: analysis, applications, and prospects,” *IEEE transactions on neural networks and learning systems*, 2021.

[8] M. Al-Qizwini, I. Barjasteh, H. Al-Qassab, and H. Radha, “Deep learning algorithm for autonomous driving using googlenet,” in *2017 IEEE Intelligent Vehicles Symposium (IV)*, pp. 89–96, IEEE, 2017.

[9] P. Ballester and R. M. Araujo, “On the performance of googlenet and alexnet applied to sketches,” in *Thirtieth AAAI conference on artificial intelligence*, 2016.

[10] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “Yolov4: Optimal speed and accuracy of object detection,” *arXiv preprint arXiv:2004.10934*, 2020.

[11] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*, pp. 740–755, Springer, 2014.