A novel detection fusion network for solid waste sorting

Yan Yu, Siyu Zou and Kejie Yin

Abstract
Vision-based object detection technology plays a very important role in the field of computer vision. It is widely used in many machine vision applications. However, in the specific application scenarios, like a solid waste sorting system, it is very difficult to obtain good accuracy due to the color information of objects that is badly damaged. In this work, we propose a novel multimodal convolutional neural network method for RGB-D solid waste object detection. The depth information is introduced as the new modal to improve the object detection performance. Our method fuses two individual features in multiple scales, which forms an end-to-end network. We evaluate our method on the self-constructed solid waste data set. In comparison with single modal detection and other popular cross modal fusion neural networks, our method achieves remarkable results with high validity, reliability, and real-time detection speed.

Keywords
Object detection, solid waste, RGB-D information, cross modal, real-time detection

Date received: 24 March 2020; accepted: 27 May 2020

Introduction
Due to the constructions, renovations, and demolition, massive amounts of construction and demolition (C&D) wastes are generated around the world. Most of these C&D wastes comprise multiple economically valuable materials, such as concrete, bricks, and wood, that can be used in manufacturing new products, construction materials, or in energy production. Vision-based solid waste sorting systems are a fundamental module in the C&D waste recycling industry. As shown in Figure 1, a vision-based solid waste sorting system consists of a server, vision sensors, industrial robot, rotational speedometer, and so on. During the sorting process, object recognition is one of the key technologies of waste sorting system.

In recent years, the use of convolutional neural networks (CNNs) provides excellent image detection and recognition performance and becomes very popular. The CNN-based methods can be divided into two categories: two-stage detection methods and one-stage detection methods. Two-stage methods such as Faster Region-based CNNs (R-CNNs) use a region proposal network to generate regions of interests in the first stage and then send the region proposals down the pipeline for object classification and bounding-box regression. The methods achieve high accuracy rates but are typically slower. One-stage methods such as You Only Look Once V3 and Single Shot multibox Detector (SSD) focus on how to accelerate the algorithm. Therefore, the coarse-to-fine process with the region proposal is removed. In these approaches, the bounding box and classification information are synchronously regressed
in the output layer. Such methods achieve lower accuracy rates but are much faster than two-stage object detectors.

All of these approaches assume the availability of large annotated data sets and clear RGB images. At present, there are few publicly labeled data sets of solid waste. In addition, in the process of construction, demolition, and transportation, solid wastes are usually badly irregular and covered with lots of dust. Therefore, the color information of the images will be seriously lost, and the accuracy of recognition will be greatly degraded. Due to these two reasons, the mainstream CNN-based detection methods are unable to achieve the perfect accuracy of solid waste detection.

Some recent work has been dedicated to improve the object detection performance by taking depth information. Compared with RGB information, depth information is less susceptible to the dust coverage, color loss, illumination, or other factors. Depth images are invariant to lighting and allow better background separation. It is obvious that combining the depth information to assist RGB inputs for object detection is a better choice. Different methods have been developed including deep features pooling, dense feature, multi-scale fusion, and so on.

In this article, we propose a deep multi-scale fusion cross modal RGB-D object detection network. Through the three-stage training method of supervision transfer training, it can solve the problem of under-fitting on small training data set. Even for the degraded images with very poor RGB information, our model can provide a high detection accuracy. Meanwhile, a real-time detection is also achieved.

To summarize, our main contributions are as follows:

- Our detection network fuses depth and RGB features in the mid-level by several multi-scale fusion layers. Compared with single modal networks, the accuracy is greatly improved by combining the depth feature.
- We propose a training strategy of fusion net. We freeze the weights of parts that have been well trained, and the training process can concentrate on the parts which after combination. Our method has a significant advantage on runtime and achieves the same as or exceeds the accuracy of other popular methods.
- Experimental results of the proposed methods are integrated into a real-life robotic system.

**Related work**

With the rapid development of affordable depth sensors and their diverse application scenarios, researchers have made great progress in the field of RGB-D object detection. To utilize the depth information and adopt the multimodal RGB-D as the input of object detection, the following two main problems need to be addressed: (1) Due to lack of large-scale depth image data set, how to obtain the input model of depth image through cross modal transfer learning and (2) how to well integrate the RGB model with the depth model. We will discuss related work about these two issues in the following part.
Supervision transfer learning

According to the literature and other related literature, several results have been achieved according to the transfer training process, and transfer learning has also been successfully applied in some fields. As a machine learning method, transfer learning can reuse a model developed for one task in another different task, and this model can be used as the starting point of the model for another task. Depth image is not like the RGB image which has big annotated data sets such as ImageNet, and this is also one of the main problems that cause the neural network model unable to conveniently utilize the depth information. When there is no large-scale annotated data set, during training of the object detection model of a depth image, the model will encounter severe under-fitting due to lack of data. Therefore, the transfer learning method can be employed to obtain the depth model through transfer training of the RGB model.

The RGB information and depth information are two kinds of modality information with different expressions. The pretraining model can be obtained by training the annotated RGB image, and then, this model can be used to conduct transfer training of corresponding paired depth image model in the data set. The transfer training from RGB image to depth image was proposed by Gupta et al., and they also provided experimental proof. However, this experiment can only achieve satisfactory results in the network built by them. And due to low speed, it cannot be used in real application scenario.

In our article, cross modal transfer training is conducted in multi-scale single modal network similar to SSD, and the model is significantly different from the model in the literature. Compared with the model proposed by Gupta et al., a new neural network layer is added to our model, and the model structure is also significantly improved. It can achieve real-time detection, and it has been used in real solid waste classification and detection system.

Fusion detection model based on the RGB-D data

There are many methods which can be used to process the RGB-D data. Deep learning methods (such as CNN) have been introduced to deal with RGB-D data. Gupta et al. introduced an RGB-D object detection method which only utilizes the CNN feature extraction once. Their method is based on Fast R-CNN, which uses AlexNet as the framework of feature extraction frame, the intermediate layer fusion (the FC6 layer of AlexNet) method is employed to obtain the CNN features of depth image and RGB, and the Softmax function is used to obtain the class score of proposal region. Although this method has good detection precision, the candidate region algorithms and the excessively FC layers of model cause it unable to provide real-time performance.

The fusion object detection method based on Faster R-CNN proposed by Chen et al. is an end-to-end model. They proposed the two-stream fusion object detection network model, in which the RGB and Depth information are input through two feature extraction modules. In the ROI Pooling layer, the two modules are fused after the FC layer, the fusion layer adopts the concatenation layer, and the detection result is obtained based on the fused CNN features. The network framework is VGG, and they adopt the anchor box design in the Faster R-CNN. The network does not use the steps of candidate frame algorithm beyond model. Therefore, by comparing with other methods, their method can provide better precision and faster speed, which can process one image in 1 s. These common object detection methods for RGB-D image generally try to improve precision, while few have the direct requirement for speed, so these methods cannot run in real time. Under the application scenario with higher requirement of speed, such as human-computer interaction, these methods are not applicable.

According to analysis of common models, it is found that the reasons causing slow running speed include using the FC layer and relatively more parameters and using the algorithm backbone network with high time-consuming. Some other parameters are supplemented in the algorithm, such as the SVM algorithm, which slow down the overall speed of algorithm. All these methods include time-consuming components, and these components may be not necessary to achieve great performance, which can be replaced by the model with balanced speed and performance to provide real-time speed.

Therefore, our proposed method focuses on the balance between speed and precision. The multimodality information is utilized to build the object detection method based on RGB-D fusion network. Our method can alleviate the problem of degraded color information of RGB image. In the meantime, attention is also paid to the running speed to adapt to the real application scenario with higher real-time requirement.

Method

We build the fusion network for object detection based on the following structures. To satisfy the requirement of detection precision, several strategies are adopted in our system, including data enhancement, which maintain consistent input and parallel prediction results based on multi-scale features; to satisfy the real-time detection requirement, the full convolution network structure and one-stage network are employed to guarantee the running speed. The network training procedure is shown in Figure 2. The realization method for entire network is introduced with the following four steps.
Cross modal transfer learning

Hinton et al. proposed the idea of extracting knowledge from the neural network and explained the distillation process in the literature.\(^\text{14}\) Gupta et al.\(^\text{10}\) extended their method. While our cross modal multi-scale deep fusion model is based on the method of Gupta et al.\(^\text{10}\), we expand the method to the SSD architecture.

Let us assume we have two modalities, one has pictures of annotated data set while the other does not. We would like to learn a rich representation of the latter modality. In this work, the former modality is RGB pictures, we denote it as \(L\), which has big data sets with annotations from that we can learn a rich representation. We use a neural network model to express the layered rich features of modal \(L\), and we assume this image representation as \(\phi = \{ \phi^i; \forall i \in \{1, \ldots, \#i\} \}\), where \(\phi^i\) is the \(i\)th layer representation which trained from annotated data set, \(\#i\) is the number of layers of the neural network, and \(W_u^{[1, \ldots, \#i]}\) is the parameters of the neural network. The latter modality is depth pictures. We assume this modality as \(U\) which without labeled data and denote an appropriate neural network as a rich representation of \(U\). \(U\) is layered as \(\psi = \{ \psi^i; \forall i \in \{1, \ldots, \#u\} \}\), where \(\psi^i\) is the \(i\)th layer representation, \(\#u\) is the number of layers of the neural network, and \(W_u^{[1, \ldots, \#i]}\) is the parameters of the neural network. Let us assume that the data set from two modalities have been paired and denote the unannotated data set as \(P_{1u}\). By match map representations \(\phi^i\) and \(\psi^i\) with paired maps \(I_u\) and \(I_t\) of modalities \(i \in \{1, \ldots, \#t\}\) refers to the partial expression of RGB network, \(I_u\) and \(I_t\) represent the images in modality \(U\) and \(L\), respectively, we can learn a rich representation of \(U\). To make the two representations have the same dimensions, we use a conversion function \(t\) and propose a loss function\(^\text{17}\)

\[
W_u^{[1, \ldots, \#i]} \sum_{(I_u, I_t) \in P_{1u}} f (t(\psi^i(I_u)), \phi^i(I_t))
\]

We transfer this theory to our one-stage network. Besides general convolutional pool layers, our network consists of the concatenate layer, the permute layer, the flatten layer, and the prior box layer. Let us analyze these layers one by one. The concatenate layer merges all features of multi-scale images. It can be deemed as the combination operation of matrix. The permute layer alters the sequence of the data dimensions. This is equivalent to multiplying by an identity matrix with row swapping. The flatten layer combines multidimensional matrices into one dimension. The prior box is dealing with bounding box which has no effect on the rich representation. All these layers can be unified into a conversion function \(s\), and the loss function of our one-stage network can be defined as follows\(^\text{25}\)

**Encoding depth images for feature learning**

During the fusion of features, it will affect the transfer training if the depth data and RGB data have inconsistent dimensionality. If two models have different input dimensionalities, the model parameters cannot correspond to each other one by one, which makes it impossible to realize direct transfer learning. Furthermore, the inputs with different dimensionalities will also result in inconsistent dimensionalities of feature image, which will be difficult to conduct model fusion.

We propose to encode the depth image with three channels, the depth image have been represented as HHA format (HHA encodes horizontal disparity, height above ground, and angle with gravity\(^\text{11}\)) which have three dimensions as RGB image. The values of data in three dimensions are mapped to the range of 0–255 and constitute the structure similar to the RGB data.\(^\text{24}\) The HHA coding is used to maintain consistent parameter dimensionality and numerical range of network before and after transfer learning, which can reduce the difference of network layer in the pretrained model and to-be-trained model, so that the transfer learning on small data can achieve optimal results.

**Figure 2.** The training process of multi-scale fusion network.
\[ W_{\mu}^{1_m} \sum_{\langle \tilde{I}_1, \tilde{I}_2 \rangle \in P_{\tilde{I}, \tilde{J}}} f (s (t (\psi_\mu (\tilde{I}_u))), \phi^f (\tilde{I}_1)) \] (2)

Through these analysis, supervision transfer can be used in our one-stage network, which is also proved by our experiments in the fourth section.

Implementation of transfer learning

The experimental results show that by utilizing the representation of labeled RGB data and the representation of unlabeled depth data obtained through transfer learning and making improvement through fine-tuning, it can effectively improve the detection performance as supplementation to the expression of RGB modal features. The model training process consists of three stages, and we will discuss the training method in the following steps in detail.

Stage I: Training the object detection network for RGB image. Both the training set and the test set use the RGB images from the self-constructed solid waste data set. The pretrained model is used to obtain the object detection model of RGB image through transfer training, and this is also the mainstream same modality transfer training process based on the object detection method of RGB image. The model obtained through training in this step can detect the object of RGB image in the solid waste data set.

Stage II: Training the object detection network for depth image. The RGB network model in stage I is obtained by training the labeled RGB image data set, but there is no pretrained model based on labeled depth images. Therefore, we utilize the RGB image network which has completed training to initialize the depth image network model and use the weight of RGB network to initialize and process the depth image detection network. Through this step, the depth data network can obtain advanced abstract semantic of object. We use the depth pictures which corresponding to the RGB pictures from the training set and the test set in step I. The depth image is in the HHA format, which has the same dimensionality as the RGB image. Through transfer training, the cross modal supervision of RGB object detection network can be transferred to the depth image detection network. The model trained in this step can detect the object in depth image separately.

Stage III: Training the RGB-D multimodal data fusion network. In this step, the pretrained models include the two object detection networks obtained in previous two steps. The training set and the test set use the RGB images and corresponding depth images from the solid waste data set established by us, and the final fusion network is obtained after transfer training. It should be noted that during training, the pixels in RGB image should correspond to the pixels in the depth image one by one. When two images are input into the fusion network, their data enhancement methods should also be consistent.

The three-stage training method can solve the problem that small-scale data set cannot completely constrain the network convergence. With this method, the depth image can help the RGB image to improve the object detection precision, and especially under degradation of RGB image, it can become the key information used to detect the object. In this method, the networks in the first two steps are based on existing network models; the third is proposed by us, and we will introduce it in detail in the next section.

Fusion network

The design of our multimodal fusion network is shown in Figure 3, which adopts the deep fusion method. The RGB and depth parts adopt normal SSD model with VGG architecture. The fusion net employs multiple scale layers to present features. The features that need to be fused come from the conv4_3, fc7, conv6_2, conv7_2, conv8_2, and conv9_2 of VGG. Among them, fc7 is the improved convolutional layer, and conv6_2 is obtained after two convolutions of conv5_3 of original VGG. Feature fusion is conducted to the same layer in two modalities. The fusion network predicts the features obtained through fusion, and the confidence of object class and the predicted value of frame can be obtained.

We adopt the high-level fusion scheme, and the eltsum layer is used in the fusion strategy. Because the eltsum layer has smaller influence on the network (which does not affect the numbers of network parameters), it has more advantages for training of small-scale data set. We place an eltsum layer after the convolution layers which need to be fused from conv4_3 to conv9_2 in two models and fuse the rich feature expression on corresponding layer.

Optimization training

The multi-scale deep fusion will generate certain training fluctuations, so it requires optimally training. During training of fusion network in stage III, we need to freeze some parameters, so that the neural network will not update the weights of these layers during back propagation. During training of fusion network, we freeze the feature extraction layers except fusion layer. We use a sufficiently small learning rate to transfer training the fusion layer. It can accelerate the training speed for stage III and improve the training stability. Without freezing the parameters, the training precision will significantly decline during the initial phase of fusion. Although the final training will be gradually improved to the level of stable precision, the experimental results show that it has poorer training performance than that the training method which freezes the parameters. If related parameters are not frozen, it will result in unstable training results and decline of final precision, so it is more effective to train the fusion model of small-scale data by freezing the parameters in stage III.
The experimental results show that we can obtain the feature representation of unlabeled depth data through transfer learning from labeled RGB data. After improvement by fine-tuning, it can effectively improve the overall detection performance as supplementation to the feature expression in source modality.

**Experiment**

In our experiments, we implement our model based on SSD, an open-source framework for traditional RGB object detection built on the Caffe platform.

**Data set and parameter setting**

**Data set and setup.** All solid waste images were captured using a Kinect v1 RGB-D camera. We collect 3921 images, train set includes 2921 images and test set has 1000 images. For all our experiments, we use annotations of the seven categories: cloth, bottle, stone, wood, plastic bag, box, and can. The image size is 336 × 263. All experiments are all worked at one single TITAN V with one GPU Intel 7700K.

All trains are limited to 70K iterations, and the learning rate is set to 0.001, except the last fusion net is trained with 0.0001 learning rate. Our fusion network uses VGG16 as feature extraction layers. We evaluate the accuracy with mean average precision (mAP) which is the common evaluation criteria of the object recognition and the information retrieval. We use the running time of one forward operation of the network as the metric of the speed performance of our network.

**Comparison with single modal methods**

The multimodal fusion object recognition method can provide some missing information in RGB images and reduce the interference caused by complex background. We compare the accuracy of the multi-scale feature fusion network with the mainstream single modal object detection network. The experimental results show that the cross modal multi-scale fusion net has higher mAP value than the single modal object recognition net.

Table 1 presents the mAP values of each category on the self-constructed solid waste data set. The methods used for comparison include RGB object recognition model based on SSD, depth object recognition method after cross modal transfer training, and fusion object recognition model. Figure 4 shows that the precision–recall curves of our method almost completely wrap the others. These results demonstrate that our net can take

| Method                  | Cloth | Bottle | Stone | Wood |
|-------------------------|-------|--------|-------|------|
| SSD on RGB              | 75.1  | 78.4   | 76.8  | 86.5 |
| SSD on depth            | 62.9  | 79.9   | 63.5  | 67.8 |
| Fusion network          | 82.5  | 84.1   | 78.9  | 83.6 |

| Method                  | Pbag  | Box    | Can   | mAP  |
|-------------------------|-------|--------|-------|------|
| SSD on RGB              | 72.4  | 62.2   | 54.2  | 72.2 |
| SSD on depth            | 68.1  | 73.5   | 42.8  | 65.5 |
| Fusion network          | 79.3  | 78.8   | 61.1  | 78.3 |

SSD: Single Shot multibox Detector; mAP: mean average precision.
**Figure 4.** Precision–recall curves of different classes: SSD on single modal RGB (dotted line in blue), the depth modal supervision transferred from RGB (dotted line in green), and our fusion net (solid line in red). SSD: Single Shot multibox Detector.

**Figure 5.** Detection results of single modal and fusion modal: The first column is the RGB object recognition based on SSD, the second one is the object recognition model based on depth after cross modal transfer training, and the third is the fusion model object recognition method. SSD: Single Shot multibox Detector.
advantage of both RGB and depth features to provide more accurate recognition.

Figure 5 shows the recognition results of the objects images with negative samples. The first column is the RGB object recognition based on SSD, the second one is the object recognition model based on depth after cross-modal transfer training, and the third is the fusion model object recognition method. It can be seen from the figure that the single modality detection modal based on RGB or depth has some error detections and leak detections, and the fusion network model has a better detection effect.

Table 2. Accuracy of cross modal architectures.

| Method                          | RGB architecture | Depth architecture | mAP (%) |
|--------------------------------|------------------|--------------------|---------|
| Gupta et al.’s fusion net with AlexNet RGB architecture | AlexNet          | AlexNet            | 75.5    |
| Gupta et al.’s fusion net with VGG RGB architecture    | VGG              | AlexNet            | 78.3    |
| Our fusion net                               | VGG              | VGG                | 78.3    |

mAP: mean average precision.

Table 3. Time cost of cross modal architectures.

| Method                          | Hardware | Time (ms) |
|--------------------------------|----------|-----------|
| Gupta et al.’s fusion net with AlexNet RGB architecture | TITAN V  | 145.4     |
| Gupta et al.’s fusion net with VGG RGB architecture    | TITAN V  | 195.7     |
| Our fusion net                               | TITAN V  | 30.1      |

Figure 6. Detection results of our fusion model and other fusion models: The first column is Gupta et al. + Fast R-CNN net, the second column is Gupta et al.’s fusion net, and the third column is our fusion net. R-CNN: Region-based convolutional neural network.
Comparison with other fusion networks

Finally, we conducted experiments to compare some other popular fusion methods which merge the RGB and depth features. We evaluate the performance in terms of accuracy and time cost. Tables 2 and 3 present our experimental results. The RGB architecture of Gupta et al.'s10 Fast R-CNN supervision transfer method adopts VGG, and the Depth architecture uses AlexNet. Gupta et al.'s fusion net supervision transfer method uses AlexNet for supervision transfer. Our method adopts VGG architecture for supervision transfer.

Figure 6 shows the recognition results of some objects compared with other networks. We choose some images as negative samples. The first column is Gupta et al. + Fast R-CNN net, the second column is Gupta et al.'s fusion net, and the third column is ours. It can be seen from the figure that when other categories of objects appear in the pictures, their fusion networks have error recognitions and leak detections, and our fusion network model has a better recognition effect.

Table 3 reveals that our method makes a great progress in runtime. The speed over 30 frames per second (FPS) is fast enough to handle most real-time situations. Other fusion models focus on accuracy which slows down the speed, such as Chen et al.13 will take more than 1 s to finish one picture.

Conclusions

We introduce a fast multimodal neural network architecture for RGB-D object recognition, which achieves the state-of-the-art performance on our self-constructed solid waste data set. Our network adopts multi-scale combining layers at mid-level to merge features from RGB modality and depth modality. We apply supervision transfer to fine-tune depth SSD model from RGB model. By this method, the problem of not having a large data set can be avoided.

Experiments on solid waste images show that we meet the real-time requirements and ensure high precision. In the C&D waste sorting application,27 the results of our method are better than other popular methods. We also evaluate our approach on NYUD2 data set which is our past work25: We achieve 49.1% mAP and process images in real time at 35.3 FPS on one single Nvidia GTX1080 GPU. In future work, we intend to extend our method to deal with the overlapping of the contours28 in the presence of multiple objects.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by scientific research foundation of Education Department of Zhejiang Province of China under Grant Number Y201738129. This work was also supported by Founded by Science and Technology Department of Zhejiang Province of China under Grant Number LGG19F020010.

ORCID iDs

Yan Yu https://orcid.org/0000-0002-4827-2270
Siyu Zou https://orcid.org/0000-0002-4171-9092
Kejie Yin https://orcid.org/0000-0003-1494-4662

References

1. Liu S, Feng Y, Zhang S, et al. L0 sparse regularization-based image blind deblurring approach for solid waste image restoration. IEEE Trans Industr Elect 2019; 66(12): 9837–9845.
2. Li B, Zhang T, and Xia T. Vehicle detection from 3D lidar using fully convolutional network. arXiv preprint arXiv: 160807916 2016.
3. Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. Adv Neur Inf Proc Syst 2015; 39(6): 91–99.
4. Soviany P and Ionescu RT. Optimizing the trade-off between single-stage and two-stage deep object detectors using image difficulty prediction. In: 2018 20th international symposium on symbolic and numeric algorithms for scientific computing (SYNASC), Timisoara, Romania, 20–23 September 2018, pp. 209–214. IEEE.
5. Redmon J and Farhadi A. YOLOV3: an incremental improvement. arXiv preprint arXiv:180402767 2018.
6. Liu W, Anguelov D, Erhan D, et al. SSD: single shot multi-box detector. In: European conference on computer vision, pp. 21–37. Berlin: Springer.
7. Aytar Y and Zisserman A. Tabula rasa: model transfer for object category detection. In: 2011 International conference on computer vision, Barcelona, Spain, 6–13 November 2011, pp. 2252–2259. IEEE.
8. Zhang J, Zhang J, and Chen S. Discover novel visual categories from dynamic hierarchies using multimodal attributes. IEEE Trans Industr Inf 2013; 9(3): 1688–1696.
9. Eitel A, Springenberg JT, Spinello L, et al. Multimodal deep learning for robust RGB-D object recognition. In: 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS), Hamburg, Germany, 28 September–2 October 2015, pp. 681–687. IEEE.
10. Gupta S, Hoffman J, and Malik J. Cross modal distillation for supervision transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, 27–30 June 2016, pp. 2827–2836. IEEE.
11. Gupta S, Girshick R, Arbeláez P, et al. Learning rich features from RGB-D images for object detection and segmentation. In: European conference on computer vision, pp. 345–360. Berlin: Springer.
12. Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning, Bellevue, WA, 2011.
13. Chen K, Li J, Lin W, et al. Towards accurate one-stage object detection with AP-loss. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Long Beach, CA, 15–20 June 2019, pp. 5119–5127.

14. Hinton G, Vinyals O, and Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:150302531 2015.

15. Gupta S, Arbeláez P, Girshick R, et al. Aligning 3D models to RGB-D images of cluttered scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, 7–12 June 2015, pp. 4731–4740.

16. Cai G, Chen J, Wu Z, et al. One stage lesion detection based on 3D context convolutional neuralnetworks. Comput Electr Eng 2019; 79: 106449.

17. Duan L, Xu D, and Tsang I. Learning with augmented features for heterogeneous domain adaptation. arXiv preprint arXiv:12064660 2012.

18. Gong B, Shi Y, Sha F, et al. Geodesic flow kernel for unsupervised domain adaptation. In: 2012 IEEE conference on computer vision and pattern recognition, Providence, RI, 16–21 June 2012, pp. 2066–2073. IEEE.

19. Kulis B, Saenko K, and Darrell T. What you saw is not what you get: domain adaptation using asymmetric kernel transforms. In: CVPR 2011, Providence, RI, 20–25 June 2011, pp. 1785–1792. IEEE.

20. Hinton GE, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv 12070580 2012.

21. Sun S, An N, Zhao X, et al. A PCA–CCA network for RGB-D object recognition. Int J Adv Robotic Syst 2018; 15(1): 1–12.

22. Zhang J, Chen J, Wang Q, et al. Spatiotemporal saliency detection based on maximum consistency superpixels merging for video analysis. IEEE Trans Ind Inform 2019; 16(1): 606–614.

23. Chen X, Kundu K, Zhu Y, et al. 3D object proposals using stereo imagery for accurate object class detection. IEEE Trans Pattern Anal Mach Intell 2017; 40(5): 1259–1272.

24. Li G, Gan Y, Wu H, et al. Cross-modal attentional context learning for RGB-D object detection. IEEE Trans Image Proc 2018; 28(4): 1591–1601.

25. Yin K, Liu S, Liu R, et al. Cross modal multiscale fusion net for real-time RGB-D detection. In: 2018 24th International conference on pattern recognition (ICPR), Beijing, China, 20–24 August 2018, pp. 2386–2391. IEEE.

26. Krizhevsky A, Sutskever I, and Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neur Inf Proc Syst 2012; 25(2): 1097–1105.

27. Zhang J, Gui M, Wang Q, et al. Hierarchical topic model based object association for semantic slam. IEEE Transon Vis Comput Graph 2019; 25(11): 3052–3062.

28. Zhang J, Liu R, Yin K, et al. Intelligent collaborative localization among air-ground robots for industrial environment perception. IEEE Trans Industr Electr 2018; 66(12): 9673–9681.