Research on Intelligent Target Recognition Integrated With Knowledge

FENG ZHANG\(^1\), HUIBING FAN\(^2\), KEJU WANG\(^1\), YONGJIN ZHAO\(^1\), XIAOXI ZHANG\(^1\), AND YANG MA\(^1\)

\(^1\)No. 208 Research Institute of China Ordnance Industries, Beijing 102202, China
\(^2\)Beijing Mechanical and Electric Research Institute, Beijing 102202, China

Corresponding author: Huibing Fan (ayrh18811592868@163.com)

ABSTRACT With the development of artificial intelligence technology, intelligent weapon systems that can automatically identify, lock on and strike targets have gradually appeared and can replace humans in executing simple decision-making commands. Target detection is a key part of intelligent weapons. At present, large-scale target detection has serious challenges such as long-tail data distributions, severe occlusion, and category ambiguity. The main detection algorithms only detect each independent area without considering the key semantic dependencies between objects. It has become a hot trend to apply deep learning to prior knowledge to form a model. This article uses both internal and external knowledge to instill a target detection system with human reasoning capabilities. Commonly used external embedded knowledge includes geometric relations, attributes, locations, etc. They have a common shortcoming in that they require large amounts of labeled data, and the integration costs are huge. The purpose of this article is to construct a general external prior knowledge module to guide network learning. By paying attention to the characteristics of each object in different semantic contexts, the characteristics of each object are adaptively enhanced, and the high-level semantics of all categories evolve on a global scale. Internal knowledge uses a convolutional attention module that can learn spatial and channel information at multiple scales. The experimental results show the superiority of our knowledge-YOLOv5. The proposed method achieved 1.7%, 2.2%, 1.1%, and 0.7% improvements over YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, respectively, on the COCO data sets; and the proposed method also achieves a 0.9% improvement on the self-built data set. The trained lightweight model Knowledge-YOLOv5s is deployed on an NVIDIA Jetson TX2 through TensorRT acceleration, and the real-time detection frame is 20 ms, which meets the real-time detection requirements. This system can also be used as a module of an intelligent weapon system, which has certain referential significance for autonomous weapons and unmanned combat systems.

INDEX TERMS Prior knowledge, NVIDIA JetsonTX2, YOLOv5, intelligent weapon system, knowledge graph.

I. INTRODUCTION

The advancement and application of informatization and intelligent technology have caused profound changes in the form of modern warfare and the battlefield environment. The development and wide application of information technology has greatly extended the scope of the modern battlefield in both time and space. In addition to traditional battlefields that target firepower to strike and destroy active forces and related facilities, a remote informatized battlefield that relies on informatized operations has unfolded. The development of informatization and intelligence has equally extensive and far-reaching influence on the traditional battlefield. Due to the application of advanced information technology, the modern combat model and battlefield environment have undergone fundamental changes.

At present, the major military powers in the world are vigorously promoting the strategy of intelligent weapons, and the military application of artificial intelligence has become a hot spot in the domestic and foreign research and application. The U.S. military has intelligent unmanned equipment such as the RQ-4 Global Hawk reconnaissance...
aircraft, the MQ-1C “Gray Eagle” unmanned reconnaissance aircraft, unmanned ground combat vehicles, and unmanned submarine vehicles. There are tens of thousands of unmanned aerial systems and ground unmanned systems. Obviously, these systems have become an indispensable and important part of the US military’s military power. Russia has achieved many results on unmanned military platforms. For example, Uran-6 demining robots were used in Syria to remove booby traps and explosive devices. China attaches great importance to the development of artificial intelligence and accelerates drones and goallessness through the “military-civilian integration” strategy and the development of military fields such as unmanned combat platforms. Grasping the direction of future war development and the pulse of the times can allow countries to take the initiative and take the lead in the future battlefield.

The army’s trend is to minimize the direct participation of combat personnel in combat and to improve the rapid response capability and actual combat effectiveness of weapon system fire and strikes. Intelligent weapon systems are the developmental trend of future wars, and they are also domestic and foreign developmental hot spots. The new generation of the Russian unmanned combat vehicle “Uranus-9” has been tested in the Syrian battlefield. It can march independently and search for targets independently. Then, it can attack targets and conduct effect evaluations to complete a series of combat tasks. Target detection is an indispensable part of intelligent weapon systems, and identifying battlefield targets faster and more accurately is crucial. Although intelligent weapon stations already exist on foreign unmanned combat platforms and although domestic research on intelligent weapon stations has been conducted, there is no finished product thus far [1].

In recent years, artificial intelligence technology has developed rapidly. As an important branch of artificial intelligence machine learning, deep learning algorithms have been widely used in various fields. Target detection, voice recognition, and autonomous driving have developed rapidly. In a video image, the problems of heavy occlusion, class blur, and small-sized objects have become more challenging. The current state-of-the-art object detection methods identify each area separately and therefore require high-quality feature representation of each area and sufficient labeled data for each category. However, this is not the case for large-scale detection problems. The existing methods are inappropriate. Unlike human beings who can, even in complex situations, the current detection system lacks the ability to reason with common sense. Therefore, a key issue is how to give the current detection system reasoning ability to imitate the human reasoning process.

Figure 1 gives images randomly selected from the COCO data set. In the left image, the surfboard is seriously occluded, and it is difficult to detect its category. The features of the seagull image in the middle image are fuzzy, and the size of the human target in the right image is so small that it is difficult to identify.

Figure 2 is a knowledge map constructed according to Figure 1. The knowledge map can predict that people are skateboarding by surfing on the sea. The knowledge graph of ambiguous category issues helps the network determine that the white target is a seagull. Small targets are also a major problem in target detection. When the target detector cannot determine the type of target, the knowledge map can judge that the black object on the ship is a person. Target detection is no longer satisfied with limited types of detection but hopes that the objects learned by detection can be extended to new categories. That is, it is necessary to learn a general target detector and use limited annotation data to complete the large-scale detection of thousands of categories.

This paper uses internal and external prior knowledge to guide the model training, based on the current advanced target detection model YOLOv5; and uses the multiscale convolutional attention mechanism to better learn the internal feature information of the video. The proposed adaptive knowledge graph module can analyze the global semantic relationships. By modifying the supervision signal, the information is transferred to the feature layer to guide training, and the molecular network is corrected to add an adaptive semantic weight to the supervision signal. This weight can be reduced. This affects the influence of a noisy, complex sample frame on detection training when it is backpropagated. It also achieves the effect of data balancing and accelerating convergence. The system diagram of Knowledge-YOLOv5s embedded with internal and external knowledge is shown in Figure 3. Knowledge-YOLOv5 can perform adaptive global reasoning on categories with certain relationships or similar attributes. Therefore, the method can effectively improve the problems of serious occlusion, class blur, and small size; furthermore, it has guiding significance for the detection of limited samples.
II. METHODOLOGY

Object detection is the most common task in the field of computer vision, and the international academic community has been studying object detection for approximately 20 years. In recent years, deep learning technology has developed rapidly, and target detection algorithms have been upgraded from traditional algorithms based on manual features to detection technologies based on deep neural networks. The task of target detection is to find the target of interest in an image or a video and simultaneously detect the position and size of the target. In actual detection, the number of targets in a video changes; and the appearance, shape, posture, and angles of targets are different. Target imaging is affected by factors such as illumination and occlusion, which cause a detection algorithm to have a certain degree of difficulty.

Target detection technology based on deep learning is currently divided into two major directions: two-stage algorithms such as the R-CNN [2], fast R-CNN [3], faster R-CNN [4], R-FCN [5], etc.; and one-stage algorithms such as YOLO [6], SSD [7], etc. Feng et al. analyzed the advantages and disadvantages of single-stage and dual-stage algorithms. They believe that the single-stage algorithm has fast real-time performance but slightly lower accuracy [8]. The recent appearance of the YOLOv5 model with better real-time performance and accuracy has been well received, so this article chooses to use YOLOv5 as the basic framework. The feature extraction networks for target detection include VGG-16 [9] and ResNet-101 [10]. The core calculation of these CNN architectures is that the convolution operator learns new feature maps from the input feature map through the convolution kernel, including spatial (H and W dimensions) and interchannel (C dimension) feature fusion.

A. CONVOLUTIONAL ATTENTION MODULE

SENet [11], the champion attention mechanism of the ImageNet 2017 competition classification task, focuses on the relationships between channels, and the model can automatically learn the importance of different channel features. However, SENet may lose spatial information by focusing on the channel information for feature recalibration. Therefore, Sanghyun Woo et al. proposed an attention mechanism module (CBAM [12]) that combines spatial and channel data. Compared with SENet’s attention mechanism, which only pays attention to the channel, the CBAM can achieve better results. Figure 4 shows the structure of the CBAM. Channel attention and spatial attention are applied in turn to focus on important features and suppress unnecessary features.

Each channel of a feature is similar to a detector, and the channel attention determines the type of feature that is the focus of the collection. As shown in Figure 5, the two methods of global average pooling and maximum pooling use different information. The input is an $H \times W \times C$ feature $F$. We first perform spatial global average pooling and maximum pooling to obtain two $1 \times 1 \times C$ channel descriptions. Then, these descriptions are sent to a two-layer neural network. In the network, the number of neurons in the first layer is $C/r$, the activation function is ReLU, and the number of neurons in the second layer is $C$. Note that this two-layer neural network is shared. Then, the two obtained features are added and passed through a sigmoid activation function to obtain the weight coefficient $M_c$. Finally, the weight coefficient and the original feature $F$ are multiplied to obtain the new feature after scaling.

Figure 6 illustrates the spatial attention. Given an $H \times W \times C$ feature, we first perform the average pooling and maximum pooling of a channel dimension to obtain two $H \times W \times 1$ channel descriptions. These two descriptions are spliced together according to the channel. Then, after a $7 \times 7$ convolutional layer, the activation function is sigmoid, and the weight coefficient is obtained. Finally, the weight coefficient and the feature are multiplied to obtain the new feature after scaling.
B. INTERNAL KNOWLEDGE BASED ON MULTISCALE CONVOLUTIONAL ATTENTION

Convolutional neural networks have been developed from the field of neuroscience. Different sized receptive fields in the same area of the animal visual cortex are stimulated; therefore, neurons can process and collect different scales of spatial information in the same stage. Some scholars have introduced this theory into convolutional neural networks. The most typical structure is InceptionNet, which linearly aggregates different convolution kernels to conduct branch learning of feature information to enhance the adaptability of the network to multiple scales. However, InceptionNet uses a hierarchical linear aggregation method. Although it aggregates the multiscale information and multibranch features of different branches, it may not be sufficient to improve the strong adaptability of neurons. Therefore, this paper proposes a nonlinear multiscale convolution. The module can dynamically and adaptively adjust the size of the receptive field to improve the learning ability. Figure 7 shows the structural diagram of the multiscale convolutional attention module.

The multiscale convolutional attention mechanism integrates information from different convolution kernel branches. There are two branches: one is a convolution branch with a $3 \times 3$ convolution kernel, and the other is a $5 \times 5$ convolution branch. Because of the calculation parameters and costs, we replace the $5 \times 5$ convolution kernel with a $3 \times 3$ convolution kernel with a hyperparameter with an expansion rate of 2. The hole convolution is a very popular and effective method. The hole convolution was first proposed in the field of image segmentation to expand the receptive field due to the reduced resolution and information loss caused by downsampling. A $3 \times 3$ convolution kernel has a $5 \times 5$ receptive field or even a larger receptive field in the case of a hyperparameter with an expansion rate of 2, which represents the spacing of each value of the input data. Figure 8 is a schematic diagram of a hole convolution with an expansion rate of 2.

Based on the multiscale convolutional attention module, CSP1_1 embedded in YOLOv5 forms CSP1_1_IK, and the structure of CSP1_1_IK is shown in Figure 9.

The CBL is composed of the Conv, BN and Leaky ReLU, as shown in Figure 10.

C. TARGET DETECTION AND RECOGNITION TECHNOLOGY EMBEDDED WITH EXTERNAL KNOWLEDGE

As the data dividend of deep learning disappears, integrating external knowledge into deep learning network training and learning has become a hot spot. At present, in the field of computer vision, external knowledge is mainly embedded in the form of object attributes, tagged relationship knowledge, regional spatial knowledge and pairwise relationship co-occurrence knowledge. There have been studies that guide knowledge to conduct large-scale target detection, which can consider the relationship and shared attributes between objects [13]–[16]. There are also studies that find the similarity of semantic spatial attributes to guide target training [17]–[19]. External knowledge such as attribute knowledge and position relationship knowledge often needs to rely on a large amount of external knowledge. Manual labeling and large-scale data sets, such as the large-scale...
A typical large-scale knowledge graph will have millions or even hundreds of millions of concepts, and there will also be complex relationship types that can be mapped to our target categories.

**Step 2: Establish an adaptive connection network.**

Directly embedding the semantic knowledge map into the detection network will have correlated redundancy and noise, so it is necessary to realize the map modeling of each image and realize the adaptive activation of the semantic knowledge map. The adaptive connection network is shown in Figure 12.

The adaptive connection network is equivalent to the gate mechanism in the recurrent neural network. The parameter W is used to generate weights for each feature channel. The parameter W is learned to exploit model the correlation between feature channels. The self-adaptive connection from the network ε formula is as follows:

\[
\varepsilon = F_{\text{adaptive}}(R, W_a) = \delta(W_2\delta(W_1R))
\]

We first multiply W1 by z, which is a fully connected layer operation. The dimension of W1 is C/r × C. This r is a scaling parameter. The purpose of this parameter is to reduce the number of channels and reduce the amount of calculations. Then, the fully connected layer passes the data through a ReLU layer, multiplies by W2. The dimension of W2 is C' × C/r, so the output dimension and input time are unchanged. Finally, through sigmoid, the function gets S1.

**Step 3: Semantic knowledge graph revision.**

The adaptive knowledge graph R ∈ R^{N×N} obtains S through the fully connected layer and then weights it according to the original feature by multiplication. When S1 > 1, the network adaptively enhances semantic features; and when S1 < 1, the network suppresses irrelevant features and filters noise. The formula is as follows:

\[
S = S_1 \cdot S_0 = \varepsilon W_e \cdot S_0
\]

**III. EXPERIMENT ANALYSIS**

**A. DATASET DESCRIPTION AND EVALUATION INDICATORS**

This article conducted experiments on the public MSCOCO 2017 and Long-distance PC datasets. MSCOCO 2017 contains 118k training images and 5k testing images in a total...
of 80 categories. We use the evaluation criteria of COCO detection [26], that is, the average accuracy (mAP) of different IoU thresholds (IoU = {0.5: 0.95, 0.5, and 0.75}) and scales (small, medium, and large). We also use the average recall rate (AR) for each image with a different number of given detections ({1, 10, and 100}) and different scales (small, medium, and large). Long-distance PC dataset is a self-built data set that is composed of the public CUHK occlusion dataset, INRIA person dataset, and the Caltech partial data and web crawler datasets. The selection criteria is to select data with a longer distance and a smaller target. The data set has a total of 2000 images, of which 1800 are randomly selected as the training set and the remaining 200 are the test set. Targets are divided into two categories: people and cars. The people are in natural scenes. Cars include cars, trucks, SUVs, and buses.

### B. TRAINING DETAILS

The hardware uses the powerful GeForce RTX 2080 Ti x4 GPU with 11 GB video memory, an Intel(R) Core(TM) i7-9800X CPU, 16 GB memory, the Windows 10 operating system, and CUDA 10.2. A total of 200 epochs are set for training, the batch_size is set to 16, the input image resolution is 640 × 640, flipping and multiscale scaling are applied to the input images (0.8, 1.2), SGD is used as the optimization function, and the initial learning rate is 0.01.

### C. EXPERIMENTAL RESULTS

We conducted tests on the MSCOCO 2017 and Long-distance PC datasets. The input image size is 640×640. Table 1 shows that on the MSCOCO dataset, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x are integrated into the internal and external knowledge modules. After the knowledge module, the target detection mAP increased by 1.7, 2.2, 1.1, 0.7 percentage points, respectively. This shows that our internal and external knowledge modules can effectively guide the target detection network.

Knowledge-YOLOv5, which incorporates knowledge, improves the problems of occlusion, rare categories and ambiguity to a certain extent. Figure 4 shows the comparison of the qualitative results between our Knowledge-YOLOv5 and YOLOv5. Our Knowledge-YOLOv5 infers the mouse near the computer and the keyboard through the adaptive category knowledge relationship. The recognition in the right picture can identify overlapping categories, windows and walls.

The NVIDIA Jetson TX2 is applied to the self-built Long-distance PC dataset as a human-vehicle recognition system. YOLOv5s is a lightweight model. Therefore, this article only conducts experiments on YOLOv5s. The experimental results are shown in Table 2. The internal knowledge module improves the performance. Because there are fewer types of self-built data sets, the embedded external knowledge module hardly improves the performance, and the embedded internal and external knowledge modules still provide the largest increase.

Table 2 shows that the internal multiscale convolutional attention mechanism performs well on the self-built Long-distance PC dataset, and thus the mechanism can effectively solve the problems of target occlusion and missed detection. Figure 14 shows the detection effect of Knowledge-YOLOv5 on the self-built data set. The real-time detection frame of YOLOv5s with embedded knowledge is almost unchanged on the GTX 2080Ti. The model is directly run

---

**TABLE 1.** Experimental results and performance of knowledge-YOLOv5 on the COCO dataset.

| Model       | Size (pixels) | mAP\textsuperscript{test} 0.5:0.95 | mAP\textsuperscript{val} 0.5 | Speed 2080Ti (ms) |
|-------------|---------------|-----------------------------------|-----------------|------------------|
| YOLOv5s     | 640           | 36.7                              | 55.4            | 20               |
| YOLOv5m     | 640           | 44.5                              | 63.1            | 28               |
| YOLOv5l     | 640           | 48.2                              | 66.9            | 38               |
| YOLOv5x     | 640           | 50.4                              | 68.8            | 66               |
| Knowledge-YOLOv5s | 640       | 38.4                              | 58.2            | 21               |
| Knowledge-YOLOv5m | 640       | 46.7                              | 64.6            | 29               |
| Knowledge-YOLOv5l | 640       | 49.3                              | 67.9            | 39               |

**TABLE 2.** Effects and performance of embedded knowledge on the self-built dataset.

| Model       | mAP\textsuperscript{test} 0.5:0.95 | Speed 2080Ti (ms) | Speed TX2 (ms) | Speed TX2-TensorRT (ms) |
|-------------|-----------------------------------|------------------|----------------|------------------------|
| YOLOv5s     | 0.875                             | 20               | 61             | 40                     |
| (External knowledge) YOLOv5s | 0.877                            | 20               | 61             | 40                     |
| (Internal knowledge) Knowledge-YOLOv5s | 0.882                           | 20               | 61             | 40                     |
| Knowledge-YOLOv5s | 0.884                           | 22               | 60             | 41                     |
on an NVIDIA Jetson TX2. The average real-time detection frame is 60 ms. After using TensorRT acceleration, the real-time detection frame is 41 ms. The real-time performance and accuracy are better.

**D. ABLATION STUDIES**

This article states that internal knowledge is composed of multiscale convolutional attention modules. We embed external knowledge and CBAM modules into four models: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The test results on the MSCOCO2017 data set are shown in Table 3. The YOLO with the CBAM module embedded has almost no effect, and the YOLOv5 with the internal knowledge module embedded will achieve some improvement, which shows the effectiveness of the external knowledge.

External knowledge is the adaptive knowledge graph embedded in YOLOv5. In this paper, the knowledge graph and the adaptive knowledge graph are directly embedded in the four YOLOv5 models. The experimental results are shown in Table 4. In contrast, the experimental results show that embedding the knowledge graph directly reduces the performance, and the model with the adaptive knowledge graph embedded greatly improved the performance, which illustrates the effectiveness of external knowledge to a certain extent.

**E. HIGH GENERALIZATION AND SMALL TARGET RECOGNITION PERFORMANCE**

The Knowledge-YOLOv5 model has a high degree of generalization. As mentioned previously in the article, the self-built data set includes ordinary people and vehicles (cars, trucks,
SUVs, and buses). We detected people in military uniforms and armored vehicles and tanks in complex environments using a good-looking video app and the training data set. Without such data, we can still detect different people and cars very well. Figure 15 is an effect map of images with large differences between the background and the targets.

The target detection industry is concerned not only with speed but also with the recognition effect of long-distance small targets. We took a set of images with a camera for a recognition test. Figure 16 show that people and car targets can be correctly identified in the image with small pixels. At a 1920*1280 resolution, 16*16 pixels can be recognized, which is the goal. The experiments prove that our model has a good effect on the recognition of small targets at long distances.

IV. CONCLUSION

We propose a novel adaptive knowledge guidance network called Knowledge-YOLOv5. The network constructs two novel general internal and external knowledge modules, namely, a multiscale convolutional attention module and an adaptive knowledge map, which enhance the classification and positioning of features. These modules express and adaptively coordinate with the visual patterns in each image. Both of these modules can be easily applied to different target detection systems, and both achieve performance improvements on the COCO dataset and self-built data set. The experiments and analysis show that Knowledge-YOLOv5 can effectively alleviate the problems of large-scale target detection occlusion and category ambiguity. The lightweight Knowledge-YOLOv5s model is run on an NVIDIA Jetson TX2 and accelerated by TensorRT to form a real-time intelligent knowledge target system. The intelligent target system is highly generalizable and can compare video targets and small targets with large differences in datasets. The method achieves a good recognition effect. The target recognition system has certain reference significance for intelligent weapon systems.

V. FUTURE OUTLOOK

This paper proposes an internal knowledge module based on an adaptive multiscale. The module cannot only learn the relationship between video spaces and channels but also adaptively learn effective information on different scales. The experiments show that the detection effect has been improved to a certain extent. Aiming at the problem that the learned features and knowledge are limited by the size and distribution of training data, this paper proposes an adaptive knowledge graph that guides network learning from high-level semantics, thereby improving the performance of the target detection system. As future work, to expand upon the present work, we can apply our knowledge framework to other tasks, such as instance-level segmentation.

The external prior knowledge used in target detection in this article is relatively limited, and we will consider applying more advanced knowledge graphs such as ConceptNet5.5 later. The target intelligence system can be applied to military intelligence fields such as intelligent weapon systems. We will optimize and upgrade the system based on the results. We will also refine the algorithm, such as by adding a target threat ranking algorithm and a target location and damage evaluation algorithm, and strive to build functions and a more comprehensive and practical intelligent identification system.

REFERENCES

[1] Z. Bowen, M. Baoquan, and Z. Mengchun, “Development status and key technology analysis of intelligent weapon station,” Fire Power Command Control, vol. 45, no. 5, pp. 1–7, 2020.
[2] K. Lenc and A. Vedaldi, “R-CNN minus R,” in Proc. BMVC, 2015, pp. 5.1–5.12.
[3] R. Girshick, “Fast R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1440–1448.
[4] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[5] J. Dai, Y. Li, K. He, and J. Sun, “R-FCN: Object detection via region-based fully convolutional networks,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2016, pp. 379–387.
[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 779–788.
[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single shot multibox detector,” 2015, arXiv:1512.02325. [Online]. Available: http://arxiv.org/abs/1512.02325
[8] Z. Feng, X. Wu, and C. Gu, “Detection of road surface identifiers based on deep learning,” in Proc. Int. Conf. Artif. Intell. Adv. Manuf. (AIAM), 2019, pp. 66–70.

[9] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. ICLR, 2015, pp. 1–14.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. CVPR, Jun. 2016, pp. 770–778.

[11] H. Jie, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[12] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. ECCV, 2018, pp. 3–19.

[13] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, “Label-embedding for attribute-based classification,” in Proc. CVPR, Jun. 2013, pp. 819–826.

[14] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, “Word spotting and recognition with embedded attributes,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 12, pp. 2552–2566, Dec. 2014.

[15] C. H. Lampert, H. Nickisch, and S. Harmeling, “Learning to detect unseen object classes by between-class attribute transfer,” in Proc. CVPR, Jun. 2009, pp. 951–958.

[16] I. Misra, A. Gupta, and M. Hebert, “From red wine to red tomato: Composition with context,” in Proc. CVPR, Jul. 2017, pp. 1792–1801.

[17] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, and T. Mikolov, “Devise: A deep visual semantic embedding model,” in Proc. NIPS, 2013, pp. 1–11.

[18] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille, “Learning like a child: Fast novel visual concept learning from sentence descriptions of images,” in Proc. ICCV, Dec. 2015, pp. 2533–2541.

[19] S. Reed, Z. Akata, H. Lee, and B. Schiele, “Learning deep representations of fine-grained visual descriptions,” in Proc. CVPR, Jun. 2016, pp. 49–58.

[20] X. Chen, L.-J. Li, L. Fei-Fei, and A. Gupta, “Iterative visual reasoning beyond convolutions,” in Proc. CVPR, Jun. 2018, pp. 7239–7248.

[21] B. Dai, Y. Zhang, and D. Lin, “Detecting visual relationships with deep relational networks,” in Proc. CVPR, 2017, pp. 3076–3086.

[22] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in Proc. ICLR, 2017, pp. 1–14.

[23] K. Marino, R. Salakhutdinov, and A. Gupta, “The more you know: Using knowledge graphs for image classification,” in Proc. CVPR, Jul. 2017, pp. 1–9.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” Comput. Sci., 2013.

[25] G. A. Miller, “WordNet: A lexical database for English,” Commun. ACM, vol. 38, no. 11, pp. 39–41, 1995.

[26] D. Ramana, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in Proc. ECCV, 2014, pp. 740–755.

FENG ZHANG received the master’s degree in signal and information processing from the Communication University of China, Beijing, China, in 2020. He is currently an Engineer with the No. 208 Research Institute of China Ordnance Industries. His research interests include deep learning algorithm research, target detection and recognition, and intelligent video analysis.

HUIBING FAN received the master’s degree in control engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2003. He is currently a Senior Engineer with Beijing Mechanical and Electric Research Institute. His research interests include computer control and artificial intelligence technology research.

KEJU WANG received the master’s degree in control engineering from Beijing Technology and Business University, Beijing, China, in 2020. He is currently an Engineer with the No. 208 Research Institute of China Ordnance Industries. His research interests include FPGA accelerated neural networks and embedded neural network realization.

YONGJIN ZHAO received the bachelor’s degree in electrical engineering and automation from the University of Electronic Science and Technology of China, in 2003. He is currently a Senior Engineer with the No. 208 Research Institute of China Ordnance Industries. His research interests include system overall design and fire control system design.

XIAOXI ZHANG received the master’s degree in power electronics and power transmission from Heilongjiang University of Science and Technology, in 2008. She is currently working with the No. 208 Research Institute of China Ordnance Industries. She is also a Senior Engineer. Her current research interest includes complex system control based on intelligence.

YANG MA received the master’s degree in computer science and technology from Beijing Institute of Technology, Beijing, China, in 2020. She is currently a Senior Engineer with the No. 208 Research Institute of China Ordnance Industries. Her research interest includes weapon system display and control.