IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 4, 2020, accepted March 22, 2020, date of publication March 25, 2020, date of current version April 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2983174

GAN-Knowledge Distillation for One-Stage Object Detection

WANWEI WANG[1], WEI HONG[2], FENG WANG[2], AND JINKE YU[2]
[1]Tianjin Key Laboratory for Advanced Signal Processing, Civil Aviation University of China, Tianjin 300300, China
[2]Beijing Zeusee Technologies Company, Ltd., Beijing 100101, China

Corresponding author: Wanwei Wang (wwwang@cauc.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB0502405, and in part by the Fundamental Research Funds for the Central Universities of the Civil Aviation University of China under Grant 3122013D020.

ABSTRACT Convolutional neural networks (CNN) have a significant improvement in the accuracy of object detection. As networks become deeper, the precision of detection becomes obviously improved, and more floating-point calculations are also needed. Because of the great amount of calculation, it is inconvenient for mobile and embedded vision applications. Many researchers apply the knowledge distillation method to improve the precision of object detection by transferring knowledge from a deeper and larger teachers network to a small student one. Most methods of knowledge distillation are needed to design complex cost functions and mainly aim at the two-stage object detection algorithm. Therefore, we propose a clean and effective knowledge distillation method called Generative Adversarial Networks - Knowledge Distillation(GAN-KD) for the one-stage object detection. The feature maps generated by teacher network and student network are employed as true and fake samples respectively, and generating adversarial training for both of them to improve the performance of the student network in one-stage object detection. The experimental result shows that our approach achieves the performance gain of 5% mAP when compared with MobilenetV1 on COCO dataset.

INDEX TERMS Object detection, generative adversarial networks, knowledge distillation.

I. INTRODUCTION

In recent years, Convolutional Neural Networks (CNN) have become ubiquitous in computer vision. With the development of deep learning, researchers have found that the accuracy of object detection has significant improvement with deeper and larger convolution neural network as the backbone of object detection. With the improvement of object detection accuracy, computer vision moves from common areas to critical areas (such as unmanned driving and medical fields). However, in order to ensure the detection accuracy, a larger convolution neural network has to be applied as backbone of object detection. It leads to the decrease of the detection speed and the increase of the cost of computing equipment, which could not meet the real-time requirements in real world. Therefore, many researchers propose many other methods to improve the detection speed on the premise of ensuring the detection accuracy, such as methods of reducing the number of floating-point operations of convolution neural network by depthwise separable convolution [1], [2], pointwise group convolution and channel shuffle [3], [4]. Although these methods have achieved considerable speed-up results, they require careful design and tuned backbone of network. Many researchers believe that although the deeper backbone network has a larger network capacity to achieve a better performance in the image classification, object detection and other tasks. However, some specific tasks cannot use large neural networks because of a small capacity, so the neural network structure should be compressed, quantized and channel pruned so as to ensure the accuracy of the CNN [5]–[9].

On the other hand, some studies [10]–[13] on knowledge distillation show that, we can employ a deeper model as teacher net and a lighter model as student net, then train student net with soft label combined with true label which is the output or intermediate result of teacher net. It can greatly improve the performance of student net on specific tasks. But most of these methods require very complex cost functions and training methods, and consequently they are mainly used for image classification, two-stage object detection and so on,but rarely applied in one-stage object detection. Therefore,

The associate editor coordinating the review of this manuscript and approving it for publication was Liangxiu Han.

a simpler and more effective knowledge distillation method is needed that can be applied to one-stage object detection.

This paper proposes a novel and effective knowledge distillation neural network architecture, which can obviously improve the performance of student net in one-stage object detection. Different from the conventional knowledge distillation method, we refer to the architecture of Generative Adversarial Networks (GAN) [14]. We separated the backbone of deeper object detection neural network and lighter object detection neural network as teacher net and student net respectively, then used the feature map generated by the teacher net as true samples, and the feature map generated by the student net as fake samples. Finally, we designed a neural network as a discriminator and utilized true samples and fake samples to conduct the generative adversarial training.
The main contributions of this paper can be summarized as follows:

1) A clean and effective architecture of knowledge distillation is proposed. There is no need to manually specify the location of knowledge distillation. It does not require the design of complex cost functions, and can be applied to one-stage object detection.
2) We use the architecture of GAN to avoid complex designs of knowledge migration. Our experiments show that this method can obviously improve the performance of student net (such as Mobilenet and ResNet50) in one-stage object detection. The efficient student network can satisfy requirements for mobile and embedded vision applications.

This paper is organized as follows. We briefly review related work in Section II. Section III presents our proposed GAN-KD algorithm. We report and discuss the results of our experiments and our future work in Section IV. Conclusions are presented in Section V.

II. RELATED WORK

A. OBJECT DETECTION

1) CNN FOR DETECTION

Object detection is still an active research area in the field of computer vision, and considerable progress and success have been achieved in this field by designing of deep convolutional neural networks for object detection. The deep learning architecture of object detection is mainly divided into two types:

1) First type is the one-stage object detection algorithm, such as the YOLO (You Only Look Once) [15], SSD (Single Shot MultiBox Detector) [16], etc. Liu W *et al.* proposed the SSD method for detecting objects in images using a single deep neural network, which directly returns the object position and category.
2) Second type is the two-stage object detection algorithm, such as Fast R-CNN (Regions with CNN features) [17], Faster R-CNN [18] and R-FCN (Region-based Fully Convolution Network) [19], etc. They usually have two steps to detect objects. Firstly they regresses the proposal boxes by the CNN, then identifies each proposal box again, and finally returns to the correct location and category.

2) SINGLE SHOT MultiBox DETECTOR (SSD)

SSD [16] is based on a feed-forward convolutional network that produces a fixed-size collection of bounding boxes and scores for the presence of object class instances in those boxes, followed by a non-maximum suppression step to produce the final detection. The network layers are based on a standard architecture used for high quality image classification (truncated before any classification layers), which will be named the backbone network. Convolutional feature layers of head network are added to the end of the truncated backbone network. These layers decrease in size progressively and allow the predication of detection at multiple scales.

B. GENERATIVE ADVERSARIAL NETWORKS (GAN)

A GAN involves generator and discriminator networks which respectively aim to map random noise to samples and discriminate real and generated samples. It attempts to generate new data from a given data set and has achieved impressive results for a variety of data types. Goodfellow *et al.* [20] formulated GAN into a two-player game, where they simultaneously train two models: a generator network G which captured the data distribution, and a discriminator network D which estimated the probability of a sample from the true data rather than the generator network G. InfoGAN [21] is an information-theoretic extension to GAN. InfoGAN decomposed the input noise vector into two parts: incompressible noise vector z and latent code vector c. Conditional GAN (CGAN) [22] added extra label information y' to generator G for conditional generation. In discriminator D, both x and y were presented as inputs and D tried to distinguish if data-label pair is from generated or real data.

C. NETWORK COMPRESSION

To further improve the forward efficiency of one-stage detectors, model compression techniques were usually introduced to reduce the model capacity and computation complexity. Many researchers believed that deep neural networks were over-parameterized, and it had too many redundant neurons and connections. He textitet al. [8] thought neurons in each layer of convolutional neural networks are sparse, and they used lasso regression to detect the most representative neuron per layer of convolutional neural networks reconstructing the output of this layer. Approaches for model compression could be classified into the following areas: network pruning [9], [23]–[25], network quantification [6], [26], network simplification [1], [4], [27] and knowledge distillation [6], [12], [28]–[30], etc.

1) NETWORK PRUNING

Network pruning aims to reduce parameter redundancy by inducing model sparsity. Existing pruning methods either train from scratch with sparsity constraints on channels, or minimize the reconstruction error between the pre-trained feature maps and the compressed ones. Zhuang *et al.* [9] believed that layer-by-layer channel pruning affected the discriminating ability of Convolutional neural networks, so the auxiliary ability of convolutional neural networks was preserved by adding auxiliary loss in the fine-tune and pruning stages. Reference [23] proposes to prune the unimportant connections with small weights in trained neural networks. The resulting network's weights are mostly zeros thus the storage space can be reduced by storing the model in a sparse format. Reference [24] imposes neuron-level sparsity in training thus some neurons could be pruned to obtain compact networks. Reference [25] proposes a Structured Sparsity Learning (SSL) method to sparsity different levels of structures (e.g. filters, channels or layers) in CNNs. Both methods utilize group sparsity regularization in training process to obtain structured sparsity. Instead of resorting to group sparsity on convolutional weights, our approach imposes simple L1 sparsity on channel-wise scaling factors, thus the optimization objective is much simpler.

2) NETWORK QUANTIFICATION

Network quantification achieves regularization by clustering parameters and activation onto discrete and reduced-precision points. Wu *et al.* [6] used the k-means clustering algorithm to accelerate and compress the convolutional layer and the fully connected layer of the model to obtain better quantization results by reducing the estimation error of the output response in each layer, and proposed an effective training scheme to suppress the multi-layer cumulative error after quantization. Jacob *et al.* [26] proposed a method that quantified weights and inputs as uint8, and bias to unit32, at the same time, the forward used quantization, and the backward correction error was not quantized to ensure that the CNN performance and speed of inference during training.

3) NETWORK SIMPLIFICATION

The target of network simplification is to design and obtain more efficient CNN models, such as SqueezeNet, MobileNet and ShuffleNet, etc. Moving towards this goal, a CNN architecture called SqueezeNet [27] was proposed which had 50x fewer parameters than AlexNet and maintains AlexNet-level accuracy on ImageNet. MobileNets [1] was based on a streamlined architecture that used depthwise separable convolutions to build light weight deep neural networks. Ma *et al.* [4] proposed that network architecture design should consider the direct metric such as speed, rather than the indirect metric like FLOPs and presented practical guidelines and a novel architecture.

4) KNOWLEDGE DISTILLATION

Knowledge distillation is an efficient method of model compression and knowledge transfer, which aims at training a smaller network to mimic a more complex teacher network. Hinton *et al.* [10] used the result of teacher net output as the soft label of student net, and advocated the use of temperature cross entropy instead of L2 loss. Romero *et al.* [28] believed that student net needed more unlabeled data to be as close as possible to teacher net. When Chen*et al.* [12] distilled the two-stage object detection, they extracted the middle feature map of teacher net and the dark knowledge of R-CNN respectively to train the student net. There were also some researchers who give the attention information of teacher network to the student network. For example, Zagoruyko and Komodakis [29] *et al.* proposed spatial-attention, which is a method of transmitting the thermal information of teacher net to student net. Yim *et al.* [30] used the relationship between layer and layer of teacher net as the learning goal of student network.

However their knowledge distillation required the design of very complex loss functions to extract complex dark knowledge. Therefore, many people propose new framework to distill knowledge with generative adversarial networks, such as KDGAN [31](Knowledge Distillation with Generative Adversarial Networks), MEAL [32] (Multi-Model Ensemble via Adversarial Learning), [33] proposed by Xu *et al.*, and [34] proposed by Liu *et al.* Existing GAN based knowledge distillation methods mainly focus on simple tasks like classification, image tag recommendation and image segmentation, while do not consider complex tasks like object detection. Even object detection algorithms mostly focus on the two-stage object detection, rarely used in one-stage object detection.

In order to have an effective way of knowledge distillation, we referred to the architecture of the GAN [14] to take the feature map generated by the teacher network and the student network as true samples and fake samples respectively. Finally we designed a neural network as a discriminator and applied true samples and fake samples to conduct the generative adversarial training to improve the performance of the student network in one-stage object detection.

In this paper, we use the one-stage object detection algorithm SSD [16] as object detection framework. The architecture of SSD is mainly divided into two parts, 1) backbone of network, used as feature extractor. 2) SSD-Head, use the features extracted by the backbone of network to detect the category and location of the object. In order to obtain a better knowledge distillation effect, it is important to make rational use of these two parts.

III. METHOD

In this section, we introduce our proposed model GAN-KD for one-stage object detection in detail, and discuss the relevant training methodology.

A. OVERALL STRUCTURE

GAN-KD, as shown in Figure 1, is composed of four modules: teacher network, student network, discriminative network, SSD-HEAD network. Figure 1 is the overall structure of our model.

We adopt the one-step object detection algorithm SSD as our framework. The structure of SSD algorithm is composed of both feature extractor and object detector. We first

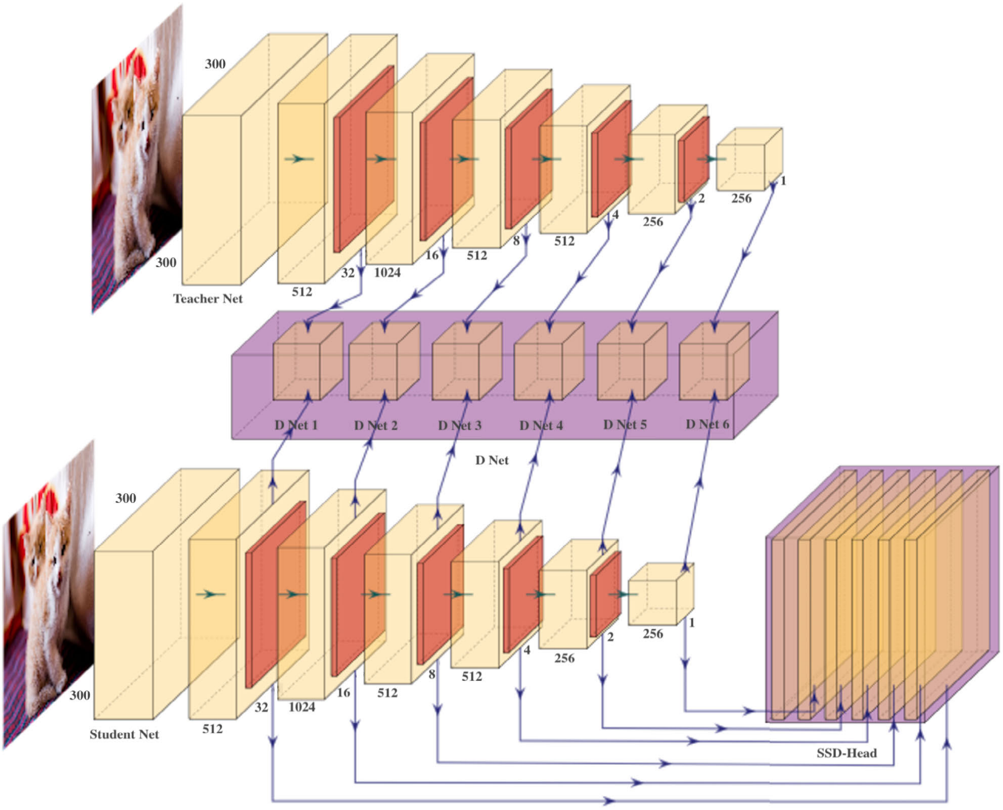


FIGURE 1. **GAN-KD network architecture. Teacher Net (the top left) are the backbone network of the larger and fully trained SSD model. Student Net (the bottom left) is a smaller network such as Mobilenet. SSD-Head (the bottom right) are the head network of SSD model. D-Net is a module consisting of six small discriminant networks.**

use a SSD model with a larger capacity and full training, and split the SSD model into backbone network and SSD-Head network. We take the backbone network as the teacher net, and pick up a smaller network as student net, such as MobileNet, ShuffleNet and ResNet. In the traditional knowledge distillation algorithm, it is necessary to design a very complex loss function. GAN algorithm is employed to complete the migration from teacher network to student network.

The teacher net and student net have six feature layers of same structure. We use multiple feature maps generated by teacher net as true samples, and multiple feature maps, produced by student net as fake samples, then send the true sample and the fake sample to each of the corresponding discriminative networks which are integrated together into D-Net (shown as Figure 2). D-Net learns to determine whether a sample is from the teacher net or the student net. The process drives student net to improve its methods until it has the same effect as the teacher net.

After feature is extracted, the output of student net become the input sample of the SSD-Head network.

B. TRAINING OBJECTIVE

Our training process involves two stages. The first stage is mainly generative adversarial training which performs on teacher net and student net. The second stage is normal SSD training which performs on student net and SSD-Head network.

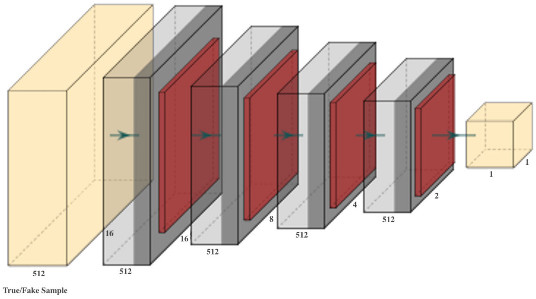


FIGURE 2. **One of the discriminative network architecture in the D-Net module. It consists of multiple downsampled convolutional layers and outputs a single scalar.**

1) GENERATIVE ADVERSARIAL TRAINING

First, we train each discriminative network in the D-Net module to identify whether the input sample is true or fake, and freeze the weights of teacher net and student net. we train the student net to generate fake samples to trick each of the discriminative networks in the D-Net module. For each sample x we employ loss function:

$$L_{Dt} = D\left(Teacher\left(x, \theta_t\right), \theta_d\right) \quad (1)$$

$$L_{Ds} = D\left(Student\left(x, \theta_s\right), \theta_d\right)] \quad (2)$$

$$L_D = \frac{1}{N}\left(L_{Dt} - L_{Ds}\right) \quad (3)$$

where Teacher and Student represent teacher net and student net, respectively. D represents the discriminative network, which is a second multi-layer perceptron and outputs a single scalar. It represents the probability that x came from the true sample. θ_t, θ_s, and θ_d represent the weights of the teacher net, the weights of the student net, and the weights of each discriminative network in the D-Net module, respectively. We train D to maximize the probability of assigning the correct label to samples from student net. N in Equation 3 represents the batch sizes.

2) OBJECT DETECTION TRAINING

In the second stage, normal SSD training is performed on student net and SSD-Head network, after many epochs of generative adversarial training. During the training process, the weights of the teacher net and the weights of each discriminative network in the D-Net module are frozen. The overall objective loss function is a weighted sum of the discriminative loss, localization loss (loc) and the confidence loss (conf), and we use the loss function:

$$L_{conf} = L_{conf}\left(Student\left(x, \theta_s\right)\right) \quad (4)$$

$$L_{loc} = L_{loc}\left(Student\left(x, \theta_s\right)\right) \quad (5)$$

$$L_G = \frac{1}{N}\left(L_{Ds} + L_{conf} + L_{loc}\right) \quad (6)$$

where L_{conf} represents the loss function of the classification in the SSD, and L_{loc} represents the loss function of the bounding box in the SSD. D represents the discriminant network, Teacher and Student represent teacher net and student net, respectively.

IV. EXPERIMENT

In this section, we evaluate the effectiveness of proposed method by distilling the knowledge of state-of-the-art methods. We experiment in the PASCAL VOC, MS COCO(Microsoft Common Objects in Context) and MPID (Multi-Purpose Image Deraining) data set to validate our approach. Our hardware device is two NVIDIA GTX 1080Ti GPUs. The software framework is GluonCV. For all the experiments, we use the same settings and input size (300 × 300).

A. TRAINING PROCESS

All models use Adam(Adaptive moment estimation) with 0.0005 learning rate in the first phase and SGD(Stochastic Gradient Descent) with 0.0005 weight decay and 0.9 Momentum in the second phase. The first stage trains epochs is 180 and the second stage epochs is 90. The student nets are MobilenetV1, MobilenetV2 and ResNet18, and teacher nets are VGG16, ResNet50 and ResNet101. These models have pre-trained under ImageNet.

B. DATA AUGMENTATION

To make the model more robust to various input object sizes and shapes, we further augmented the training data set by flipping, rotating (each image was rotated by 5, 10, 15 degrees both in clockwise and counter clockwise orientations), color and contrast enhancing, and noise-adding.

C. RESULTS

1) PASCAL VOC DATASET

PASCAL VOC 2007 data set [35] is a standard benchmark for the task of object detection. It contains 20 object categories and one background class and consists of 10,582 images for training (including VOC 2012 training set and additional data annotated in [36]), 1,449 images for validation and 1,456 for test. In the test data set, we evaluated the performance of the object detector via mean average precision (mAP), a standard metric in PASCAL VOC. We consider that a bounding box is correct if it had an IoU ratio of at least 50% with the ground-truth object annotation.

TABLE 1. Object detection mAP(%) on the PASCAL VOC 2007 test set. Test results of different student nets which are not used GAN-KD and used GAN-KD in different teacher net.

Student net	Teacher net	VOC 2007 test
MobilenetV1	-	75.4
	VGG16	77.3(**+1.9**)
	ResNet50	77.6(**+2.2**)
	ResNet101	77.6(**+2.2**)
MobilenetV2	-	75.9
	VGG16	77.2(**+1.3**)
	ResNet50	77.7(**+1.8**)
	ResNet101	77.5(**+1.6**)
ResNet18	-	74.8
	VGG16	77.2(**+2.4**)
	ResNet50	77.6(**+2.8**)
	ResNet101	77.3(**+2.5**)

From Table 1, we compared the native SSD with the SSD of GAN-KD under different teacher nets, which could improve student net 2.8% mAP at the highest level on PASCAL VOC data set. When the teacher net was ResNet101 and student net was ResNet18, the improvement was not as great as teacher net of ResNet50. In Figure 3, we showed some detection examples on PASCAL VOC with the ResNet50 as teacher net and MobilenetV1 as student net.

2) MS-COCO DATASET

The MS-COCO dataset is a challenging object detection data set. Experiments were carried out on MS-COCO 2017 [37], which contains 118k images for training, 5k for validation (val) and 20k for testing without provided annotations (test-dev). Detectors were trained on COCO training set, and evaluated on the validation set.

From the Table 2, we found that the lower mAP on the data set is, the more obvious effect of the student net after GAN knowledge distillation is. So we used MobilenetV1_0.75 as

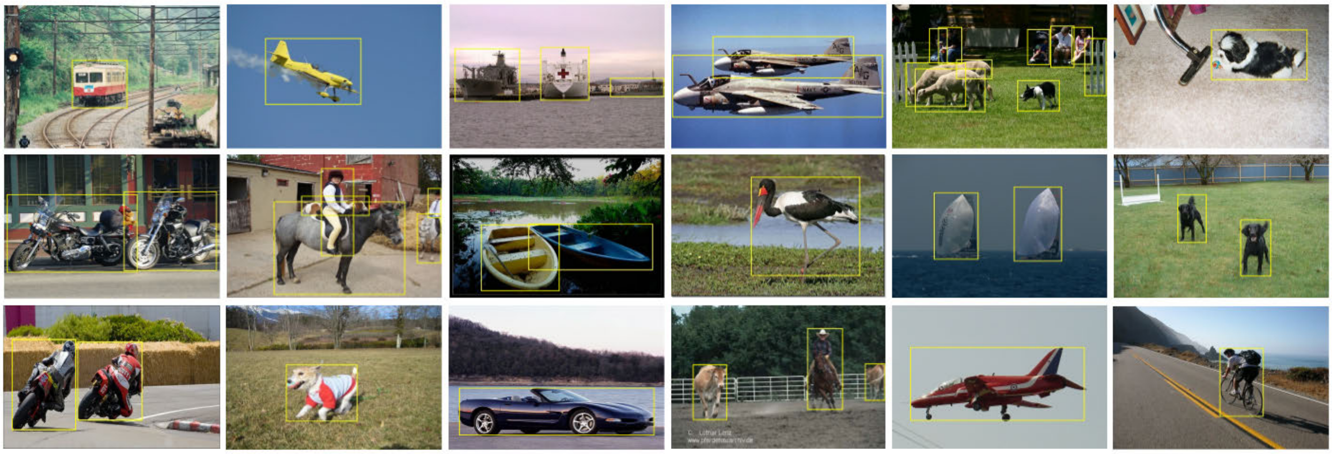


FIGURE 3. Detection results on Pascal VOC 2007 with our GAN-KD model. We show detection results with scores higher than 0.6.

TABLE 2. Object detection mAP(%) on the MS COCO 2017 validation set. MoblienetV1 and MoblienetV2 use GAN-knowledge distillation in different teacher net.

Student net	Teacher net	COCO 2017 val
MobilenetV1	-	21.7
	VGG16	24.7(**+3.0**)
	ResNet50	25.4(**+3.7**)
	ResNet101	25.0(**+3.3**)
MobilenetV1_0.75	-	18.0
	VGG16	21.8(**+3.8**)
	ResNet50	23.0(**+5.0**)
	ResNet101	22.6(**+4.6**)
MobilenetV2	-	22
	VGG16	24.1(**+2.1**)
	ResNet50	24.6(**+2.6**)
	ResNet101	24.2(**+2.2**)

TABLE 3. Object detection mAP(%) on the MPID (RID and RIS) data set employing GAN-KD and not using GAN-KD in different teacher net.

Student net	Teacher net	MPID
MobilenetV1	-	16.7
	VGG16	18.2(**+1.5**)
	ResNet50	20.1(**+3.4**)
	ResNet101	19.8(**+3.1**)
MobilenetV1_0.75	-	16.7
	VGG16	18.4(**+1.7**)
	ResNet50	20.6(**+3.9**)
	ResNet101	20.1(**+3.4**)
MobilenetV2	-	17.2
	VGG16	18.7(**+1.5**)
	ResNet50	20.9(**+3.7**)
	ResNet101	20.2(**+3.0**)

the student net, and we have increased 5% mAP on COCO data set. In Figure 4, we showed some detection examples on COCO with the ResNet50 as teacher net and MobilenetV1 as student net.

3) MPID DATASET

The images of PASCAL VOC and COCO data set are gained in normal environment, we also experimented our method in extreme environment to further test the generality of this approach. Li *et al.* [38] conducted a study of object detection in rainy days. It is a difficult problem because rain is a complicated atmospheric process and could cause several different types of visibility degradations, due to a magnitude of environmental factors including raindrop size, rain density, and wind velocity. In addition, the author opened a rainy day data set called Multi-Purpose Image Deraining (MPID). MPID data set is a new large-scale benchmark consisting of both synthetic and real-world rainy images of various rain types, such as rain streak, raindrop, and rain and mist. Li *et al.* compared the precision of detection algorithms with Faster-RCNN, YOLOv3, RetinaNet, VGG and other state-of-the-art object detection algorithms on MPID data set.

Therefore, in order to prove the availability of our network, we also applied MPID data set to test on rainy day images. MPID data set mainly contains two parts: a Rain in Driving (RID) set collected from car-mounted cameras when driving in rainy weathers, and a Rain in surveillance (RIS) set collected from networked traffic surveillance cameras on rainy days. From the Table 3, we found that our approach still achieved better results. It could be seen, 3.9% mAP on MPID data set increase was showed by means of MobilenetV1 as the student net, and 3.6% mAP on MPID data set increase was achieved by means of MobilenetV2 as the student net. Figure 5 shows detection examples using Mobilenet without GAN-KD and Mobilenet with different GAN-KD model in comparison.

D. FUTURE WORK

We also use our method to improve the two-stage object detection, such as Faster-RCNN. We employ ResNet50 as student net, and ResNet101 as teacher net. In addition, we compare our detection results with the method proposed by Wang *et al.* [39]. We use PASCAL VOC 2007 train val

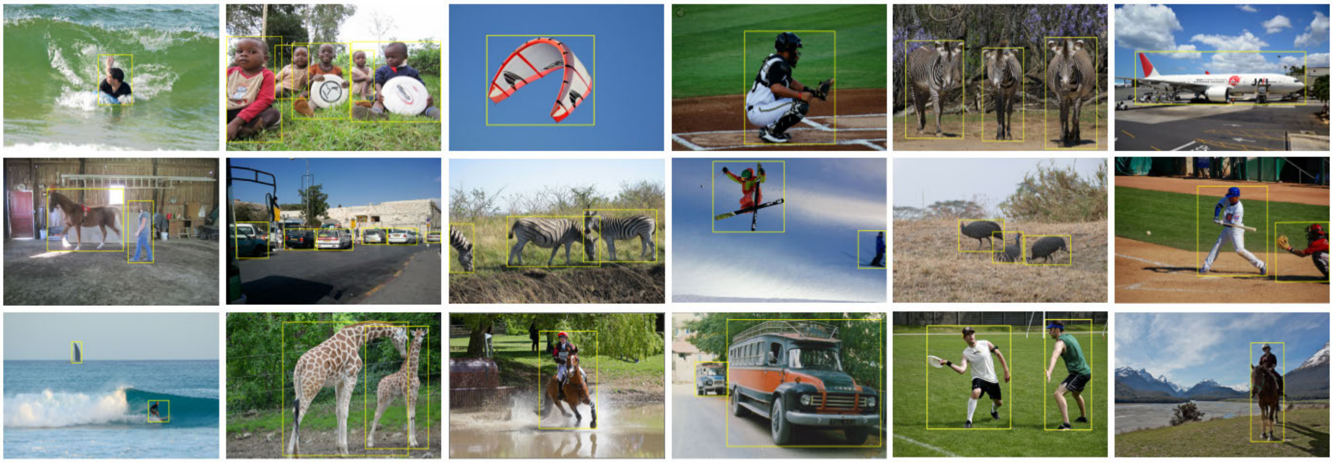
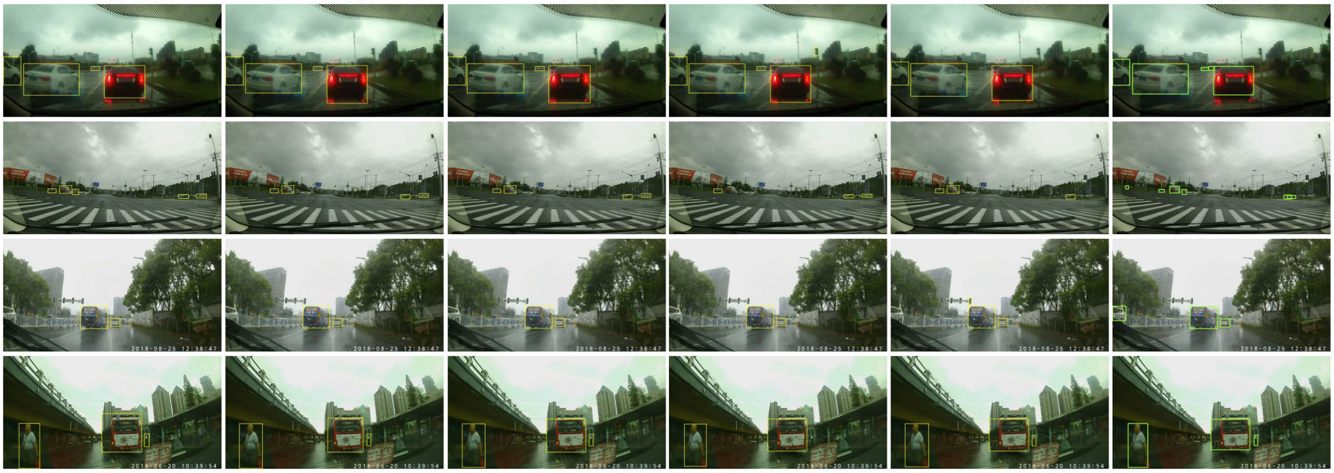


FIGURE 4. Detection results on MSCOCO 2017 with our GAN-KD model. We show detection results with scores higher than 0.6.



(a) GAN-KD(ResNet50) (b) GAN-KD(ResNet101) (c) GAN-KD(VGG16) (d) MobilenetV1 (e) MobilenetV2 (f) Ground-truths

FIGURE 5. Detection results on MPID dataset using MobileNet without GAN-KD and MobileNet with different GAN-KD models on two images (first two rows) from the RID dataset and two examples (last two rows) from the RIS dataset.

TABLE 4. Teacher net is the Faster-RCNN based on ResNet101, and regard PASCAL VOC 2007 train validation as the training set. The mAP is 73.8%+ on the PASCAL VOC 2007 test set. The first row and the second row is our method, and the third row is the method proposed by Wang *et al.* [39].

Student net	VOC 2007 test
ResNet50(ROI Pooling)	67.0
ResNet50(ROI Pooling + GAN-KD)	73.8(**+6.8**)
ResNet50(ROI Align)	71.7
ResNet50(ROI Align + GAN-KD)	74.0(**+2.3**)
ResNet50	69.0
ResNet50(Imitation) [39]	72.0(**+3.0**)

as the training set (without VOC 2012 training set), and the mAP in the PASCAL VOC 2007 test set was 73.8%+.

The shallow student net with knowledge distillation all gets significant improvement. Compared with the 3% absolute gain of Wang *et al.* [39], we get 6.8% absolute gain in mAP for Faster-RCNN based on ResNet50. We also find that Faster-RCNN of ROI Align is 4.5% mAP higher than Faster-RCNN of ROI Pooling in PASCAL VOC 2007 test, as shown in Table 4.

V. CONCLUSION

At present, most of the knowledge distillation methods are aimed at the two-stage object detection. We propose a clean and effective knowledge distillation method for the one-stage object detection. The feature map generated by the teacher net is taken as true samples, and the feature map generated by the student net is taken as fake samples. Then we make the distribution of student net and the teacher net to match each other via true samples and fake samples to perform the generative adversarial training. Through unsupervised learning of generative adversarial training, we avoid manual design and extract feature from teacher net to adapt the student net. Therefore, the whole training process is more efficient, and greatly improve the detection accuracy of the student net. Our approach provides new ideas for the further development of knowledge distillation.

REFERENCES

[1] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," Tech. Rep., 2017.

[2] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[3] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[4] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 116–131.

[5] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," 2015, *arXiv:1510.00149*. [Online]. Available: http://arxiv.org/abs/1510.00149

[6] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng, "Quantized convolutional neural networks for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4820–4828.

[7] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-net: ImageNet classification using binary convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 525–542.

[8] Y. He, X. Zhang, and J. Sun, "Channel pruning for accelerating very deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1389–1397.

[9] Z. Zhuang, M. Tan, B. Zhuang, J. Liu, Y. Guo, Q. Wu, J. Huang, and J. Zhu, "Discrimination-aware channel pruning for deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 875–886.

[10] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*. [Online]. Available: http://arxiv.org/abs/1503.02531

[11] Q. Li, S. Jin, and J. Yan, "Mimicking very efficient network for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6356–6364.

[12] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 742–751.

[13] K. Shmelkov, C. Schmid, and K. Alahari, "Incremental learning of object detectors without catastrophic forgetting," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3400–3409.

[14] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434

[15] 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)*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 21–37.

[17] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[19] 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.*, 2016, pp. 379–387.

[20] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[21] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Neural Inf. Process. Syst.*, 2016, pp. 2172–2180.

[22] M. Mirza and S. Osindero, "Conditional generative adversarial nets," Tech. Rep., 2014.

[23] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural network," in *Proc. Neural Inf. Process. Syst.*, 2015, pp. 1135–1143.

[24] H. Zhou, J. M. Alvarez, and F. Porikli, "Less is more: Towards compact CNNs," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 662–677.

[25] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li, "Learning structured sparsity in deep neural networks," in *Proc. Neural Inf. Process. Syst.*, 2016, pp. 2074–2082.

[26] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2704–2713.

[27] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and <0.5 mb model size," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Feb. 2016.

[28] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," 2014, *arXiv:1412.6550*. [Online]. Available: http://arxiv.org/abs/1412.6550

[29] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," 2016, *arXiv:1612.03928*. [Online]. Available: http://arxiv.org/abs/1612.03928

[30] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4133–4141.

[31] X. Wang, R. Zhang, Y. Sun, and J. Qi, "KDGAN: Knowledge distillation with generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 775–786.

[32] Z. Shen, Z. He, and X. Xue, "MEAL: Multi-model ensemble via adversarial learning," in *Proc. 33rd Conf. Artif. Intell. (AAAI), 31st Innov. Appl. Artif. Intell. Conf. (IAAI), 9th AAAI Symp. Educ. Adv. Artif. Intell., (EAAI)*, Honolulu, HI, USA, Jan./Feb. 2019, pp. 4886–4893.

[33] Z. Xu, Y. Hsu, and J. Huang, "Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, Apr./May 2018.

[34] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 2604–2613.

[35] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.

[36] B. Hariharan, P. Arbeláez, L. D. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 991–998.

[37] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[38] S. Li, I. B. Araujo, W. Ren, Z. Wang, E. K. Tokuda, R. Hirata, Jr., R. Cesar, Jr., J. Zhang, X. Guo, and X. Cao, "Single image deraining: A comprehensive benchmark analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3838–3847.

[39] T. Wang, L. Yuan, X. Zhang, and J. Feng, "Distilling object detectors with fine-grained feature imitation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4933–4942.

WANWEI WANG received the M.S. degree in signal and information processing from the Civil Aviation University of China, in 2010. He is currently the Deputy Director of the Flight Tracking and Surveillance Technology Research Center and an Assistant Director of the Tianjin Key Laboratory for Advanced Signal Processing, Civil Aviation University of China. Since 2010, he has been a Lecturer with the Institute of College of Electronic Information and Automation, Civil Aviation University of China. His current research interest includes signal and image processing.

WEI HONG received the M.S. degree in software engineering from Jiangnan University, Jiangsu, in 2018. From 2017 to 2018, he was an Algorithm Research Intern with Beijing Zeusee Technologies Company, Ltd. His research interests include computer vision and pattern recognition.

FENG WANG received the B.S. and M.S. degrees in software engineering from Tianjin University, Tianjin, in 2015. Since 2017, she has been an Algorithm Researcher with Beijing Zeusee Technologies Company, Ltd. Her research interests include deep learning and object detection.

JINKE YU received the B.S. degree in computer science from the Dalian University of Science and Technology, Liaoning, in 2019. Since 2019, he has been an Algorithm Engineer with Beijing Zeusee Technologies Company, Ltd. His research interests include computer vision and deep learning framework design.