

# When Object Detection Meets Knowledge Distillation: A Survey

Zhihui Li , Pengfei Xu , Xiaojun Chang , Senior Member, IEEE, Luyao Yang, Yuanyuan Zhang,  
Lina Yao , Senior Member, IEEE, and Xiaojiang Chen

(Survey Paper)

**Abstract**—Object detection (OD) is a crucial computer vision task that has seen the development of many algorithms and models over the years. While the performance of current OD models has improved, they have also become more complex, making them impractical for industry applications due to their large parameter size. To tackle this problem, knowledge distillation (KD) technology was proposed in 2015 for image classification and subsequently extended to other visual tasks due to its ability to transfer knowledge learned by complex teacher models to lightweight student models. This paper presents a comprehensive survey of KD-based OD models developed in recent years, with the aim of providing researchers with an overview of recent progress in the field. We conduct an in-depth analysis of existing works, highlighting their advantages and limitations, and explore future research directions to inspire the design of models for related tasks. We summarize the basic principles of designing KD-based OD models, describe related KD-based OD tasks, including performance improvements for lightweight models, catastrophic forgetting in incremental OD, small object detection, and weakly/semi-supervised OD. We also analyze novel distillation techniques, i.e. different types of distillation loss, feature interaction between teacher and student models, etc. Additionally, we provide an overview of the extended applications of KD-based OD models on specific datasets, such as remote sensing images and 3D point cloud datasets. We compare and analyze the performance of different models on several common datasets and discuss promising directions for solving specific OD problems.

Manuscript received 30 April 2022; revised 11 January 2023; accepted 13 March 2023. Date of publication 15 March 2023; date of current version 30 June 2023. This work was supported in part by the Natural Science Outstanding Youth Fund of Shandong Province under Grant ZR2021YQ44, in part by National Natural Science Foundation of China under Grants 61972315, 61973250, 62073218, 61973249, 61902316, 61902313, 62002271, 82150301, 62133012, 62273232, and 62273231, in part by the Young Science and Technology Nova of Shaanxi Province under Grant 2022KJXX-73, in part by the Fundamental Research Funds for the Central Universities under Grant XJS210310. Recommended for acceptance by J. Li. (*Corresponding author: Pengfei Xu.*)

Zhihui Li is with the Shandong Artificial Intelligence Institute, Shandong Academy of Sciences, Qilu University of Technology, Jinan, Shandong 250316, China (e-mail: zhihuilics@gmail.com).

Pengfei Xu, Luyao Yang, Yuanyuan Zhang, and Xiaojiang Chen are with the School of Information Science and Technology, Northwest University, Kirkland, WA 98033 USA (e-mail: pf xu@nwu.edu.cn; rui.monash.liu@outlook.com; yuanyuan.zhan@nwu.edu.cn; 79852780@qq.com).

Xiaojun Chang is with the Australian Artificial Intelligence Institute, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: cxj273@gmail.com).

Lina Yao is with the Data61, CSIRO, Canberra, ACT 2601, Australia School of Computer Science and Engineering, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: lina.yao@data61.csiro.au).

Digital Object Identifier 10.1109/TPAMI.2023.3257546

**Index Terms**—Feature distillation, incremental object detection, knowledge distillation, model compression, object detection, weakly supervised object detection.

## I. INTRODUCTION

OBJECT detection (OD) is one of the most important and practical tasks in the field of computer vision. A large number of methods/algorithms and network models have been designed to detect multiple objects more accurately and quickly [1], [2]. Research into OD began towards the end of the 20th century. For the earlier OD algorithms, various types of artificial features were designed to describe the visual information of the objects and background in the images; classifiers (such as Support Vector Machine(SVM), AdaBoost, and so on) can be used to recognize the objects; object locations in images are then provided by the algorithms. In recent years, following the development of region-based convolutional neural networks (RCNN), several OD models based on deep neural networks (DNN) have been successively proposed [1], [3]. Fig. 1 provides an overview of the milestones in OD and knowledge distillation (KD) research. Although significant performance improvements have been achieved by DNN-based OD models, these models are also increasingly reliant on complex network architectures and large-scale parameters, meaning that the real-time applications of most OD models are limited [4], [5]. This problem has resulted in the development of model compression, which aims to learn compact OD models with fewer parameters, but without reducing model performance.

At present, methods based on model pruning [6], low-rank factorization, quantization, and KD [7], [8] are commonly used for model compression [9]. The former three types of model compression methods focus primarily on how to simplify the network architectures and reduce the parameters with as little performance degradation as possible, while rarely considering knowledge transformation. The latter method (model compression using KD, first introduced by Hinton et al. in 2015 [7]) takes both model performance and efficiency into account and can be used as a novel model training strategy to improve the performance of lightweight networks through the use of knowledge transfer. The relevant knowledge is mined from a large-scale dataset by introducing a high-performance teacher model, which is used as guidance when training a lightweight student model to improve the performance of the latter [10].

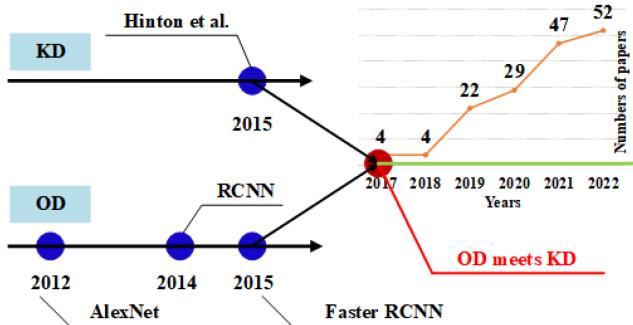


Fig. 1. Milestones of OD and KD. Taking 2012 as the demarcation line, the early OD methods were implemented using artificial features. Later, deep neural networks gradually became the mainstream methods for OD. Since RCNN [40] was proposed in 2014, various two-stage OD models successively emerged, followed by one-stage OD models, such as YOLO [41], DetectNet [42], and so on. At present, most OD models have huge numbers of network parameters and high computational complexity. Subsequently, model compression and knowledge transfer based on KD were proposed by Hinton in 2015 [7], [43]. In 2017, KD-based OD models were first proposed; these approaches tended to be more lightweight than their predecessors [14], [15]. Later, related works increased rapidly, especially in 2021 and 2022. Moreover, KD-based OD models are being constantly optimized and applied to various OD tasks, such as KD-based conventional OD tasks [20], [44], KD-based 3D OD tasks [45], [46], KD-based incremental OD tasks [47], [48], and so on.

In general, the teacher and student models in KD models can have similar or different network structures. Most KD models achieve knowledge transformation by using different types of loss functions or a combination of loss functions, which are calculated based on the visual features or the final outputs of the teacher and student models. In addition, advanced specific KD strategies/techniques have also been proposed for specific visual tasks [11].

KD was initially applied to image classification [7], [12], [13], until several KD-based OD works began to emerge in 2017 [4], [14] (see Fig. 1). For KD-based OD in images, KD is also used to compress the OD models and improve the detection performance of lightweight student models through knowledge transfer [14], [15], which is similar to KD-based image classification. The introduction of KD solved a range of problems affecting various OD tasks; as a result, an increasing number of works on KD-based OD have been presented, especially in 2021 and 2022 [4], [16], [17], [18], [19], [20]. However, there still are great challenges associated with applying KD to OD. For example, OD models involve fewer categories, meaning that less knowledge can be distilled from the category information of teacher models' outputs. OD models may pay more attention to the object locations in images, while category distillation cannot be used to distill localization knowledge. Furthermore, distilling knowledge from multimodal data for OD in 2D/3D images is also a challenging task. Therefore, this survey focuses on presenting and analyzing existing KD-based OD methods. Our goal is to discuss the technological improvements brought by KD to OD, summarize the scientific problems of KD-based OD, and explore possible future research, so as to provide inspiration and incentive to researchers when designing related models.

The existing KD-based OD models mainly solve the problems of model compression and performance improvement, which is

achieved through feature interactions between a teacher model and student model, along with the guidance provided by the soft label information output by the teacher model. Other works have been undertaken to solve new problems, such as S-OD [21], [22], remote sensing OD [23], [24], incremental OD [25], [26], [27], weakly supervised OD [28], [29], 3D OD [30], [31], and so on. In addition, a series of novel ideas or network modules have been proposed to solve a range of OD problems. For example, special information [32], [33], multi-modality [34], [35], multiple teacher models [36], [37], self-feature distillation [38], [39], etc. have been introduced to improve the performance of student models. Overall, KD-based OD has become an important branch of OD in the field of computer vision and has attracted increasing research attention in recent years. At present, a large number of related models have been designed and successfully applied in various areas. However, to date, no reviews have been conducted to summarize and analyze the current relevant research on this topic. Therefore, our paper will provide the first comprehensive literature survey of KD-based OD models and highlight several possible future trends.

This paper is organized as follows. We first classify the related KD-based OD methods in Section II. A detailed description, summary, and analysis of the different types of existing methods is then provided in Section III. Several related benchmark datasets and evaluation parameters of model performance are provided in Section IV, along with the performance comparison of different models. The relevant problems that need to be solved and suggestions for future research on KD-based OD are discussed in Section V. Finally, Section VI concludes this paper.

## II. THE CATEGORIZATION OF KD-BASED OD METHODS

OD methodologies can be categorized in various ways, such as one-stage and two-stage OD models, or fully, weakly, and semi-supervised OD models, and so on. Similarly, KD-based OD methods also have their own categorization. In this section, we list these methods in categories according to the related OD tasks and KD strategies employed (see Table I), including KD-based OD models for conventional OD tasks and solving specific OD problems, and we also elaborate these methods according to different KD strategies. In the below, we first give a brief statement of these methods in main categories, and a detailed analysis and description of the corresponding methods are given in Section III.

### A. KD-Based OD Methods Based on Different OD Tasks

#### (1) Novel KD-Based OD Models for Conventional OD Tasks

The conventional OD models are mainly used to detect large numbers of common objects in natural images, and KD is introduced to obtain lightweight student models with better performance. However, many novel ways to use KD-based conventional OD models have also been devised. For example, some methods use an adversarial learning-based strategy to make the student model more accurately learn knowledge from the teacher model, while other methods have designed relevant

TABLE I  
THE CATEGORIZATION OF RELATED KD-BASED OD METHODS BASED ON OD TASKS AND KD STRATEGIES

Methods based on different OD tasks	Conventional OD tasks	Traditional OD model compression using KD Personalized KD-based OD models
	Solving specific problems	S-OD and light-limited object detection KD-based OD in remote sensing images KD-based incremental OD KD-based 3D object detection Video-based OD by introducing KD KD-based weakly supervised OD The extended OD tasks based on KD
Methods based on different KD strategies	-	Conventional methods using distillation loss and soft labels Feature distillation Various network structures of teacher-student models Multiple teacher models Self-feature distillation Specific information guidance for OD

loss functions to reduce the gap between the student and teacher models.

#### (2) KD-Based OD Models to Solve Specific Problems

Several KD-based OD models have been proposed to solve specific OD problems. For example, high-resolution images are used to distill knowledge in the teacher model in order to solve the problem of S-OD, and multiple-modality information is introduced into teacher models to improve the OD performance of the student models under low light conditions. In addition, the KD branches of the attention mechanism, using high-resolution images and feature distillation modules between different network layers, are exploited to solve the problem of multi-scale OD in remote sensing images. Furthermore, new categories have emerged, and the problem of catastrophic forgetting during model updating has also been taken into consideration; related incremental KD-based OD methods have additionally been proposed by introducing pyramid networks and prior knowledge. KD has also been introduced for 3D OD and video-based OD tasks, yielding better performance. Moreover, weakly supervised OD based on KD is also a specific OD task. While the existing related works on this subject mainly focus on how to improve model performance using different KD strategies, less attention has been paid to model compression. Therefore, we review the relevant works on how to introduce semantic information, utilize unlabeled images, and design related strategies to improve the corresponding OD model performance. In addition to the aforementioned KD-based OD models, there are many other extended OD tasks based on KD technologies, including relationship detection, human-object interaction (HOI) detection, lane detection, face detection, person search, object segmentation, etc.

#### B. KD-Based OD Methods Based on Different KD Strategies

The KD strategies used in visual tasks vary widely. The novel and optimized KD strategies can significantly improve the effects of knowledge transfer and knowledge learning. In these KD-based OD methods, a variety of advanced KD strategies have been proposed to improve the performance of OD models. For example, some methods have been developed that employ feature distillation of knowledge transfer at different network

layers of the teacher and student models, and various network structures have been designed that enable teacher and student models to learn the features from multimodal data. In addition, similar to human teaching activities, the performance of student models can be improved by designing multiple teacher models to jointly or gradually train one student model. Teacher and student models can also learn visual features from each other, or perform self-feature distillation using various loss functions. Finally, different types of prior knowledge (object masks, semantic context, textual information, etc.) can be used as a guide for training lightweight student models.

### III. KD-BASED OD METHODOLOGIES

#### A. The Basic Principles

KD has been introduced into OD models based on deep learning, and has been used to improve model compression and model performance. A common way to distill knowledge is to use the different types of distillation loss to guide the student models in learning the knowledge of the teacher models. The general loss function for KD-based OD models can be expressed as follows:

$$L = L_{det} + \gamma L_{dis}, \quad (1)$$

where  $L_{det}$  is the loss function for OD,  $L_{dis}$  is the loss for KD, and  $\gamma$  is a weight coefficient. Fig. 2 presents the basic framework of KD-based OD models. Here, the output of the teacher network is used as soft labels for the student network. Various feature distillation modules have been proposed for specific OD tasks. However, a survey of the the models or algorithms employed in many relevant existing works reveals that the basic network frameworks continue to be designed based on OD models [49], [50], [51], [52]. Therefore, we derive the basic principles of KD-based OD models through analyzing the methods proposed by introducing KD technologies into traditional two-stage and one-stage OD models.

At present, two-stage OD models take Faster R-CNN as the mainstream network framework, while YOLO and SSD are classic and widely used one-stage OD models [16]. Therefore, the current public KD-based OD models are explored on these basic OD networks. As shown in Equation (1), the basic principle

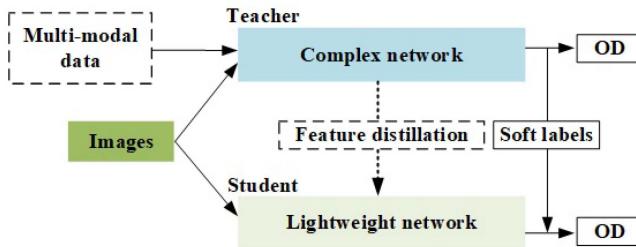


Fig. 2. The basic framework of KD-based OD models. Conventional KD-based OD methods are largely applied to 2D images and distill the knowledge using soft labels. We represent the framework of these methods using solid lines and boxes. Moreover, some improvements to existing methods have been made by introducing multi-modal data or feature distillation (marked with dotted lines/boxes) to the conventional framework.

of these early KD-based OD models lies in the distillation loss functions. The total loss function of KD-based OD models on both two-stage and one-stage OD networks can be summarized as follows:

$$L = \alpha L_{class} + \beta L_{reg} + \gamma L_{hint}, \quad (2)$$

where  $L_{class}$ ,  $L_{reg}$ , and  $L_{hint}$  represent the classification loss function, the regression loss function used to determine the bounding box, and the loss function of hint feature distillation in KD-based OD models, respectively. Notably, Equation (2) presents only the basic form of a distillation loss function; numerous different types of loss functions have been proposed to deal with the problems encountered by existing KD-based OD models.

In addition, various types of distillation modules, mechanisms, and strategies have also been explored in an attempt to improve the performance of KD-based OD models. For instance, there are multiple prior information guided-modules [21], multi-teacher networks [37], [53], [54], multi-modal data distillation networks [14], [35], [55], self-distillation networks [38], [56], [57], [58], weakly supervised OD networks [28], [29], [59], [60], and so on. These will be elaborated upon and discussed in more detail later in this paper.

## B. Challenges

### (1) The Balance Between Model Compression and Performance

KD is an effective model compression technique, and effective knowledge transfer helps with the construction of a lightweight network model; such lightweight models are more suitable for practical applications with particular model efficiency requirements and device performance limitations [4], [61], [62]. Complex networks have large-scale network parameters and a long running time, making them unsuitable for such practical applications [63]. However, it is also challenging for lightweight networks to learn effective visual features from large-scale complex datasets [4]. A balance must therefore be struck between model compression and model performance improvement.

### (2) The Imbalance Between Multi-Modal Data Features

At present, most KD-based OD models focus on transferring knowledge within the image domain; only a few works have attempted to extract additional features from other modal data (infrared images, depth images, text, index data, etc.) [19], [35], [55]. The introduction of multi-modal features is beneficial to KD. However, a more challenging problem is that of how to deal with the imbalance between multi-modal data features. Here, “imbalance” means that significant differences exist between the feature dimensions and semantic information of different modalities. For example, the dimensions of features extracted from visual data and index data were significantly different. In addition, the semantic feature gaps between the textual and visual features of RGB images also make it difficult to use text information to guide the visual feature learning of student models. Therefore, another challenge is that of combining imbalanced multi-modal features to guide student models for feature learning [19], [64]; new feature fusion mechanisms or multi-modal information-guided mechanisms need to be designed for student model learning.

### (3) Designing or Selecting Superior Teacher and Student Models

KD technology is utilized to transfer the knowledge learned by complex teacher models to lightweight student models. An optimal teacher model or model combination has a very favourable influence on guiding the feature learning of a student model. Therefore, the selection of teacher and student models makes a very important contribution to the performance of the final student models [53], [65]. However, as there are many complex and lightweight models for OD, it is difficult to choose the appropriate teacher and student models for specific OD problems.

## C. Novel KD-Based OD Models for Conventional OD Tasks

For conventional OD tasks, researchers attempt to design OD methods/models with superior performance and optimized model structures. Therefore, KD is used to simplify traditional object detection models—which typically have large numbers of network parameters—while still ensuring high detection accuracy. In addition, various types of feature and prior knowledge are extracted from multi-modal data by KD technology to continuously improve the performance of OD models.

**1) Traditional OD Model Compression Using KD:** At present, CNN-based OD models achieve superior performance, and increasingly complex network models with more network parameters have been designed to improve detection performance. However, these models are not suitable for practical applications or running on embedded devices due to performance limitations. Therefore, KD technology and lightweight OD models capable of achieving reliable performance are good choices for solving these problems. When KD was initially introduced, very few works applied KD to OD [66]. Around 2017, many KD-based OD models began to be proposed, with the primary motivation of developing a lightweight network model. Li et al. proposed a feature mimic architecture, which can be viewed as a special case of the KD-based OD framework [49]. Most of these early methods are designed to obtain compression models

through feature distillation or soft label distillation between teacher and student models [44], [63], [67], [68], while some other applications use KD for network pre-training [69].

Some KD-based OD methods have also been devised to solve the problem of category imbalance by designing different category-balanced focal loss functions [50]. Multi-scale feature distillation loss functions have additionally been proposed to handle the task of multi-scale OD [16], and multiple KD has also been explored to empower the student model to learn features from multi-level feature maps, enabling the extraction of both low-level details and high-level abstractions simultaneously [70]. In 2017, Chen et al. [4] proposed a KD-based detection model, which may represent the first attempt to introduce KD into multi-class OD tasks. Two types of loss functions (a weighted cross-entropy loss and a teacher bounded loss) are utilized to address class imbalance and the regression component respectively. In addition, KD can also be applied to multi-object tracking. For example, in [71], an end-to-end KD framework implements multi-object tracking through a multi-task network with a shared backbone network.

In addition, the network structures of these models have many similarities. For instance, KD using the soft labels of the teacher models is the most common method and has been employed in almost all KD-based OD models. Feature distillation between the convolution layers of the teacher and student models has also been introduced to improve OD performance.

Different from image classification, OD involves not only the category of the objects, but also the object location labels. Therefore, Label Assignment Distillation (LAD) [72] and Localization Distillation (LD) [73] were explored to further improve the accuracy of object location. LAD [72] is a simple and effective KD method capable of being applied to most object detectors. LAD makes a student model indirectly learn the knowledge from teacher models using the teacher network to generate and assign soft labels, which is a significantly different approach from that adopted by traditional KD methods. Notably, localization ambiguity is widespread in OD tasks; hence, Zheng et al. [73] introduced distillation learning into the localization branch of a OD network, which uses LD to improve object location. To make use of the objects' location information, the bounding box distributions generated by a teacher network are distilled to the student network. Therefore, to achieve location information distillation, Zheng et al. [74] developed the concept of the valuable localization region (VLR), which is different from the main distillation region determined by label assignment. These authors designed an algorithm to obtain these VLRs, then carried out localization distillation by region weighting. Similarly, to solve the problem of traditional KD being less effective at distilling 1-bit detectors due to the neglect of the information discrepancy between the teacher and 1-bit student model, an information discrepancy-aware strategy (IDa-Det) [75] is proposed. IDaDet can be used to select the representative proposal pairs according to information discrepancy, and are beneficial for distilling 1-bit detectors. Furthermore, Yang et al. [76] considered the difference between the features in focal and global regions, then unequally distilled these different features to the student

model. The focal and global feature maps are obtained from the neck of the teacher model, and only the distillation loss calculated on feature maps is used for feature distillation. Also using soft labels, Umer et al. [77] proposed a Pseudo-KD (PKD) training method for saliency prediction. PKD utilises a classical KD architecture, and the pseudo-labels provided by the teacher network are used to distill the predicted saliency map to the student network.

The aforementioned methods use a single teacher model to conduct KD for a single student model. Subsequently, KD-based OD models that use multiple teacher models or student models to achieve mutual KD were introduced. For instance, collaborative learning is used to conduct mutual KD among multiple student models [78]. This method can continuously improve the generalization ability of different student networks. It can also be further extended to other visual tasks and used to optimize collaborative learning for multiple related tasks.

**2) Personalized KD-Based OD Models:** In recent years, to improve the performance of OD, researchers have implemented multiple expensive deep networks, which require a lot of storage space and consume significant running time. The complex design of these networks also makes them difficult to apply to real-world scenarios, such as performing OD on devices with limited resources. One effective way of solving these problems, an effective way is to assemble large-scale networks into lightweight networks via model compression [51]. The lightweight student model learns the knowledge of multiple teacher models and compresses the model using the KD strategy to alleviate the memory problem. Notably, it also takes much less time to assemble the implicit knowledge from multi-teacher models. In particular, the so-called assembled lightweight network [51], supervised by the adversarial learning-based strategy, improves the reasoning speed. The methods proposed in [79], [80] take the feature maps generated by the teacher and student models as the true and false samples respectively, and conduct adversarial model training between the teacher and student models to improve the performance of the student model during single-stage OD. There is another way to undertake single-stage OD, and to improve performance by adversarial training [79], [81]. Sometimes, if the data for teacher model training is not available, this will result in a missing data problem. Depth inversion is a method of synthesizing data from neural networks to solve this problem. The method proposed in [81] is a KD-based OD network designed to tackle the problem of missing data. This method synthesizes images from a pre-trained model through a model inversion process called "DIODE," then carries out KD for OD on these synthesized images.

In addition, to achieve fast OD, the KD strategy can be applied to transferring the knowledge obtained by large-scale and slow teacher models to smaller and faster student models [82]. To solve the problem of useful knowledge being unevenly distributed in complex networks, instance-conditional KD [83] is proposed to locate the useful knowledge, as well as to decode the network description of knowledge retrieval with an auxiliary optimization of knowledge retrieval.

To further optimize OD networks, we can start from the teacher-student relationships. From the perspective of the rigor of teacher network training, strict teacher networks have higher accuracy. Moreover, researchers believe that teachers should be more tolerant (less peak confidence distribution). Therefore, by adding an extra loss into the teacher network, the student network can better learn the similarity between categories, which helps to prevent over-fitting [84]. In addition, it is known that the predictions of neural networks often lack interpretability; however, we sometimes need these explanations to understand and become able to trust the decisions of certain systems, such as driver assistance systems. At present, researchers have attempted to explain the predictions of neural networks with reference to semantic concepts. However, these methods change the basic networks, which affects the model's performance. Therefore, we need to provide meaningful explanations without modifying the underlying network. In an attempt to achieve this, a post-explainable black-box model [85] was developed, which may be the first explainable method of generative KD that does not impact the original performance. Furthermore, Lee et al. [86] identified the problem of feature imbalance between teacher and student models, which is often ignored by traditional KD-based OD models. This led to the development of the Shared Knowledge Encoder (SKE). Under this approach, multi-layered features of the teacher and student models are input into SKE for unified encoding, after which an auxiliary decoder is adopted for feature decoding to facilitate knowledge distillation with balanced features.

#### D. KD-Based OD Models for Solving Specific Problems

In addition to the conventional OD tasks, there are also more specific detection tasks and related works, such as S-OD, OD in remote sensing images, incremental OD, 3D OD, video-based OD, weakly supervised OD, and other extended OD tasks. In this subsection, we will provide an overview of these specific detection tasks, which are organized into categories.

1) *Small Object Detection and Light-Limited OD*: S-OD, which is an important branch of OD research, is uniquely difficult because the objects to be detected occupy very few pixels in the image, which makes accurate detection challenging. Traditional S-OD models include the method based on an image pyramid, the method that increases the number of small objects, the method of multi-scale feature fusion, and so on. However, in the OD model based on KD, the problem of S-OD is solved by the feature pyramid method. The distillation model transfers the features learned from a high-resolution object to a S-OD model to improve the S-OD performance. For example, Deng et al. [21] proposed an extended feature pyramid network (EFPN) with an extra high-resolution pyramid level specialized for S-OD, and further designed a pivotal feature reference-based super-resolution (SR) module, named feature texture transfer (FTT), to endow the extended feature pyramid with credible details in order to facilitate more accurate S-OD. FTT enables EFPN to learn more reliable details, while the cross-resolution distillation strategy introduced supports the student model in learning the teacher model's perception of object details. Another strategy

to facilitate small sample detection is to increase the number of samples detected [87]. In addition, Zhao et al. [87] further improved the performance of S-OD by introducing FPNLite, which can ensure that the network remains focused on more detailed information at the bottom of the small object. However, the KD method they used is mainly for optimizing model performance, and most of these methods use conventional distillation strategies. However, fewer methods take feature distillation between teacher and student models into account; this may lead to significant performance degradation of student models after the model is made lightweight.

In the context of OD tasks, another problem is performing an OD task when the object has been captured in a weak light or under poor conditions. To solve these problems, the KD-based OD method introduces data of other modes (such as photo-abundant images [22], infrared images [14] and depth images [55], etc.), so that the model can learn more auxiliary detection information, which improves the detection performance. A new non-local feature aggregation method is introduced to KD-based OD to facilitate OD for low-light images [22]. This approach uses photo-abundant images as the input of the teacher model and regularizes the features through the KD of the teacher-student model. The student network can imitate the pre-trained teacher model to learn more effective non-local feature information. In addition, thermal images [14] and depth images [55] are also introduced into the template detection model of KD. The acquisition of the depth image is not affected by light, meaning that the depth image can assist the network in obtaining effective information in the modal data of the depth image in the OD of the weak light image [55] to improve the OD performance. Different types of modal data are introduced into the KD-based OD framework, and multimodal data can be used to learn more comprehensive and effective information about the object to be detected. For example, in [35], four types of multi-modal data (RGB, heat map, depth image, and audio) are simultaneously introduced into the OD framework based on audio data. A self-supervised MM-DistillNet framework is then introduced; this framework consists of multiple teachers that leverage the diverse image modalities to simultaneously exploit complementary cues and distill knowledge into a single audio student network, whereas the audio student network detects and tracks objects in the visual frame using only sound as an input. Therefore, the work in [35] proposes distillation learning OD based on audio signals rather than images, making it unique.

2) *KD-Based OD in Remote Sensing Images*: To date, a range of OD methods designed for remote sensing images have been developed. KD has also been introduced into this field to solve the problems of multi-scale, multi-resolution, and multi-direction OD.

First, a common problem encountered when performing OD in remote sensing images relates to the characteristics of multi-scale objects. To deal with this problem, the related KD-based OD models use large-scale remote sensing images with clear or high-resolution objects as the training data for teacher models, while the training data used for student models is the remote sensing images with small-scale or low-resolution objects. Some relevant KD-based OD models carry out feature distillation

in different layers [16], [24], [88], some corresponding loss functions are designed for multi-scale feature distillation at different convolution layers to facilitate KD-based multi-scale OD [24], and some feature distillation modules and corresponding loss functions are designed to carry out knowledge transfer between multiple convolution layers of the teacher and student models [89]. In addition, an OD distillation model based on a feature pyramid network is also a good method to solve the object scale inconsistency problem in remote sensing images. For example, the method proposed by Chen et al. [23] deals with the problem of incremental OD, which requires dealing with the appearance of new objects in remote sensing images [90], and also designs a feature pyramid network to solve the problem of multi-scale objects existing in remote sensing images. The pyramid network is also applied for detecting multi-scale objects in UAV (unmanned aerial vehicle) images [91]. In general, large-sized objects can obtain better activation values in networks. Therefore, Liu et al. [91] carried out interactive feature learning at different scales between teacher and student models, and designed a position-aware L2 loss function to facilitate the joint training of teacher and student models. In this way, the cross-scale features of large-sized objects can be introduced into student models. In addition, feature distillation modules and corresponding loss functions are designed to carry out knowledge transfer between multiple convolution layers of the teacher and student models [89]. These distillation modules comprehensively consider the multi-scale features of multiple convolution layers in the teacher model and distill them to a certain layer of the student model, enabling achieve multi-scale feature distillation to be achieved.

The second problem related to remote sensing objects detection is OD in multiple directions. For example, the work in [92] contends with the issue that the student model can detect objects in multiple directions and angles. An auxiliary loss function is used to support the student model in learning the visual features of multi-directional objects processed in the teacher model. In addition, Chen et al. [9] attempt to solve this problem by introducing an asymmetric convolution module (ACM) into YOLOV3, and proposed Tiny Yolo-Lite. Tiny Yolo-Lite is compressed from YOLO by network pruning and KD; here, the KD is designed to compensate for the performance degradation caused by network pruning, as well as to improve the generalization performance of the model. Similar to the model compression method employed in [9], the method in [6] also uses model pruning for model compression and KD for knowledge transfer to improve the model performance. These two methods employ a similar adaptive pruning strategy for model pruning, along with a different pruning strategy for network pruning. For the methods in [93], [94], the relevant networks are designed to solve the problem of multi-directional OD in remote sensing images. The common KD technique is used to obtain a lightweight OD model through the use of feature and prediction distillation loss functions.

Furthermore, the self-attention mechanism is introduced into the distillation learning model. For instance, Chai et al. designed a new method, called bidirectional self-attention (Bi-SAD) [18], to deal with automatic cloud detection in remote sensing images.

Bi-SAD conducts feature distillation at different levels, and can extract the texture and semantic information of the cloud through self-distillation learning, thereby improving the cloud detection performance in remote sensing images. To optimize the lightweight remote sensing object detectors, two distillation modules are proposed to distill the multiscale features and provide more precise regression results to the student model [95].

However, although cross-domain OD may sometimes occur in remote sensing images, few existing methods consider how to use KD technology to solve this problem [90]. In addition, incremental OD tasks often appear in the context of remote sensing images, while the detection performance of most existing methods used to perform these incremental tasks is largely dependent on the similarity of visual features between the samples of old and newly added tasks. Finally, few existing KD algorithms take the regression branch into account. If more knowledge distillation strategies are proposed to optimize the regression branch, the performance of the lightweight student models will improve [6].

3) *KD-Based Incremental OD*: In recent years, with the increasing application of deep learning methods, the performance of OD has been significantly improved. However, traditional OD methods require the annotation of a large amount of image data; thus, when a new category of object appears, the original model cannot detect this new category because there is no annotation information available. A simple way to resolve this new object category detection problem is to annotate the related data of this new category, then train the whole network using the whole dataset after adding the annotated data; however, this approach is very computationally costly. There is another way to fine-tune the original model, namely by applying incremental learning, which is called incremental OD. However, there exists a problem by the name of "catastrophic forgetting," in which the model learns the features of new categories of objects, after which the detection effect of old object categories will fall sharply after model fine-tuning. The reason is that the network weights, network parameters, and relevant features extracted by networks tend to detect new categories of objects, but "forget" the feature information of old object categories. We can accordingly mitigate catastrophic forgetting by using a KD strategy that migrates knowledge from the initial network to the incremental network [96], [97].

Fig. 3 presents the basic model structure of KD-based incremental OD. To alleviate the forgetting phenomenon in incremental OD, incremental OD models based on Faster R-CNN are proposed in [98] [26]. To reduce feature forgetting by preventing changes to the predictions of older categories of objects, Liu et al. [47] used a continuous general-purpose detector, which enables continuous learning from different areas without forgetting. An end-to-end incremental learning OD model is also realized by designing several loss functions and a loss of confidence so that the old data information is retained to the maximum extent. The class-incremental Faster R-CNN (CIFRCN) model [98], which can dynamically add new categories with a small amount of labeled images, was subsequently proposed. In CIFRCN, the prospect domain of RPN is extended to generate precise boundary box proposals. The classifier in Fast R-CNN is

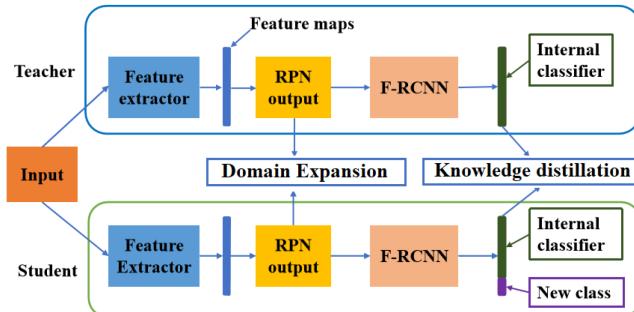


Fig. 3. The structure of KD-based incremental OD.

extended via KD to improve the accuracy of proposal classification. Similarly, Dong et al. [99] also introduced the commonly used KD strategy to solve the problem of catastrophic forgetting in the incremental OD model.

Attentional characteristic distillation (AFD) is proposed to take advantage of attentional distillation while avoiding the disadvantages of both methods: that is, using both top-down (TD) and bottom-up (BU) attentional mapping to filter noise information. Bottom-up attention captures the key context information of the input image, which is used to guide the current model to imitate the attention features of the previous model. Furthermore, a new incremental object detector based on Faster R-CNN, which can continuously learn the features from new categories of objects without using old data, is proposed in [26]. In this method, with the aid of both old models and residual models, a triple network is designed to help incremental models learn the objects' visual features in new stages without forgetting the existing knowledge. Similarly, Yang et al. [100] proposed a dual network, consisting of old models and incremental models, and try to transfer the knowledge learned from the old categories of objects to the new model to detect new object categories.

The strategy of combining incremental learning and KD can also be applied to OD in remote sensing images [101]. Beginning from the intuition that remote sensing images have different sizes and rich categories, an incremental OD architecture based on the feature pyramid network (FPN) and KD is proposed in [23]; this may be the first research work on incremental OD in the remote sensing field. This method uses FPN to detect objects of different scales, adds a new branch to the last layer of FPN to detect new categories, and carries out KD on the output of the old branch. By implementing these operations, the model maintains the ability to learn the features from older categories. In addition, vending machines are also an important application. For example, in [102], a large retail image data set for incremental OD is built and made public, and an incremental object detector (FCIOD) based on ftP-RCNN is proposed. An image-level example management strategy is introduced to prevent forgetting in incremental learning. The sample set size is fixed: the model can access up to  $K$  images of previous training data.

KD helps the models to reduce forgetting, but it also slows the rate of adaptation to new categories. To solve this problem, Joseph et al. [27] proposed a meta-learning method for OD.

In this method, the feature information is automatically shared across incremental tasks through prerequisite gradient updates during model learning. The proposed method not only helps the model to retain old knowledge, but also has the flexibility to adapt to new categories. The meta-learning incremental object detector performs better than other current methods. It formulates a new loss function that combats intransigence due to KD by learning a general set of gradient directions, thereby alleviating the problem of "catastrophic forgetting" and improving the models' adaptability. In addition, more related OD tasks are employed to share their effective visual features in [103]. These shared features are selected by using task-aware gates among these different OD tasks for specific tasks. Therefore, these methods discussed above essentially attempt to use the knowledge extracted by teacher models to avoid student models significantly forgetting the previous objects. However, these models (excepting the method in [104]) rarely take the possibility of overconfidence in the teacher models into account; as a result, some incorrect prediction results of the teacher models may lead to incorrect guidance for the student models. Verwimp et al. [104] used current ground-truth labels to judge the predictions of teacher models, then selectively guide the knowledge transfer for student models.

Although KD technologies can alleviate the problem of detecting new categories of objects in incremental OD task to a large extent, there are still some limitations in related existing methods. First, the ability of KD to avoid catastrophic forgetting is significantly diminished as the number of new categories increases [97]. Second, if there are only a small number of objects in the new categories, then detecting these new objects at an optimal level of performance will be a challenging task [23]. Finally, few existing methods consider dealing with both the catastrophic forgetting problem of the models and improving the performance of detecting objects in known categories.

4) *KD-Based 3D OD*: In recent years, KD has also been widely used in 3D OD. Most KD-based 3D OD models aim to obtain lightweight 3D models using KD [105]. For example, Felix et al. [30] presented a method to reduce the complexity of the 6-DOF detection network while maintaining accuracy. These authors used KD to teach a lightweight convolutional neural network (CNN) and learn from a real-time 6-DOF detection CNN. This method allows real-time applications to use only RGB images while reducing hardware requirements. Moreover, to compress the network parameters of the student model, a shared autoencoder is designed by taking the overall network parameters of the teacher and student models as the input parameters [106]. Subsequently, combined with a compressed representation loss, the compression feature parameters shared by the teacher and student models are extracted. KD has been applied not only to image-based 3D OD, but also to the optimization of point clouds based on a 3D OD method. Wang et al. [107] designed a two-stage training approach for point cloud OD. First, they trained an OD model on a dense point cloud, which is generated by multiple frames and uses the additional information that is available only during training. They then trained the same counterpart of the model on the sparse single-frame point cloud, and uniformly regularized the characteristics of the two models.

It was proven that this process improves the performance of low-quality data without introducing additional overhead. In 3D OD tasks, we are often required to solve the problem of inconsistent scales of objects. Similarly, this problem also exists in 3D OD tasks. Therefore, KD-based 3D OD methods are designed to distill the 3D visual features with high-resolution voxels to the student models, which deal with 3D objects with low-resolution voxels. For example: in the method of LiDAR distillation [108], 3D point cloud data containing voxels with different resolutions are divided into beams; the low-beam point clouds can be obtained by downsampling the high-beam point clouds. Multiple teacher models with 3D network backbones are then used to perform KD from the higher-beam point clouds to the student model dealing with the low-beam point clouds.

In 3D OD, the point cloud is used as the detection data. If the corresponding 2D image is considered at the same time, the detection effect is improved, and multimodal data will be helpful to improve the model. Qin et al. [31] considered using 2D images for KD, and proposed a method of detecting the point clouds of 3D objects with weak supervision. Although this method does not require a ground truth 3D bounding box for training, it needs to make full use of common data formats (paired images and point clouds) for weak supervision. Moreover, in the process of performing KD from 2D images to a 3D point cloud, Qin et al. proposed a cross-modal transfer learning method. This approach regards the detection network based on point cloud as the students, learns knowledge from the existing pre-trained image recognition network, and then employs the teacher network as the medium to transfer knowledge from 2D images to the object point cloud domain, which can reduce the labeling cost of 3D OD on unlabeled datasets. Similarly, Sautier et al. [46] used the object segmentation and detection results in 2D images based on self-supervised learning to obtain 2D image expressions, after which superpixel-driven contrastive distillation is designed and applied on a 3D OD task from 2D images to 3D point clouds. This method does not require the 3D point cloud data to be annotated, which greatly reduces the cost of data annotation. The methods in [31], [46] employ the KD-based 3D OD pattern of 2D-3D. Other methods proposed in [45], [109] try to generate the required 2D images based on three-dimensional point cloud data (LiDAR), then use these generated 2D images to train network models. The 2D visual representation of the objects is then distilled to 3D feature space through feature-level and object-level distillation, thereby facilitating more efficient 3D OD.

Most existing KD-based 3D OD methods use KD technology to solve the problem of 3D model compression, and other modal data is also introduced to reduce the high cost of labeling 3D objects. However, in 3D OD tasks (especially for automatic driving), complex and changeable 3D scenes have always been a significant challenge for 3D OD. Therefore, using multimodal data and design-optimized KD strategies to obtain lightweight and high-performance 3D OD models will remain a very challenging field of research in the future.

*5) Video-Based OD by Introducing KD:* Videos are more complex than images. To date, few research works have been conducted in this field, and there is a lot of scope for introducing

KD to video-based OD and related tasks. Because video data contains spatial visual information and temporal information, it is particularly important to obtain the temporal information contained in videos. Some researchers learn from the human visual system (HVS), which relies heavily on the temporal dependence between visual input frames, to effectively identify the objects in videos. In 2019, Farhadi et al. [110] proposed a new framework called temporal knowledge distillation (TKD), which can distill the temporal knowledge from a heavy neural network over selected video frames into a light-weight model.

When a human visual system is mentioned, it inevitably brings to mind salient OD (SOD), which simulates a human visual perception system to locate the most attractive object in the scenes. This approach has been widely used in various computer vision tasks, but most are applied to the image level [111]. In 2020, Piao et al. applied KD to SOD and proposed a depth distiller (A2dele) [34], which is mainly inspired by KD and privileged information [112]. Through A2dele, the lack of privileged depth knowledge can be transmitted to an RGB stream. The transmitted knowledge is divided into two parts: (1) the first part is designed to achieve the expected control of pixel depth knowledge transmitted to the RGB stream prediction; (2) the second part is designed to transform the location knowledge of significant objects into RGB features. Therefore, by embedding A2dele during training, a lightweight architecture without a deep stream can be realized [34].

In the latest research works, based on the unique characteristics of video data, Tang et al. proposed a lightweight network for SOD in videos [113], in which KD is carried out for spatial and temporal dimensions respectively. Specifically, the spatial dimension combines a saliency-guided feature embedding structure and spatial KD to refine the spatial features. In the temporal dimension, a temporal KD strategy is proposed by introducing an attention-based inferred frame feature coding module. In combination with temporal distillation, the sequence information is fully propagated from adjacent frames to the current frame. This strategy allows the network to learn robust temporal features by inferring frame feature coding, as well as to extract the features from adjacent frames.

*6) KD-Based Weakly Supervised OD:* In traditional OD tasks, the cost of data annotation is very high, which has motivated the development of weakly supervised OD (WSOD). In recent years, some researchers have applied the KD strategy to WSOD and semi-supervised OD to further improve the accuracy of models by improving the quality of the extracted features. Typical applications include the Class-Aware Object Detection network (CADN) proposed by Zhang et al. in 2021 to solve the problem of surface defect detection [28]. The network uses only image markers for training, image classification, and defect location simultaneously. To coordinate real-time performance and accuracy, a KD strategy is adopted to ensure that the lighter CADN has similar characteristics to the larger CADN, meaning that the lighter CADN can maintain high real-time performance while improving accuracy. A new semi-supervised OD formula proposed in [114] uses a small number of seed box-level annotations and a large number of image-level annotations to train the detector. For their part, Cao et al. [115] also found that

the features of tiny abnormal objects learned by the teacher and student models exhibited great similarity. Semi-supervised learning is then introduced into the tiny OD based on weakly supervised learning, and the ground-truth labels of a small number of abnormal samples are used to increase the feature differences between abnormal and normal objects, which improves the performance of abnormal tiny OD. Notably, a large amount of label noise is introduced in the process of data mining, which has a severe negative impact on the test results. Accordingly, a new anti-noise integrated RCNN (Note-RCNN) object detector was proposed to deal with the noise. Two classification heads and a set of distillation heads are used to avoid the hazards of over-fitting noise labels and false negative labels. The training box regression head only marks seeds, eliminating the harm of the inaccurate excavation box boundary. In addition, the works in [59][60] also conducted a related study. By separating teacher data from ground truth data, the researchers solved the problem of the previous OD refinement strategy being too dependent on the availability of the teacher model and the ground truth tag. The key to this approach is that detection based on observed objects involves non-maximum suppression (NMS) steps.

However, the accuracy of weakly supervised OD still needs to be improved. The problems are as follows: First, because only image-level tags are available, the WSOD detector tends to detect more significant objects and different parts of objects; second, the object location information in WSOD is seriously insufficient. The first problem is addressed in [29], in which adaptive training strategies and bounding box regressors were designed using an object-level refinement of the WSOD2 framework, combined with bottom-up object evidence and top-down classification reliability scores. Similarly, a weakly supervised OD model that also focuses on the partial features of combined objects employs the common KD strategy to obtain a lightweight student network for combined OD [116]. The comprehensive attention self-distillation (CASD) mechanism proposed in [117] can also be used in this context. To balance feature learning across all object instances, CASD calculates the combined attention aggregated from multiple transformations and feature layers of the same image. To maintain consistent spatial monitoring of all objects, CASD self-extracts the WSOD network to ensure that consistent attention is paid to different transformation results of the same image. In this way, the detection performance of WSOD can be further improved. A WSOD framework named the Spatial Likelihood Voting and Self-KD Network proposed in [118] can also be used for this purpose.

In addition, a self-knowledge distillation (SKD) module based on spatial likelihood voting (SLV) is proposed to refine the feature representation of a given image. Under this approach, the likelihood graph generated by SLV is used to supervise the feature learning of the backbone network. As a result, the backbone network can focus on a wider area and improve the detection model performance. In response to the second problem, the approach in recent years has been to use multi-instance detection networks (MIDN), which select the best instance from the candidate instances and then aggregate the others based on similarity. The work in [119] found that the accuracy of the detector could be greatly improved by selecting a better

polymerization standard. Thus, Zeni et al. proposed a refined KD method in 2020, which improved the performance of OICR by making additional improvements to the existing method; this was named boosting-OICR (BOICR).

This method of combining the KD strategy with weakly supervised learning is often used in conjunction with multi-label image classification [120], phrase grounding [19], and 3D OD [31]. The work in [120] was the first to apply knowledge distillation to multi-label image classification. A depth image classification framework based on multi-knowledge tags was subsequently proposed. First, a weakly supervised detection (WSD) model was designed, after which WSD knowledge was used to guide the teacher classification model to construct an image classification framework and rectification module based on prediction and object-level visual features, which improved the classification effect. For phrase-based problems, it is necessary to learn region-phrase pairs from image-sentence pairs. The work in [19] designed a contrastive learning framework and implemented an object detector, which is able to extract weakly supervised phrase grounding in image regions. Its novelty lies in the learning of a regional phrase scoring function and image sentence scoring function. This function works by comparing each region in the image with the candidate phrase without detecting the object. Based on this, [31] realized 3D OD by using paired images and point clouds. It proposes an unsupervised 3D object proposal module (UPM) based on standardized point cloud density to generate an object candidate box. In this case, only the object is identified, not the category. The student model projects the 3D candidate frame forward (into 2D). After cutting, the candidate frames are classified and optimized, and the final prediction is generated using the teacher model trained on the image dataset. The student model also generates predictions using the RoIAlign and Full Connection layers.

Although KD has improved the performance of the existing weakly supervised OD models through the use of various technologies, an obstacle to high detection accuracy for these OD models remains the weak labeling information of the objects in the samples (especially objects that are difficult to detect) [29], [118]. Therefore, it is worth to explore the future research on weakly supervised OD based on KD. In addition, KD technologies should also consider how to reduce the gap between the weak labeling information and the actual accurate labeling of the objects, as this will improve the performance of the weakly supervised OD models [19].

*7) Extended OD Tasks Using KD:* In the field of computer vision, there are a large number of related tasks and extended OD tasks. For example, lane detection, face detection, and person search are also related to OD tasks; moreover, relationship detection and HOI detection can also be considered as extended OD tasks. Table II lists the related methods and datasets for these tasks. Of course, there may be other relevant OD tasks in existence. At present, as far as our investigation has shown, KD technologies have also been successfully applied to these extended OD tasks. Accordingly, in this section, we provide a brief overview of these works.

The core difficulty with lane detection is the need for large amounts of expensive labeled data. While numerous existing

TABLE II  
THE LIST OF EXTENDED OD TASKS USING KD. THE INTRODUCTION OF KD TECHNOLOGY SIGNIFICANTLY IMPROVES THE PERFORMANCE OF THESE EXTENDED OD METHODS VIA KNOWLEDGE TRANSFER, AND MULTIPLE EXPERIMENTS ARE PERFORMED ON SEVERAL RELATED DATASETS.

Methods	Extended OD tasks	Datasets
Peng et al. [121]	Lane Detection	CULane, LLAMAS
Jin et al. [122]	Face Detection	Fddb, WIDER FACE
Munjal et al. [123]	Person Search	CUHK-SYSU, PRW-mini
Yu et al. [15]	Relationship Detection	Visual Relationship Detection (VRD), Visual Genome
Plesse et al. [124]		Large VG, VRD, Visual Genome
Moutik et al. [125]	HOI Detection	V-COCO, HICO-DET

works have utilized active learning-based methods to address it, their performance is unsatisfactory. Peng et al. [121] tried to introduce the KD strategy into active learning, and accordingly designed a KD-based active learning model to evaluate the uncertainty of the data. This new method can effectively solve the problem caused by noisy labels and inappropriate entropy. Moreover, while the development of facial detection is well established, it also faces the problem of increasing model complexity. To maintain reasonable face detection accuracy while achieving a lightweight model, KD can be utilized. For example, Jin et al. [122] used a KD-based loss function to deal with the problem of category imbalance, thereby improving the competitiveness of lightweight facial detectors. Person search consists of two parts, namely person detection and re-identification, and also incorporates a goal-finding process. In order to better optimize the detection and re-identification aspects of the people search model, Munjal et al. [123] proposed a KD-based end-to-end people search model, which applied KD to supervising the model training.

Visual relation detection is designed to detect relationships between objects in order to facilitate a deeper understanding of the image. Relationship detection requires first locating the objects in the image, then identifying the relationships between them. However, the complex interactions between objects and the lack of abundant available data lead to unsatisfactory relationship detection results. To improve these results, KD strategies are introduced into visual relation detection. For instance, in order to solve the problems of the large semantic space of object relationship and the lower amount of available model training data, Yu et al. [15] used a data mining method to extract the linguistic knowledge from training annotations and publicly available texts. This extracted linguistic knowledge is then distilled into the student model (an end-to-end deep neural network) to predict the visual relationships from visual and semantic representations. Similarly, the commonly used visual models perform poorly when learning the knowledge of each predicate due to the complex object interaction in the image. Subsequently, a KD method based on spatial feature statistics is used to distill the semantic knowledge from the precomputed models and training annotations, which helps the visual model to estimate the relevance of object pairs [124]. Therefore, KD-based methods are mainly used to extract the semantic information from multi-modal data; we can subsequently can distill this semantic information into object relationship detection models to improve these visual models' ability to understand the semantic relationship between objects. Similar to relationships detection, HOI detection also needs to locate people and objects in the video first, which is

followed by detecting the interactions between them. To achieve a better HOI detection result, we should also consider the scene information while considering the interactions between human and objects. Moutik et al. [125] used a deep neural network (such as AlexNet, VGG or ResNet) for scene recognition to extract the scene information, then transferred this scene information to the human-object relationship detection task through a knowledge graph neural network. The scene information introduced by KD technology significantly improves the performance of the HOI detection models. In addition, language knowledge is also distilled to HOI detection models [126]. A pre-trained visual and textual model is used as the teacher model to extract interactive relational knowledge from visual and textual data, so as to guide the student transformer-based HOI model to better detect the potential (unseen) relationships in provided images.

In general, whether it is a conventional, specific or extended OD task, KD techniques/strategies can be used to improve the performance of OD models, or to obtain lightweight OD models via knowledge transfer. Therefore, we believe that KD may become a mainstream trend to address the limitations of OD model in the future.

#### E. OD methods/models Based on Different KD Strategies

In Section III-D, we have provided a review of the KD-based OD methods/models based on different types of OD tasks. However, we can also consider that these related works are designed according to different KD strategies: these include conventional methods using soft labels and distillation loss, some other methods based on feature distillation, multiple teacher distillation, self-feature distillation, specific information guidance, and so on. In this section, we will analyze and review related KD-based OD models from the perspective of different KD strategies.

1) *Methods Using Distillation Loss and Soft Labels:* At present, the utilization of different types of distillation loss and soft labels are common fundamental KD strategies [127], [128], [129]. The classical KD loss can be divided into two parts: target-class knowledge distillation (TCKD) and non-target-class knowledge distillation (NCKD). Moreover, the coupling of TCKD and NCKD limits the flexibility of balancing these two parts and suppresses the effectiveness of knowledge transfer; thus, decoupling knowledge distillation (DKD) is required during KD [130]. In addition, the softened probabilities output by the teacher model are referred to as soft labels [122], [131], which can provide richer information to the student model than hard labels, enabling the lightweight student model to achieve better performance. In this section, we collate related

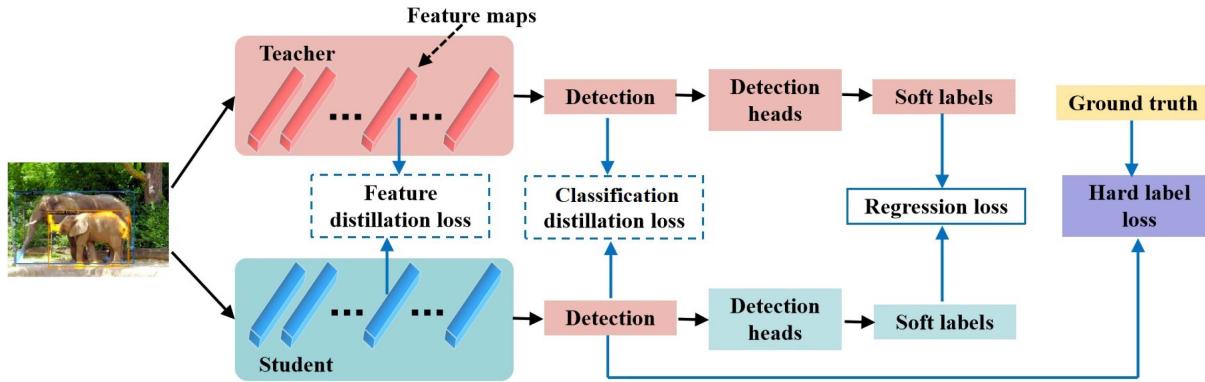


Fig. 4. The utilization of different types of loss in a detailed structure of KD-based OD models. Hard label loss and regression loss are commonly used, while feature distillation loss and classification distillation loss are selectively applied in some methods for specific tasks. We represent these loss functions using dotted lines and boxes.

works employing distillation loss-based and soft label-based KD strategies, analyze the basic principles of these methods, and identify the shortcomings and challenges. Fig. 4 presents the detailed structure of KD-based OD models and the utilization of different types of loss, while more complex network structures are designed to solve specific OD problems.

#### (1) Distillation Loss-Based KD for OD

In traditional KD-based visual tasks, the proper use of distillation loss can significantly improve the performance of student models [44], [126], [132], [133]. In commonly used KD-based OD models, the student models can be optimized using a traditional cross-entropy loss [102], [134]. Due to the huge amount of computation introduced by OD models in the pursuit of high performance, a loss function based on mutual information was proposed by liu et al. the goal of which was to make the features learned by student models closer to that of teacher models [135]. In addition, Amik et al. [136] proposed Dynamic Corrective Knowledge Distillation (DR-KD) for OD. DR-KD transforms the student model into its own teacher model. If the teacher model (which is also the student model) makes an incorrect prediction during the guidance process, then the error is corrected before knowledge learning. Therefore, the logit is constantly cross-checked with the labels to determine whether the highest logit predicted by teacher models is mapped to the labels during distillation. DR-KD has comparable performance to existing state-of-the-art teacher-less KD frameworks.

Furthermore, novel types of distillation loss are designed and optimized to improve the accuracy of OD models [99], [137]. For example, focal loss and its variants are introduced to solve the problem of label imbalance [63], while class-balanced focal loss is used to solve the problem of class imbalance and the imbalance between positive and negative samples [50]. Similarly, weighted cross-entropy loss is designed to solve the problem of class imbalance [4]. Distillation loss can also be designed by introducing attention mechanism [138]. For example, a loss function with attention mechanism is designed to enable the object detector to learn a strong mapping relationship from a small number of samples, thereby improving the ability of the object detector to detect the foreground [139]. These methods

usually guide the training of student models by minimizing the task-dependent loss and KD loss, which requires loss weights to balance the two terms of these loss functions. However, it is very difficult to choose appropriate weights. Therefore, a step-wise knowledge distillation (SSKD) strategy is proposed, which operates by transferring the useful knowledge from the teacher model to the student model rather than using the ground truth. A KD training strategy of this kind avoids the need for loss weight selection [140].

Another way of improving the loss function is to jointly optimize the distillation loss of the old branch and the detection loss of the new branch [23]. For example, to avoid catastrophic forgetting in KD-based incremental OD models, the improved cross-entropy loss is used to replace the hard ground truth labels; in addition, the distillation loss on the old categories and the cross-entropy loss on the new categories are jointly optimized, allowing the model to achieve a good prediction of both the old and new categories [141]. Moreover, combining multiple distillation loss functions is another good way to improve the performance of KD-based OD models [25], [65], [81]. For example, in the incremental object detector based on triple residual networks [26], the two-level residual distillation loss and the joint hierarchical distillation loss are combined, which enables distinguishing the features between the old and new categories; the knowledge learned from the old and new categories is also maintained respectively.

#### (2) Soft Label-Based KD for OD

In order to improve the performance of student models and reduce their dependence on ground truth labels, related works typically guide student models using soft labels or pseudo-labels output by teacher models [16], [19], [78]. In addition, we need to successfully use the predictions of teacher models as soft labels for student models, and we should also know how to reasonably assign the labels of teacher models, even the hard labels. For example, LAD in [72] can significantly improve the performance of a student model with a lightweight teacher model, and the performance of the student model will in fact be superior to that of its teacher. LAD requires a trained teacher to provide guidance (soft and hard labels) for its student. However, in

practice, it is difficult to obtain excellent teacher models capable of providing effective guidance information. Therefore, Zhang et al. [57] proposed a self-distillation framework for OD, called label-guided self-distillation (LGD), which can generate the required guidance information using only the internal relationship between objects. Moreover, there are also some other methods that have tried to obtain better OD results by combining soft and hard labels [39], [71].

In label guided KD-based OD models, pseudo-labels can also play important roles in the training of student models. For example, Feng et al. [97] proposed an adaptive pseudo-label selection strategy to selectively calculate the distillation loss using the pseudo-labels. The student model carries out feature learning by using pseudo-labels of the teacher model first, and fine-tunes the network according to the ground truth labels. This KD strategy can not only reduce the demand for labeled samples [59], [60], but also achieves higher OD performance than traditional methods based on the single-stage KD strategy [142]. In addition, for KD-based 3D OD models, using high-beam point clouds with pseudo-labels to train student models is a good way to solve the problem of high-cost 3D sample data labeling; thus, low-beam pseudo-LiDAR needs to be generated by down-sampling high-beam point clouds [108].

2) *Feature Distillation for OD*: Another KD strategy for feature distillation in intermediate feature layers can also effectively improve the performance of OD models. In this subsection, we will provide an overview of different feature distillation strategies embedded in OD models, including the basic ideas of feature distillation, full trust feature distillation, selective trust feature distillation, and so on.

In general, the methods based on feature distillation use the features output from the middle layers of teacher models to supervise the training of student models, so that the student models can mimic the features output from the teacher models to the greatest extent possible. The essence of the idea is to continuously optimize the loss function consisting of both the activation functions of the teacher and student models' feature layers. Therefore, feature distillation is carried out by using a loss function to train the student model. A general formulation of the loss function for feature distillation has been provided by Gou et al. [10]:

$$L = L_{FeaD}(f_t(x), f_s(x)) = L_F(\Phi_t(f_t(x)), \Phi_s(f_s(x))), \quad (3)$$

where  $f_t(x)$  and  $f_s(x)$  are the output features from the middle layers of the teacher and student models respectively. Considering that the different network structures of the teacher and student models may lead to different sizes of the output features from their middle layers, a transformation function  $\Phi(\cdot)$  is used to match these features.  $L_F(\cdot)$  denotes the loss function between the features of teacher and student models.

According to the related works on feature distillation published in recent years, there are two main distillation strategies employed for feature distillation on OD models: full trust feature distillation and selective trust feature distillation. In the below, we will provide a description of these methods in detail.

### (1) Full Trust Feature Distillation

Full trust feature distillation means that the student model learns all the knowledge from the teacher model unconditionally, without considering whether the knowledge to be learned is correct or not. Methods based on full trust feature distillation can further be divided into global feature distillation and local feature distillation.

*Global Feature Distillation* is an approach in which the student model imitates the entire feature maps of the middle layers of the teacher model. For example, the fast scene text detector uses all feature maps of the teacher model to guide the training of the student model [92]. In addition, the related works in [115], [116], and [93] also designed their models through the strategy of full trust feature distillation. These methods are in essence the applications of the most basic feature distillation strategy; they do not make any improvements to the feature maps, but instead simply guide the student models to learn all feature information directly from the teacher models. However, there are limitations on the capacity to improve the performance of student models by learning the global features of the teacher model indiscriminately. Therefore, researchers have gradually explored optimizing the output feature maps to further improve the performance of student models. For example, Qi et al. [24] aligned the feature maps at different resolutions using a feature pyramid, dynamically fused these features, and finally extracted these fused features from the teacher model to provide better guidance to the student model. The work of He et al. [88] ranks the feature maps by calculating the channel strengths of these feature maps in the teacher and student models for feature distillation. The feature distillation strategies proposed in [24] and [88] perform knowledge learning between the same layers of the teacher and student models; moreover, the multi-layer feature maps of the teacher model can also be used to guide the single-layer feature learning of the student model [28], [89].

*Local Feature Distillation* refers to the simulation of a student model learning those local features that are more helpful for the final prediction, rather than the entire feature maps of the teacher model. In recent years, an increasing number of related works have explored corresponding distillation strategies for local feature learning; these methods mainly try to learn the visual features at key locations in the feature maps. Chen et al. [70] used a region distillation strategy to train a lightweight pedestrian detector that crops features corresponding to RoI regions, after which the cropped local feature maps are used as the guidance information for the student model. In addition, the anchors in object detectors are widely used to locate the key local features for training student models [17], [91], [95]; these anchors can also be ranked to enable the student model to learn feature maps with different significance [143]. It can be readily observed that the local feature distillation in the above mentioned methods is performed directly around the feature maps. The attentional mechanisms can also be used to make the student models pay more attention to the key local visual features. For example, spatial attention on feature maps is introduced for local feature distillation [144], while the attention mechanism used to highlight foreground regions as well as contextual information is also helpful for OD [47]. Furthermore, Yang et al. [76] considered that the teacher and student models pay different

levels of attention to the foreground and background, while the uneven differences in the feature maps in turn impact the effect of KD. Therefore, these authors proposed a strategy combining focal distillation and global distillation, in which focal distillation is guided to the student model by using spatial and channel attention masks during model training. The goal here is to make the student model focus only on the key pixels and channels on the feature map, thereby improving the performance of the student model. There are also some other local feature distillation strategies, such as the key proposal generation of the student and teacher models for local feature distillation [75].

These above mentioned OD methods based on whether global feature distillation or local feature distillation place full trust in the guidance information of the teacher model. However, it should be noted that the guidance feature information used by the teacher model to supervise the training of the student model may be incomplete or even incorrect, which has a negative impact on the performance improvements of the student model.

### (2) Selective Trust Feature Distillation

To address the detrimental effects on the student model of incorrect information provided by the teacher model, the strategy of selective trust feature distillation is introduced in KD-based OD models. Selective trust feature distillation means that the guidance information provided by the teacher model to the student model should be selected first: in short, it is necessary to remove the incorrect information and leave only the feature information that has a positive impact on the detection performance. For example, Heo et al. [145] proposed a margin ReLu to suppress the unfavorable feature information from the teacher model, so that the student model learns only the favorable features and thus achieves performance improvement.

In summary, whether full trust feature distillation or selective trust feature distillation is employed, this is ultimately optimized through the loss function. However, it is challenging to quickly and accurately select the substantial and beneficial features from the large amount of prior guidance information provided by the teacher model. There are thus many scientific problems worthy of further study associated with OD models based on selective trust KD.

**3) Various Network Structures of Teacher-Student Models:** This section will explore the KD strategy from a new perspective. Specifically, we found that different network structures can be designed for the teacher and student models respectively, and the knowledge extracted from multimodal data can facilitate significant performance improvements on the part of the student model. Therefore, this section will summarize and analyze the network structures of the teacher-student models and the feature learning from multimodal data.

### (1) Similar Teacher-Student Network Structures

It is a common KD strategy that the teacher and student models have similar network structures. Many different backbone networks have been adopted by teacher and student models, such as ResNet [18], [58], ResNext [146], SSD [80], VGG [146], and so on. In addition, some studies do not directly use the classical network model as the backbone network, but instead adjust existing networks [32], [138]. However, these methods are KD-based OD models using traditional distillation strategies,

whether they directly use typical networks or employ adjusted networks as the backbone of the teacher and student models. Most KD-based OD models with similar teacher-student network structures extract the knowledge from single-modal data (RGB images), although there also are some methods that try to learn knowledge from other modalities for guiding the lightweight student model. For example, the student model can learn semantic knowledge provided by the teacher model [111], learn the textual and visual features extracted by the teacher model on text information [147], or jointly learn the visual features from RGB images and heat-like images under the guidance of the trained teacher model on these two modes of data [14].

Furthermore, for 3D OD, there are also common approaches involving teacher and student models using similar network structures as their backbones. For example, Wei et al. [108] opt to use the KD strategy to generate a lightweight 3D detector; here, the network structure of teacher and student models in the distillation framework is the same 3D convolutional neural network. In addition, ItKD, designed by Cho et al. [106], uses an autoencoder in combination with KD to improve the performance of a 3D object detector. The teacher and student models in ItKD are composed of the same backbone CenterPoint and autoencoder, and the same point cloud data is used for training the teacher and student networks. In the KD-based 3D OD tasks, multi-modal data is also used as the input of the distillation models. In methods employing this strategy, the student model is trained using 3D point cloud data, while the teacher model is trained using other modalities. For example, Qin et al. [31] proposed a cross-modal KD method, in which RGB images are used to train the teacher model and the point cloud is used to train the student model. This method aims to transfer the knowledge from the RGB domain to the point cloud domain, thereby reducing the labeling cost of 3D OD. Moreover, multimodal data can be used to train the teacher model, which is more beneficial to the performance of the student model [45], [109]. Multimodal data (LiDAR-image, which consists of point clouds and RGB images after segmentation) is used for training the teacher model in [45], [109]; here, the student model is expected to learn the knowledge from the teacher model and to obtain the similar outputs to the teacher model using only LiDAR.

### (2) Different Network Structures of Teacher and Student Models

Another KD strategy involves the teacher and student models using different network structures as their backbones. For example, the method in [82] uses DarkNet-53 based SSD as the teacher model's backbone and MobileNet v2/ShuffleNet v1 as the student model's backbone. Su et al. [55] use ResNet-based networks as the teacher model's backbone and a self-built 3-layer CNN as the student model's backbone. A similar strategy is used in [5]. In addition to those listed above, there are also many more similar methods with different combinations of teacher-student models. Notably, while the teacher and student models can choose various networks as their own backbones, it is necessary to choose the appropriate networks according to the specific problems to be solved, especially given the lack of capacity of the student model. Similarly, different network structures can

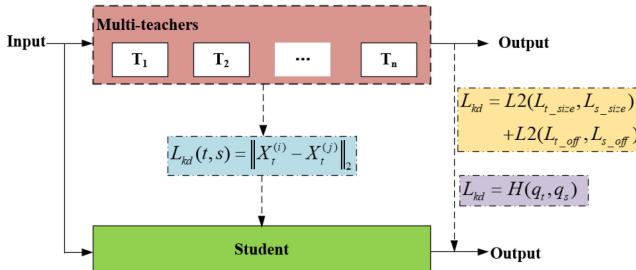


Fig. 5. The structure of multiple teacher models guiding one student model to learn the knowledge.

also be used for KD-based 3D OD tasks. Sautier et al. [46] use a 2D-to-3D distillation strategy to improve 3D OD in an autonomous driving context. The backbone of the teacher model is ResNet50 trained with RGB images, and the student model uses U-Net as its backbone trained with LiDAR data. The final experiments show that the model with this strategy outperforms the state-of-the-art methods.

This section lists several KD-based OD models using various teacher-student model network structures. Similar/different networks are used as the backbones of the teacher and student models to extract features from multimodal data. Through analysis of the existing relevant methods, we determine that the methods using different network structures as the backbone of the teacher model to extract knowledge from the multi-modal data have relatively more advantages, when it comes to guiding the feature learning of the student models. However, there are no fixed KD strategies of combining teacher-student model structures and multimodal data, we should design/choose appropriate networks for KD according to the specific tasks.

**4) Multiple Teacher Models:** KD is similar to the learning processes used by humans. The traditional technology of KD involves a teacher model guiding a student model to learn the knowledge. Notably, however, human teaching activities involve more patterns than traditional KD. Therefore, several KD models based on human teaching patterns have been proposed.

#### (1) Multiple Teachers Guiding One Student

As previously discussed, the first commonly used human teaching mode is that in which multiple teachers teach one student. For example, if teachers with different areas of specialization all teach one student, the student can acquire higher-quality knowledge. Similarly, we can use different types of teacher models to learn different knowledge from large-scale datasets, then try to transfer the learned knowledge to one student model, so as that this student model can learn more comprehensive and significant visual features, as shown in Fig. 5. For example, in [37], [53], [54], multiple teacher models are used to guide one student model to improve its OD performance. There are some key differences between these three works: Kuang et al. [53] carried out weighted fusion of different teacher models to improve the accuracy of OD networks; Chen et al. [37] used two teacher networks trained with different strategies to ensure that the knowledge could be fully transferred to the student network; Li

et al. [54] designed an asymmetric two-path learning framework to train the student model.

#### (2) One Teacher Guiding Multiple Students

The second commonly used teaching pattern is that one teacher teaches multiple students. For example, multiple student models are guided by one teacher model, after which one student model with the best performance is selected from all the student models. The work in [65] uses a KD framework in which one teacher model guides multiple student models to solve the problem of Siamese trackers being limited by high cost.

#### (3) Students Guiding Each Other

As is evident, students in real-world scenarios can also learn from each other. Accordingly, the works in [36], [65] apply this approach to KD to improve the performance of student models in OD. Kang et al. [36] proposed an oracle knowledge extraction framework based on a neural structure search to solve the capacity and complexity problems of the integration model. Moreover, teacher-student KD [65] involves not only a teacher model guiding multiple students, but also mutual guidance between student models. For knowledge sharing between students, the final objective function is as follows:

$$L_{s_1}^{kD} = L_{s_1}^{kT} + \sigma(s_1) L^{ks}(s_1 \| s_2), \quad (4)$$

$$L_{S_2}^{KD} = L_{S_2}^{KT} + \beta \cdot \sigma(S_2) L^{KS}(S_2 \| S_1), \quad (5)$$

where  $S_1$  and  $S_2$  are a slow and a smart student model respectively. The smarter the student model is, the better it will learn the knowledge.

**5) Self-Feature Distillation:** Traditional KD methods usually require a pre-trained teacher model to train a student model. To reduce the dependence of the student model on the teacher model, many scholars have designed so-called teacher-free KD models. The existing teacher-free KD methods include self-distillation, cooperative learning, and label regularization, among others. In addition, there is another type of KD, namely feature distillation, which is dissimilar to logits methods (object distillation). The student model in logits methods only learns the teacher model's logits as the resulting knowledge, rather than the middle-layer features. Many scholars have applied feature distillation to self-feature distillation, enabling a KD scheme without teachers to be devised.

Self-feature distillation does not require teacher networks, which reduces training cost. Moreover, these methods carry out one-to-one feature transformation learning between layers. For example, LGD [57] is a self-distillation model used for OD. The knowledge learned by networks can be transmitted between different layers through top-down distillation [38], and attention maps are used to capture more effective feature information and thereby promote the development of OD. An auxiliary classifier is added to the intermediate feature layers to strengthen self-supervision and enable student models to learn more effective feature representations [56]. Progressive self-knowledge distillation (PS-KD) [39] carries out feature learning by gradually extracting the knowledge from the features the model obtains. Notably, since all the lower layers imitate the attention maps of upper layers, the attention information of the lower layers may be lost. Accordingly, Bi-SAD [18] is

proposed to solve this problem. Moreover, in YOLOv6 [148], a simple self-distillation technique using a pre-trained student model as the teacher model is introduced to minimize the data distribution difference between the predictions of the teacher and student models.

In order to solve the problem of models easily falling into overfitting due to limited sample availability during few-shot OD, Li et al. [139] utilized the self-feature KD strategy to improve the generalization ability of the student model by designing an attention loss, which includes classification, regression, and class-specific features in a small number of examples. The core concept is that the student model can learn the feature mapping functions to approximate the original model through location and category feature transformation. Moreover, to improve the generalization ability of OD models, Wu et al. realized cross-domain OD based on Single-Domain through self-distillation technology [58]. Cyclic-disentangled self-distillation is proposed to continuously strip the scene information of the objects during model training, as well as to extract the shared feature expression suitable for OD in different domains. Specifically, the fine-grained location and classification information contained in multiple convolution layers of the teacher model is used to guide the backbone network (student model) to learn cross-domain features, thereby improving the generalization ability of OD models and achieving cross-domain OD. Similarly, self-distillation technology was used by He et al. for cross-domain OD [149]. These authors designed two self-distilling branches to learn the shared proposal features from the source and target domain. In addition, there are also some distillation frameworks that use self-distillation as part of their models, such as the methods in [129] and [118], which have successfully integrated self-feature distillation patterns into their models to achieve performance improvements.

6) *Specific Information Guidance for OD*: General OD models based on KD make full use of a heavyweight teacher model to transfer the complex knowledge learned from a specific dataset to a lightweight student model. In this section, we will introduce several models in which some specific knowledge information extracted by the teacher models is transferred to the student models. Among these models are the mask-guided OD models that focus on the local features of images, along with the textual guided models that introduce textual prior knowledge information for improving the student models' ability to learn the visual features; there are also some other models that pay attention to the semantic information of different object relationships in images through semantic guidance, etc.

#### (1) Mask-Guided KD-Based OD

The mask-guided network [32] is designed based on the two-stage OD model structure. Here, the mask information is used to guide the student model to pay attention to the global and local features, and the loss function of KD is given by combining global and local loss. Similarly, the method proposed by Wang et al. [17] uses the ground truth in the initial generation model as a mask to guide student models in learning the features of the objects of interest and the adjacent objects. For fine-feature imitation, a combination of the adaptation layer and the generated mask enables the student model to simulate the teacher

model's attention to local features and nearby objects. Limitation loss combined with testing loss is proposed for training the student model. In current KD-based OD models, student models largely depend on the outputs of teacher models, or transitional trust teacher models. However, in real-world scenarios, teacher models may not be able to provide very reliable output features or prediction results. Therefore, student models should learn the knowledge from teacher models selectively. Several methods have been developed to rank the outputs (including feature maps, proposals or predictions) of the teacher models based on their own quality measurements, and select a few key predictive regions or reliable predictions to guide the student OD models [128], [143], [150].

#### (2) Attention-Guided KD-Based OD

There are two issues that are not considered in the traditional KD-based OD models: (1) there is an imbalance between foreground and background pixels; (2) the relationship between different pixels lacks distillation [146]. Therefore, to carry out attentional feature learning for the key areas in images, some methods have opted to introduce the attention mechanism to guide the student model in learning local visual features. Attention mechanisms are introduced in KD to automatically locate the ROIs, the features in which are easily overlooked by student models [138], [146]. For example, a KD framework with the attention-guided mechanism is proposed based on the two-stage OD model (Faster RCNN) in [146]. The attention mechanism is used to guide the distillation model in finding key pixels and channels from the whole feature map, which enables the student model to pay more attention to these key pixels and channels rather than the whole feature map; this also inhibits the student model's learning of background visual features to a certain extent. In addition, the attention-guided mechanism can also be introduced into a one-stage OD model (SSD [151]). An end-to-end attention-guided KD method is designed in [138]. Similar to the model in [146], in which a Mask L2 Loss for local feature distillation is used in ROIs, the model in [138] uses an attention-guided distillation loss (the weighted euclidean loss) to reduce the gap between the features extracted by the teacher and student models. While, Chu et al. [144] designed a distillation loss based on the difference in spatial attention information and the predictions between teacher and student networks, in which the spatial attention information is extracted from the feature maps as knowledge for distilling to the student model. This KD-based OD method is used to recover the performance of the student model after model pruning. However, different from [138], [146], the KD-based OD models with an attention mechanism can also be used for feature fusion during model training [33]. To enable student models to learn both the abstract and simple knowledge, the model can focus on learning the abstract knowledge during subsequent model training.

The above-mentioned KD-based OD models use attentional guidance modules for student models to pay more attention to the local details of the objects. Mask-based methods were first developed for outdated detectors, such as vanilla Faster-RCNN, which failed to extend to modern detectors equipped with FPN. Specifically, these methods perform direct one-to-one matching

between pyramid levels of the student-teacher pair, which leads to two issues: (1) indiscriminately applying the same mask on all levels may introduce noise from unresponsive feature levels; (2) mask-based methods are not extendible to heterogeneous detector pairs, since their feature levels may not be strictly aligned.

#### (3) Semantics-Guided KD-Based OD

To more fully utilize various pieces of information associated with OD, in addition to making full use of the information of the single object or area, semantic context information [17], [52], [135] and other prior knowledge of the outside world (such as text messages [19], [64], [111]) can also be introduced to KD-based OD models. The context information of images makes for effective features in computer vision tasks. The teacher models can fully learn the context information, and then guide the student models to pay attention to other relevant things around the objects to be detected [17], [52], [135], [152] etc. A simple way to use the contextual features is to make the model focus on the relative position between relevant objects, then use the visual features of these relevant objects to assist with the current visual task. Therefore, different mechanisms have been developed to support the teacher model in guiding the student model to estimate the locations of the objects and their surroundings. Notably, these models do not depend on the soft support outputs of teacher models [17], [152]. Another way to utilize image context information is feature interaction learning between different network layers. Since the object and its surrounding areas have similar patterns, KD should be carried out at multiple network layers (rather than at one or the last layer of the network). For example, Yao et al. [52] proposed a KD strategy of semantic-guided feature imitation to extract the relationship features between different regions. Similarly, Liu et al. [135] also observed the relationships (i.e., semantic context information) between the features extracted by teacher and student models at different network layers. As a result, mutual information loss was introduced into the distillation network to make the student model extract as much information similar to that provided by the teacher models as possible. In addition, the semantic information contained in the generated image captioning is distilled to guided the student model of salient object detection [111]. The semantic features are extracted from texts by a transformer decoder, and used to guide the visual encoder in different convolution layers so that it focuses on multiple salient objects. In addition, the relationship between instances can also be considered as semantic features of objects. VS et al. [153] construct an instance relation graph neural network using the proposals provided by the teacher model, and design a graph distillation loss and a graph contrastive loss for distilling the instance relationship information to the student model. In this way, the student cross-domain OD model can better avoid overfitting to noise introduced by the teacher models' unreliable predictions.

#### (4) Text-Guided KD-Based OD

Textual information is another kind of prior knowledge that can be used for OD. However, it is difficult to accurately match these two modes of textual and visual features; moreover, it is challenging to guide the network to better extract visual features

and make more indirect inferences. Therefore, it is beneficial for the teacher model to learn valid prior knowledge from text data, and further to guide the student model to effectively extract the visual features. For example, to reduce the expansion cost and increase the reasoning speed of OD models, a method called ViLD (Vision and Language KD) [64] is proposed for transferring the prior knowledge extracted from open vocabulary datasets into student models. ViLD may be the first method that transfers the textual knowledge extracted by a pre-trained image classification model (a teacher model) from LVIS datasets to a two-stage OD model (a student model). Specifically, the text and image regions of the categories suggested by the object are encoded by a teacher model, after which the region embeddings of the detection box produced by the student detector are aligned with the text and image embeddings inferred by the teacher model. For their part, to make use of text information, Wang et al. [19] used the matching relations between images and sentences as supervisory information to improve the visual feature extraction of the whole OD model. This method detects the related objects by using the related text statement information as input information, which is a more difficult weakly supervised OD task. These two methods utilized the textual features to guide the student model to detect the objects in previously observed categories. However, Ma et al. [132] tried to use texts (image captions) to guide the student model to focus on the unseen objects. In this method, a weakly supervised global level language-to-visual knowledge distillation method (GKD) is proposed to exploit distilling knowledge from the visual captions of novel categories. GKD learns the matching pairs of visual and textual features based on contrast learning with multi-layer cross-attention, and distills the caption representation to the global image representation for unseen OD.

#### (5) Lipschitz Continuity-Guided KD-Based OD

The aforementioned models use specific information or mechanisms to guide the student model to learn specific visual features. However, they ignore the functional characteristics of neural networks, which makes these technologies unreliable when applied to new tasks. To alleviate this problem, Lipschitz continuity can be used to better characterize the functional features of neural networks and guide the knowledge extraction process [133]. By minimizing the distance between the Lipschitz constant of the two neural networks, the teacher network can regularize the student network better and improve the OD performance.

## IV. EXPERIMENTAL ANALYSIS

Two common OD datasets (MS COCO [154] and PASCAL VOC [155]) are used to verify the validity of OD models based on KD. Average precision (AP), Accuracy (ACC), Intersection over Union (IoU), and other common model performance evaluation parameters are used to evaluate the performance of KD-based OD models, and the rare evaluation indicator CorLoc is also applied. In this section, we briefly describe the commonly used datasets for KD-based OD utilized in our experiments, then conduct a performance comparison analysis of the different models.

TABLE III

COMPARISON OF MODELS ON PASCAL VOC 2007(%). THE ITEMS IN **BOLD** ARE THE  $\text{MAP}_{0.5}$  RESULTS OBTAINED BY THE METHODS/MODELS ACHIEVING BETTER PERFORMANCE ( $\text{MAP}_{0.5} > 80.0\%$ )

Methods	Teacher	Student	$\text{mAP}_{0.5}$	$\text{mAP}$	$\text{CorLoc}$
Chen et al. [4]	VGG16	Tucker	59.4	—	—
	VGG16	AlexNet	60.1	—	—
	VGG16	VGGM	63.7	—	—
IDa-Det [75]	FRCNN(R-18)	—	76.9	—	—
	FRCNN(R-34)	—	78.0	—	—
	SSD300(VGG16)	—	72.5	—	—
PAD [156]	FR(R101)	FR(R50)	<b>81.98</b>	54.00	—
	FR(VGG16)	FR(VGG11)	72.99	42.18	—
	RetinaNet(R101)	RetinaNet(R50)	<b>81.21</b>	55.52	—
ROI mimic [49]	RetinaNet(VGG16)	RetinaNet(VGG11)	69.57	43.13	—
	FR(R101)	FR(R50)	<b>82.25</b>	55.52	—
	FR(VGG16)	FR(VGG11)	75.04	45.00	—
PAD-ROI-mimic [156]	FR(R101)	FR(R50)	<b>82.46</b>	56.41	—
	FR(VGG16)	FR(VGG11)	75.79	45.62	—
	FR(R101)	FR(R50)	<b>81.93</b>	54.91	—
Fine-grained [157]	FR(VGG16)	FR(VGG11)	74.59	44.60	—
	RetinaNet(R101)	RetinaNet(R50)	<b>81.46</b>	56.57	—
	RetinaNet(VGG16)	RetinaNet(VGG11)	71.99	45.53	—
PAD-Fine-grained [156]	FR(R101)	FR(R50)	<b>82.29</b>	55.39	—
	FR(VGG16)	FR(VGG11)	75.17	45.21	—
	RetinaNet(R101)	RetinaNet(R50)	<b>81.94</b>	57.27	—
KD-Mobilenet-SSD [44]	RetinaNet(VGG16)	RetinaNet(VGG11)	73.18	46.92	—
	CI-Mobilenet-SSD	Mobilenet-SSD	68.00	—	—
	R50	R50	78.5	56.1	—
LD for OD [129]	R50-DCN	R50	<b>80.2</b>	58.4	—
	(CTN-101(R101) and CTN-DLA(DLA-34))	CTN-34(R34)	75.8	—	—
Kuang et al. [53]	CTN-50(R50)	CTN-50(R50)	77.1	—	—
	SLV-SD Net [118]	SLV-SD(VGG)	54.0	—	72.0
	SSD_DarkNet53	SSD_Lite_MV2	74.00	—	—
KD using GAN [82]	SSD_DarkNet53	SSD_Lite_ShuffleNetV1	65.6	—	—
	R101	MV1	79.8	—	—
	R50	MV2	79.5	—	—
GAN-KD(SSD) [79]	R50	R18	<b>80.4</b>	—	—
	R101	R50	<b>80.6</b>	—	—
	VGG16-fSSD	MV1-SSD-lite	70.24	—	—
Dong et al. [51]	VGG16-fSSD	MV2-SSD-lite	73.55	—	—
	VGG16	1/2VGG16	75.05	—	—
	VGG16	1/8VGG16	56.88	—	—
Mask-guided KD [32]	R50	1/2R50	63.23	—	—
	MobileNet	1/2MobileNet	54.53	—	—
	ALIGN	ViLD(R50)	78.9	—	—
ViLD(Finetuning) [64]	ALIGN	ViLD(R50)	78.5	—	—
	ViLD(Supervised) [64]	ViLD(R50)	—	—	—
OORD [152]	FR(R101)	FR(R50)	—	73.1	—
	FR(VGG16)	FR(VGG11)	—	68.9	—
LONDON [133]	R50-SSD	R18-SSD	73.82	—	—
		MobileNet-SSD	69.09	—	—
AttentionDis [138]	VGG16	1/8-VGG16	57.96	—	—
	VGG16	1/4-VGG16	69.48	—	—
	MobileNet	1/2-MobileNet	63.92	—	—

### A. Commonly Used Image Datasets and Evaluation Parameters

PASCAL VOC 2007 [155] has 20 categories of objects, which includes a the training subset of 5 K images and a testing subset of 5 K images. MS COCO 2017 [154] is also a large image dataset for OD, object segmentation, image caption, and other visual tasks. The images in MS COCO are captured from complex daily scenes, and objects in these images are calibrated by accurate segmentation. There are 91 categories of objects in 328,000 images marked with 2,500,000 object labels.

$\text{AP}_S$  is the value of AP for small-sized objects (sizes of areas  $<32^2$  pixels).  $\text{AP}_M$  is the value of AP for medium-sized objects ( $32^2$  pixels  $<$  sizes of areas  $<96^2$  pixels).  $\text{AP}_L$  is the

value of AP for large-sized objects (sizes of areas  $>96^2$  pixels). **mAP** is mean average precision, which is the mean value of AP across all categories. **mA<sub>0.5</sub>** is mean average precision with  $\text{IOU}_{\text{threshold}}=0.5$ . **CorLoc** is the location accuracy rate, which is the object positioning accuracy of the training set, and represents the proportion of images in which at least one object is correctly detected. CorLoc is commonly used for weakly supervised OD.

### B. Comparison of Different Models

Tables III and IV list the performance of recent classical models on the datasets of PASCAL VOC 2007 and MS COCO 2017. We here provide the abbreviations

TABLE IV  
COMPARISON OF MODELS ON MS COCO 2017(%). THE ITEMS IN BOLD ARE THE MAP<sub>0.5</sub> RESULTS WITH MAP<sub>0.5</sub> > 60.0%

Methods	Teacher	Student	AP <sub>S</sub>	AP <sub>M</sub>	AP <sub>L</sub>	mAP <sub>0.5</sub>	mAP
Chen et al. [4]	VGG16	Tucker	—	—	—	28.3	—
	VGG16	AlexNet	—	—	—	35.8	—
	VGG16	VGGM	—	—	—	37.2	—
SSKD(FPN) [140] SSKD(RetinaNet) [140]	R152	R50	23.3	44.9	52.8	<b>61.8</b>	40.5
	R152	R50	20.3	41.8	51.6	58.0	38.7
	FRCNN(R-18) FRCNN(R-34) SSD300(VGG16)	—	16.7 17.7 3.7	29.8 31.3 21.1	39.9 40.6 35.0	— — —	29.3 30.5 19.4
IDa-Det [75]	R50	MV2	5.5	22.6	35.1	—	—
	R101	R18	6.5	25.4	39.2	—	—
	R101	R50	19.6	50.1	63.7	—	—
Chen et al. [33]	FR w/R101	FR w/R18-HCL	49.58	39.51	19.42	56.72	36.75
	FR w/R101	FR w/R50-HCL	52.87	43.81	23.60	60.97	40.36
	FR w/R50	FR w/MV2-HCL	46.47	35.81	16.77	53.15	33.71
Zhang et al. [146]	FR(R50)	—	23.5	45.0	55.3	<b>62.2</b>	41.5
	Cascade RCNN(R50)	—	24.8	48.0	59.3	<b>62.7</b>	44.4
	RetinaNet(R50)	—	22.7	43.3	52.5	58.8	39.6
Dong et al. [21]	FRCNN(RNX-101)FPN	FRCNN(RNX-101)EFPN	28.0	47.5	54.2	<b>64.7</b>	44.6
CoLAD [72]	PAA-R50	PAA-R101	27.9	49.9	57.3	<b>64.4</b>	46.0
	PAA-R50	PAA-RNX-64x4d-101	29.8	51.0	59.1	<b>66.4</b>	47.5
	PAA-R50	PAA-RNX-64x4d-101-DCN	30.6	52.8	61.9	<b>68.3</b>	49.2
Qi et al. [24]	FCOS [158](RNX-101)	FCOS [158](RNX-50)	22.8	44.8	57.0	59.9	41.6
	RetinaNet(RNX-101)	RetinaNet(RNX-50)	21.6	43.7	55.5	59.4	40.3
LD for OD [129]	R50	R50	23.3	44.4	51.1	58.8	41.2
	R50-DCN	R50	27.2	49.5	55.7	<b>63.7</b>	45.6
LGD [57]	FRCNN(R-50)	—	—	—	—	—	40.3
	FRCNN(R-101)	—	—	—	—	—	42.1
	FRCNN(R-101 DCN)	—	—	—	—	—	44.4
SFD for OD [38]	FR(R50)	—	21.4	41.7	50.1	58.7	38.2
	RetinaNet(R50)	—	20.8	40.50	50.0	56.7	37.40
SLV-SD Net [118]	SLV-SD(VGG16)	—	—	—	—	27.3	12.4
	SLV-SD(R50)	—	—	—	—	28.1	13.4
Wang et al. [17]	FR(R101)	FR(R101h-I)	13.2	35.9	47.5	51.2	31.6
	FR-FPN(R50)	FR-FPN(R50h-I)	21.0	34.9	45.5	55.8	34.8
ViLD(Finetuning) [64] ViLD(Supervised) [64]	ALIGN	ViLD(R50)	—	—	—	59.8	39.1
	ALIGN	ViLD(R50)	—	—	—	<b>67.6</b>	46.5
OORD [152]	FCOS(RNX-101)	FCOS(MV2)	19.9	37.8	43.8	54.2	35.6
G-DetKD [52]	FR-R152-FPN	FR-R50-FPN	23.7	45.0	53.7	—	41.0
	FR-R50-FPN	FR-R50(1/4)-FPN	18.1	36.6	44.5	—	33.7
AID [150]	RetinaNet(R101)	RetinaNet(R50)	—	—	—	—	70.1
	RetinaNet(R101)	RetinaNet(R101)	—	—	—	—	72.6
Li et al. [143]	RetinaNet(R50)	RetinaNet(R18)	17.9	38.1	47.6	—	—
	FCOS(R50)	FCOS(R18)	19.3	38.9	45.9	—	—
	ATSS(R50)	ATSS(R18)	21.0	40.7	48.1	—	—
	GFL(R50)	GFL(R18)	21.8	41.7	51.0	—	—
ICD [83]	FR(R50)	FR(R50)	24.5	44.2	53.5	—	40.9
	RetinaNet(R50)	RetinaNet(R50)	24.2	45.0	52.7	—	40.7
LD [74]	R101	R50	24.5	46.2	54.8	60.3	42.1
FGD [76]	RetinaNet(RNX101)	RetinaNet(R50)	22.9	45.0	54.7	—	40.7
	Cascade Mask RCNN(RNX101)	FR(R50)	23.8	46.4	55.5	—	42.0
	RepPoints(RNX101)	RepPoints(R50)	24.0	45.7	55.6	—	42.0
	Cascade Mask RCNN(RNX101)	Mask RCNN(R50)	23.7	46.2	55.7	—	42.1

of some networks: R152/R101/R50/R34/R18 stand for ResNet-152/ResNet101/ResNet50/ResNet34/ResNet18; MV1/MV2 stand for MobileNetV1/MobileNetV2; FR stands for Faster R-CNN; CTN stands for CenterNet; and RNX stands for ResNeXt.

Table III lists the performance comparison of models on PASCAL VOC 2007. From these listed evaluation parameters, we can see that the performance of the final student models in conventional KD-based OD models is largely proportional to the network parameters. For instance, most student models can achieve an mAP<sub>0.5</sub> of over 80% by using R50 as their backbone and R101 as their teacher models' backbone, while

the methods using VGG11/VGG16 as their student and teacher models' backbone have lower mAP<sub>0.5</sub>. The performance of these conventional models is similar in terms of mAP [49], [156], [157]. Most conventional models listed in Table IV use R50 as their student models' backbone, while it is still clear that model performance is proportional to model complexity for these models [33], [72], [86], [140].

Of course, the ultimate goal of designing KD-based OD models is to obtain a more lightweight student model. Therefore, advanced KD techniques have been proposed and applied to OD models. Models of this kind, which are based on adversarial learning, try to achieve comparable OD performance to

conventional models while being considerably more lightweight [51], [79], [82]. However, these methods need to design a reasonable strategy to control adversarial learning between the teacher and student models. In order to further improve the performance of lightweight student models, related methods introduce external information or multi-modal data to guide the feature learning of student models. These methods can help the lightweight student models to achieve superior performance by introducing assistance information [32], [64], [133], [152]. The situation is similar in Table IV [17], [52], [64], [143], [150]. There are relevant methods in Table III that work to obtain a lightweight student model by introducing the attention mechanism, but their performance is the worst when compared with other related methods [138]. The reason may be that the network structures of both the teacher and student models used by these methods are relatively simple.

For the dataset of MS COCO 2017 (as shown in Table IV), in addition to the methods mentioned in Table III, there are also some methods that guide the feature learning of student models by selecting valuable features during KD, and have achieved relatively superior performance [73], [76], [83]. Moreover, methods based on the self-distillation strategy can obtain considerable OD performance if they choose R50 or other networks with similar complexity as the backbone of their student models [57], [129]. However, once the network structures of student models have been greatly simplified, the performance will decline rapidly [38], [118].

In this section, Tables III and IV only list some methods with common evaluation indicators on PASCAL VOC 2007 and MS COCO 2017. There are many other KD-based OD methods/models that have achieved considerable results on different datasets for specific OD tasks and visual problems.

## V. FUTURE RESEARCH WORKS

In summary, a variety of KD strategies have been applied to OD, and have achieved significant improvements in model compression and detection accuracy. However, there are still many problems to be further studied and solved. On the basis of the previous content, we here point out some existing problems in the current works, and explore the future research directions of KD-based OD.

### (1) KD-Based Incremental Learning for OD

In the practical OD applications, it is often necessary to add some new categories of objects into the OD task (that is incremental OD). To solve this issue, traditional methods retrain the model using entirely labeled data, although this is prohibitively costly; moreover, fine-tuning the model with a small amount of data leads to catastrophic forgetting [97], [98]. The OD framework based on KD can transfer the historical object information from the teacher model to the student model; as a result, the student model not only learns the visual features of new categories of objects, but also maintains the ability to detect the historical objects. Commonly, KD-based incremental OD methods are proposed by designing corresponding loss functions [25], gradually supplementing the data of new objects [26], [98], [100], or proposing some new distillation mechanisms [47].

These methods have achieved good results in many aspects. However, there are still some problems that need to be solved for incremental OD, such as the labeling of new categories of objects, model performance degradation caused by the addition of new categories of objects similar to historical objects, how to identify the emergence of new objects, and real-time updating of the model. Therefore, adjunct networks capable of identifying the emergence of new objects in a timely manner should be added to the whole model framework; student models with stronger object discrimination ability need further consideration to improve their object recognition performance; and models based on weakly supervised or semi-supervised learning can be considered to solve the problem of labeling new objects.

### (2) Weakly Supervised or Semi-Supervised Learning OD Models Based on KD

Similar to other computer vision tasks, the high cost of data annotation is also a problem in OD tasks; this is especially true for incremental OD due to the high continuous annotation cost of increasing new categories of objects. Therefore, weakly supervised, semi-supervised, or unsupervised learning may be an effective way to solve this problem in the future. At present, there are also weakly supervised and semi-supervised learning OD models based on KD. These methods aim to achieve accurate OD by using data with weak annotations [28], multi-label annotations [120], or a small number of images with accurate annotations [114]. In addition, some visual feature extraction mechanisms [117], [118], [120] have been proposed to improve the performance of OD. Therefore, future research might investigate how to design and introduce related mechanisms or network modules into OD frameworks for analyzing image regional features, etc., or how to introduce relevant prior knowledge to further improve the performance of KD-based OD models employing weakly supervised or semi-supervised learning.

### (3) Interactive Distillation Between Multi-Teacher and Multi-Student Models for OD

Distillation learning is similar to the pattern of human learning. Thus, as in real-world scenarios, a student model can be guided by multiple teacher models for feature learning. At the same time, a task can also be completed by multiple student models and student models can learn from each other. In existing methods, a teacher model can guide multiple student models to conduct feature learning, then select the student model with superior performance [65]. Multi-teacher models can also jointly train student models in two stages [37]; alternatively, multiple teacher models can be combined to train student models by means of weight fusion [53]. These methods make full use of the advantages of multiple teacher models to train a superior student model and achieve good results. However, it is worth conducting further research into how knowledge might be seamlessly transferred from multiple teacher models to student models. In addition, it is important to investigate how the teacher model and student model interact with each other through certain mechanisms or technologies during joint model training, which may also be an effective way to improve the performance of both teacher and student models. Finally, as multiple teacher models can train multiple different types of student models, the question

of how to select the optimal student model or combine multiple student models for OD should receive attention in the future.

#### (4) New Knowledge and Multiple Modal Features Distillation for OD

The introduction of new knowledge is a very effective way to improve OD models' performance, and the ability to extract new knowledge from other modes is a major advantage of KD. Existing methods of this kind take a variety of multimodal data (RGB images [35], thermal images [55], depth images [14], [35], textual information [19], [64], etc.) as the input of the teacher model, then use the teacher model to extract relevant features from these multimodal data for guiding the student model to learn the visual features from 2D RGB images. However, a question that merits future exploration is that of how to minimize the gaps between the different types of features learned by the teacher model from multimodal data and the visual features for OD. In addition, using the features of other modal data as the prior knowledge of weakly supervised OD could also be considered when designing weakly supervised OD models based on KD.

#### (5) Model Compression for 3D OD

3D OD is mostly used in automatic driving and other fields. The OD models dealing with 3D image data are more complex than those of 2D data. It is therefore of great practical significance to use KD technology to compress the 3D OD models, and there are many innovative works on this topic worth exploring. First, as there are large-scale network parameters in 3D OD models, model compression based on KD is an important research field that needs to be considered. For example, we could use a complex deep neural network as the teacher model to guide a finely designed lightweight student model. However, another key question is that of how to improve the accuracy of OD models. Future research into KD-based 3D OD could focus on the image data and the models. For 3D image data, the models using the point cloud data as input can adopt self-distillation to improve the accuracy of 3D OD. For the models using multiple forms of data as input, the initial detection results can first be obtained from simple 2D images, after which they could be used as weakly supervised tags to optimize the 3D OD models with complex 3D data. In addition, we can use more sophisticated complex models to optimize 3D OD models, such as the KD method that combines the segmentation model and detection model. It would also be beneficial to design a suitable KD framework that is specifically tailored for the unique context of 3D OD.

Finally, KD has great advantages in model compression and model performance improvement, and has been widely used in multiple computer vision tasks. In recent years, the use of KD technology for OD-related tasks has attracted increasing research attention; at the same time, KD-based OD has also encountered many challenges. In the future, KD should be extended to a wide range of visual detection tasks, such as 3D OD, weakly supervised/unsupervised OD, visual relationship detection, social relationship detection, and so on. In addition, KD-based OD can be further applied to other specific types of data, such as multi-source remote sensing images, multi-modal images, textual data, audio data, etc.

## VI. CONCLUSION

This survey reviews KD-based OD models. First, we detailed the basic principles for designing OD models based on KD. We then summarized and analyzed the previous works in terms of the KD-based OD tasks, the KD strategies employed in OD models, the related problems to be solved, and the datasets associated with model application. Finally, we discussed the promising possible research directions to be further explored in the future. As shown by the above comprehensive analysis of current KD-based OD models, KD brings great potential to traditional OD models in terms of model compression and performance improvement. Therefore, there are many novel ideas and techniques in this research field that merit further exploration.

## REFERENCES

- [1] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.
- [2] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," 2019, *arXiv:1905.05055*.
- [3] L. Liu et al., "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, 2020.
- [4] 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.
- [5] Z. Xing, X. Chen, and F. Pang, "DD-YOLO: An object detection method combining knowledge distillation and differentiable architecture search," *IET Comput. Vis.*, vol. 16, pp. 418–430, 2022.
- [6] Z. Li et al., "A compression pipeline for one-stage object detection model," *J. Real-Time Image Process.*, vol. 18, pp. 1949–1962, 2021.
- [7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
- [8] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3048–3068, Jun. 2022.
- [9] S. Chen, R. Zhan, W. Wang, and J. Zhang, "Learning slimming SAR ship object detector through network pruning and knowledge distillation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1267–1282, 2021.
- [10] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, 2021.
- [11] J. Gou, L. Sun, B. Yu, L. Du, K. Ramamohanarao, and D. Tao, "Collaborative knowledge distillation via multiknowledge transfer," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 20, 2022, doi: [10.1109/TNNLS.2022.3212733](https://doi.org/10.1109/TNNLS.2022.3212733).
- [12] J. Gou, L. Sun, B. Yu, S. Wan, and D. Tao, "Hierarchical multi-attention transfer for knowledge distillation," *ACM Trans. Multimedia Comput., Commun. Appl.*, 2022.
- [13] D. Walawalkar, Z. Shen, and M. Savvides, "Online ensemble model compression using knowledge distillation," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 18–35.
- [14] S. S. Kruthiventi, P. Sahay, and R. Biswal, "Low-light pedestrian detection from RGB images using multi-modal knowledge distillation," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 4207–4211.
- [15] R. Yu, A. Li, V. I. Morariu, and L. S. Davis, "Visual relationship detection with internal and external linguistic knowledge distillation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1974–1982.
- [16] R. Mehta and C. Ozturk, "Object detection at 200 frames per second," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 1–15.
- [17] 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.*, 2019, pp. 4933–4942.
- [18] Y. Chai et al., "Compact cloud detection with bidirectional self-attention knowledge distillation," *Remote Sens.*, vol. 12, no. 17, 2020, Art. no. 2770.
- [19] L. Wang, J. Huang, Y. Li, K. Xu, Z. Yang, and D. Yu, "Improving weakly supervised visual grounding by contrastive knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14 090–14 100.

- [20] M. Bharadhwaj, G. Ramadurai, and B. Ravindran, "Detecting vehicles on the edge: Knowledge distillation to improve performance in heterogeneous road traffic," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3192–3198.
- [21] C. Deng, M. Wang, L. Liu, Y. Liu, and Y. Jiang, "Extended feature pyramid network for small object detection," *IEEE Trans. Multimedia*, vol. 24, pp. 1968–1979, 2022.
- [22] C. Li, X. Qu, A. Gnanasambandam, O. A. Elgendi, J. Ma, and S. H. Chan, "Photon-limited object detection using non-local feature matching and knowledge distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3976–3985.
- [23] J. Chen, S. Wang, L. Chen, H. Cai, and Y. Qian, "Incremental detection of remote sensing objects with feature pyramid and knowledge distillation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.
- [24] L. Qi et al., "Multi-scale aligned distillation for low-resolution detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14 443–14 453.
- [25] L. Chen, C. Yu, and L. Chen, "A new knowledge distillation for incremental object detection," in *Proc. Int. Joint Conf. Neural Netw.*, 2019, pp. 1–7.
- [26] D. Yang, Y. Zhou, D. Wu, C. Ma, F. Yang, and W. Wang, "Two-level residual distillation based triple network for incremental object detection," 2020, *arXiv:2007.13428*.
- [27] K. Joseph, J. Rajasegaran, S. Khan, F. S. Khan, and V. Balasubramanian, "Incremental object detection via meta-learning," 2020, *arXiv:2003.08798*.
- [28] J. Zhang, H. Su, W. Zou, X. Gong, Z. Zhang, and F. Shen, "CADN: A weakly supervised learning-based category-aware object detection network for surface defect detection," *Pattern Recognit.*, vol. 109, 2021, Art. no. 107571.
- [29] Z. Zeng, B. Liu, J. Fu, H. Chao, and L. Zhang, "WSOD2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8292–8300.
- [30] H. Feliz et al., "Squeezed deep 6DoF object detection using knowledge distillation," in *Proc. Int. Joint Conf. Neural Netw.*, 2020, pp. 1–7.
- [31] Z. Qin, J. Wang, and Y. Lu, "Weakly supervised 3D object detection from point clouds," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 4144–4152.
- [32] Y. Zhu, C. Zhao, C. Han, J. Wang, and H. Lu, "Mask guided knowledge distillation for single shot detector," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2019, pp. 1732–1737.
- [33] P. Chen, S. Liu, H. Zhao, and J. Jia, "Distilling knowledge via knowledge review," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5008–5017.
- [34] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, "A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9060–9069.
- [35] F. R. Valverde, J. V. Hurtado, and A. Valada, "There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11 612–11 621.
- [36] M. Kang, J. Mun, and B. Han, "Towards oracle knowledge distillation with neural architecture search," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 4404–4411.
- [37] X. Chen, J. Su, and J. Zhang, "A two-teacher framework for knowledge distillation," in *Proc. Int. Symp. Neural Netw.*, Springer, 2019, pp. 58–66.
- [38] Z. Wu and Z. Hu, "Object detection based on self-feature distillation," *J. Phys. Conf. Ser.*, vol. 1982, no. 1, 2021, Art. no. 012081.
- [39] K. Kim, B. Ji, D. Yoon, and S. Hwang, "Self-knowledge distillation with progressive refinement of targets," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6567–6576.
- [40] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.
- [41] 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.*, 2016, pp. 779–788.
- [42] A. Tao, J. Barker, and S. Sarathy, "DetectNet: Deep neural network for object detection in DIGITS," *Parallel Forall*, vol. 4, pp. 323–336, 2016.
- [43] J. Gou, L. Sun, B. Yu, S. Wan, W. Ou, and Z. Yi, "Multi-level attention-based sample correlations for knowledge distillation," *IEEE Trans. Ind. Informat.*, early access, Sep. 26, 2022, doi: [10.1109/TII.2022.3209672](https://doi.org/10.1109/TII.2022.3209672).
- [44] J. G. Ko and W. Yoo, "Knowledge distillation based compact model learning method for object detection," in *Proc. Int. Conf. Inf. Commun. Technol. Convergence*, 2020, pp. 1276–1278.
- [45] Z. Chong et al., "MonoDistill: Learning spatial features for monocular 3D object detection," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–17.
- [46] C. Sautier, G. Puy, S. Gidaris, A. Boulch, A. Bursuc, and R. Martel, "Image-to-Lidar self-supervised distillation for autonomous driving data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9891–9901.
- [47] X. Liu, H. Yang, A. Ravichandran, R. Bhotika, and S. Soatto, "Continual universal object detection," 2020, *arXiv:2002.05347*.
- [48] D. Yang, Y. Zhou, W. Shi, D. Wu, and W. Wang, "RD-IOD: Two-level residual-distillation-based triple-network for incremental object detection," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 18, no. 1, pp. 1–23, 2022.
- [49] Q. Li, S. Jin, and J. Yan, "Mimicking very efficient network for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6356–6364.
- [50] J. Yu, H. Xie, M. Li, G. Xie, Y. Yu, and C. W. Chen, "Mobile CenterNet for embedded deep learning object detection," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2020, pp. 1–6.
- [51] N. Dong, Y. Zhang, M. Ding, S. Xu, and Y. Bai, "One-stage object detection knowledge distillation via adversarial learning," *Appl. Intell.*, vol. 52, pp. 4582–4598, 2022.
- [52] L. Yao, R. Pi, H. Xu, W. Zhang, Z. Li, and T. Zhang, "G-DetKD: Towards general distillation framework for object detectors via contrastive and semantic-guided feature imitation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3591–3600.
- [53] H. Kuang and Z. Liu, "Research on object detection network based on knowledge distillation," in *Proc. 4th Int. Conf. Intell. Auton. Syst.*, 2021, pp. 8–12.
- [54] H.-T. Li, S.-C. Lin, C.-Y. Chen, and C.-K. Chiang, "Layer-level knowledge distillation for deep neural network learning," *Appl. Sci.*, vol. 9, no. 10, 2019, Art. no. 1966.
- [55] K. Su, C. M. Intesar, Q. Zhao, and Y. Tomioka, "Knowledge distillation for real-time on-road risk detection," in *Proc. IEEE Int. Conf. Dependable Autonomic Secure Comput. Int. Conf. Pervasive Intell. Comput. Int. Conf. Cloud Big Data Comput. Int. Conf. Cyber Sci. Technol. Congr.*, 2020, pp. 110–117.
- [56] C. Yang, Z. An, L. Cai, and Y. Xu, "Hierarchical self-supervised augmented knowledge distillation," 2021, *arXiv:2107.13715*.
- [57] P. Zhang, Z. Kang, T. Yang, X. Zhang, N. Zheng, and J. Sun, "LGD: Label-guided self-distillation for object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 3309–3317.
- [58] A. Wu and C. Deng, "Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 847–856.
- [59] A. Banitalebi-Dehkordi, "Knowledge distillation for low-power object detection: A simple technique and its extensions for training compact models using unlabeled data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 769–778.
- [60] A. Banitalebi-Dehkordi, "Revisiting knowledge distillation for object detection," 2021, *arXiv:2105.10633*.
- [61] X. Dai et al., "General instance distillation for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7838–7847.
- [62] A. Musa, M. Hassan, M. Hamada, and F. Aliyu, "Low-power deep learning model for plant disease detection for smart-hydroponics using knowledge distillation techniques," *J. Low Power Electron. Appl.*, vol. 12, no. 2, 2022, Art. no. 24.
- [63] T. Gao, Y. Gao, Y. Li, and P. Qin, "Revisiting knowledge distillation for light-weight visual object detection," *Trans. Inst. Meas. Control*, vol. 43, no. 13, pp. 2888–2898, 2021.
- [64] X. Gu, T. Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *Proc. 10th Int. Conf. Learn. Representations*, 2022.
- [65] Y. Liu, X. Dong, X. Lu, F. S. Khan, J. Shen, and S. Hoi, "Teacher-students knowledge distillation for Siamese trackers," 2019, *arXiv:1907.10586*.
- [66] J. Shen, N. Vesdapunt, V. N. Boddeti, and K. M. Kitani, "In teacher we trust: Learning compressed models for pedestrian detection," 2016, *arXiv:1612.00478*.
- [67] A. Mishra and D. Marr, "Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy," 2017, *arXiv:1711.05852*.

- [68] J.-M. Guo, J.-S. Yang, S. Seshathiri, and H.-W. Wu, "A light-weight CNN for object detection with sparse model and knowledge distillation," *Electronics*, vol. 11, no. 4, 2022, Art. no. 575.
- [69] X. Chen, C. Xu, M. Dong, C. Xu, and Y. Wang, "An empirical study of adder neural networks for object detection," in *Proc. 35th Conf. Neural Inf. Process. Syst.*, 2021, pp. 6894–6905.
- [70] R. Chen, H. Ai, C. Shang, L. Chen, and Z. Zhuang, "Learning lightweight pedestrian detector with hierarchical knowledge distillation," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1645–1649.
- [71] W. Zhang et al., "Boosting end-to-end multi-object tracking and person search via knowledge distillation," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1192–1201.
- [72] C. H. Nguyen, T. C. Nguyen, T. N. Tang, and N. L. Phan, "Improving object detection by label assignment distillation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 1005–1014.
- [73] Z. Zheng, R. Ye, P. Wang, J. Wang, D. Ren, and W. Zuo, "Localization distillation for object detection," 2021, *arXiv:2102.12252*.
- [74] Z. Zheng et al., "Localization distillation for dense object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9407–9416.
- [75] S. Xu et al., "IDa-Det: An information discrepancy-aware distillation for 1-bit detectors," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2022, pp. 346–361.
- [76] Z. Yang et al., "Focal and global knowledge distillation for detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4643–4652.
- [77] A. Umer, C. Termirthikun, T. Qiu, P. H. Leong, and I. Lee, "On-device saliency prediction based on pseudo knowledge distillation," *IEEE Trans. Ind. Informat.*, vol. 18, no. 9, pp. 6317–6325, Sep. 2022.
- [78] Q. Guo et al., "Online knowledge distillation via collaborative learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11 020–11 029.
- [79] W. Wang, W. Hong, F. Wang, and J. Yu, "GAN-knowledge distillation for one-stage object detection," *IEEE Access*, vol. 8, pp. 60 719–60 727, 2020.
- [80] N. Dong, Y. Zhang, M. Ding, S. Xu, and Y. Bai, "One-stage object detection knowledge distillation via adversarial learning," *Appl. Intell.*, vol. 52, no. 4, pp. 4582–4598, 2022.
- [81] A. Chawla, H. Yin, P. Molchanov, and J. Alvarez, "Data-free knowledge distillation for object detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 3289–3298.
- [82] E. Finogeev, V. Gorbatsevich, A. Moiseenko, Y. Vizilter, and O. Vygolov, "Knowledge distillation using GANs for fast object detection," *Int. Arch. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. 43, pp. 583–588, 2020.
- [83] Z. Kang, P. Zhang, X. Zhang, J. Sun, and N. Zheng, "Instance-conditional knowledge distillation for object detection," in *Proc. 35th Conf. Neural Inf. Process. Syst.*, 2021, pp. 16468–16480.
- [84] C. Yang, L. Xie, S. Qiao, and A. L. Yuille, "Training deep neural networks in generations: A more tolerant teacher educates better students," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 5628–5635.
- [85] A. Haselhoff, J. Kronenberger, F. Kuppers, and J. Schneider, "Towards black-box explainability with Gaussian discriminant knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 21–28.
- [86] S. Lee, S. Lee, and B. C. Song, "Balanced knowledge distillation for one-stage object detector," *Neurocomputing*, vol. 500, pp. 394–404, 2022.
- [87] Y. Zhao et al., "Real time object detection for traffic based on knowledge distillation: 3rd place solution to pair competition," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2020, pp. 1–6.
- [88] S. He et al., "Enhancing mid-low-resolution ship detection with high-resolution feature distillation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [89] T. Ma, W. Tian, and Y. Xie, "Multi-level knowledge distillation for low-resolution object detection and facial expression recognition," *Knowl.-Based Syst.*, vol. 240, 2022, Art. no. 108136.
- [90] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2018.
- [91] B.-Y. Liu, H.-X. Chen, Z. Huang, X. Liu, and Y.-Z. Yang, "ZoomInNet: A novel small object detector in drone images with cross-scale knowledge distillation," *Remote Sens.*, vol. 13, no. 6, 2021, Art. no. 1198.
- [92] P. Yang, F. Zhang, and G. Yang, "A fast scene text detector using knowledge distillation," *IEEE Access*, vol. 7, pp. 22 588–22 598, 2019.
- [93] Z. Huang, W. Li, and R. Tao, "Extracting and distilling direction-adaptive knowledge for lightweight object detection in remote sensing images," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 2360–2364.
- [94] C. Shiqi, W. Wei, Z. Ronghui, Z. Jun, and L. Shengqi, "A lightweight, arbitrary-oriented SAR ship detector via feature map-based knowledge distillation," *J. Radar*, vol. 11, pp. 1–14, 2022.
- [95] Y. Yang et al., "Adaptive knowledge distillation for lightweight remote sensing object detectors optimizing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.
- [96] W. Zhou, S. Chang, N. Sosa, H. Hamann, and D. Cox, "Lifelong object detection," 2020, *arXiv: 2009.01129*.
- [97] T. Feng and M. Wang, "Response-based distillation for incremental object detection," 2021, *arXiv:2110.13471*.
- [98] Y. Hao, Y. Fu, Y.-G. Jiang, and Q. Tian, "An end-to-end architecture for class-incremental object detection with knowledge distillation," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2019, pp. 1–6.
- [99] N. Dong, Y. Zhang, M. Ding, and G. H. Lee, "Incremental-DETR: Incremental few-shot object detection via self-supervised learning," 2022, *arXiv:2205.04042*.
- [100] D. Yang, Y. Zhou, and W. Wang, "Multi-view correlation distillation for incremental object detection," 2021, *arXiv:2107.01787*.
- [101] Y. Peng, L. Yuxuan, and L. Ming, "In defense of knowledge distillation for task incremental learning and its application in 3D object detection," *IEEE Trans. Robot. Autom.*, vol. 6, no. 2, pp. 2012–2019, Apr. 2021.
- [102] Y. Hao, Y. Fu, and Y.-G. Jiang, "Take goods from shelves: A dataset for class-incremental object detection," in *Proc. Int. Conf. Multimedia Retrieval*, 2019, pp. 271–278.
- [103] B. Yang et al., "Continual object detection via prototypical task correlation guided gating mechanism," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9255–9264.
- [104] E. Verwimp et al., "Re-examining distillation for continual object detection," 2022, *arXiv:2204.01407*.
- [105] J. Yang, S. Shi, R. Ding, Z. Wang, and X. Qi, "Towards efficient 3D object detection with knowledge distillation," 2022, *arXiv:2205.15156*.
- [106] H. Cho, J. Choi, G. Baek, and W. Hwang, "ItKD: Interchange transfer-based knowledge distillation for 3D object detection," 2022, *arXiv:2205.15531*.
- [107] Y. Wang, A. Fathi, J. Wu, T. Funkhouser, and J. Solomon, "Multi-frame to single-frame: Knowledge distillation for 3D object detection," 2020, *arXiv: 2009.11859*.
- [108] Y. Wei, Z. Wei, Y. Rao, J. Li, J. Zhou, and J. Lu, "Lidar distillation: Bridging the beam-induced domain gap for 3D object detection," 2022, *arXiv:2203.14956*.
- [109] W. Zheng, M. Hong, L. Jiang, and C.-W. Fu, "Boosting 3D object detection by simulating multimodality on point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13 638–13 647.
- [110] M. F. Bajestani and Y. Yang, "TKD: Temporal knowledge distillation for active perception," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 953–962.
- [111] B. Xu, G. Liu, H. Huang, C. Lu, and Y. Guo, "Semantic distillation guided salient object detection," 2022, *arXiv:2203.04076*.
- [112] V. Vapnik and R. Izmailov, "Learning using privileged information: Similarity control and knowledge transfer," *J. Mach. Learn. Res.*, vol. 16, pp. 2023–2049, 2015.
- [113] Y. Tang, Y. Li, and W. Zou, "Fast video salient object detection via spatiotemporal knowledge distillation," 2020, *arXiv: 2010.10027*.
- [114] J. Gao, J. Wang, S. Dai, L.-J. Li, and R. Nevatia, "NOTE-RCNN: NOISE tolerant ensemble RCNN for semi-supervised object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9508–9517.
- [115] Y. Cao et al., "Semi-supervised knowledge distillation for tiny defect detection," in *Proc. IEEE 25th Int. Conf. Comput. Supported Cooperative Work Des.*, 2022, pp. 1010–1015.
- [116] W. Qian, Z. Yan, Z. Zhu, and W. Yin, "Weakly supervised part-based method for combined object detection in remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5024–5036, 2022.
- [117] Z. Huang, Y. Zou, V. Bhagavatula, and D. Huang, "Comprehensive attention self-distillation for weakly-supervised object detection," 2020, *arXiv: 2010.12023*.
- [118] Z. Chen et al., "Spatial likelihood voting with self-knowledge distillation for weakly supervised object detection," *Image Vis. Comput.*, vol. 16, 2021, Art. no. 104314.
- [119] L. F. Zeni and C. R. Jung, "Distilling knowledge from refinement in multiple instance detection networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 768–769.
- [120] Y. Liu, L. Sheng, J. Shao, J. Yan, S. Xiang, and C. Pan, "Multi-label image classification via knowledge distillation from weakly-supervised detection," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 700–708.

- [121] F. Peng, C. Wang, J. Liu, and Z. Yang, "Active learning for lane detection: A knowledge distillation approach," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15 152–15 161.
- [122] H. Jin, S. Zhang, X. Zhu, Y. Tang, Z. Lei, and S. Z. Li, "Learning lightweight face detector with knowledge distillation," in *Proc. Int. Conf. Biometrics*, 2019, pp. 1–7.
- [123] B. Munjal, F. Galasso, and S. Amin, "Knowledge distillation for end-to-end person search," 2019, *arXiv: 1909.01058*.
- [124] F. Plesse, A. Ginsca, B. Delezoide, and F. Prêteux, "Visual relationship detection based on guided proposals and semantic knowledge distillation," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2018, pp. 1–6.
- [125] O. Moutik, S. Tigani, R. Saadane, and A. Chehri, "Hybrid deep learning vision-based models for human object interaction detection by knowledge distillation," *Procedia Comput. Sci.*, vol. 192, pp. 5093–5103, 2021.
- [126] M. Wu et al., "End-to-end zero-shot HOI detection via vision and language knowledge distillation," 2022, *arXiv:2204.03541*.
- [127] P. Yun, Y. Liu, and M. Liu, "In defense of knowledge distillation for task incremental learning and its application in 3D object detection," *IEEE Trans. Robot. Autom.*, vol. 6, no. 2, pp. 2012–2019, Apr. 2021.
- [128] C. Yang, M. Ochal, A. Storkey, and E. J. Crowley, "Prediction-guided distillation for dense object detection," 2022, *arXiv:2203.05469*.
- [129] M. Zheng et al., "End-to-end object detection with adaptive clustering transformer," 2020, *arXiv: 2011.09315*.
- [130] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11 953–11 962.
- [131] Z. Tian et al., "Adaptive perspective distillation for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1372–1387, Feb. 2023.
- [132] Z. Ma et al., "Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14 074–14 083.
- [133] Y. Shang, B. Duan, Z. Zong, L. Nie, and Y. Yan, "Lipschitz continuity guided knowledge distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10 675–10 684.
- [134] X. Qu, C. Ding, X. Li, X. Zhong, and D. Tao, "Distillation using oracle queries for transformer-based human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19 558–19 567.
- [135] X. Liu and Z. Zhu, "Knowledge distillation for object detection based on mutual information," in *Proc. 4th Int. Conf. Intell. Auton. Syst.*, 2021, pp. 18–23.
- [136] F. R. Amik, A. I. Tasin, S. Ahmed, M. Elahi, and N. Mohammed, "Dynamic rectification knowledge distillation," 2022, *arXiv:2201.11319*.
- [137] Y. Liu, C. Shu, J. Wang, and C. Shen, "Structured knowledge distillation for dense prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jun. 12, 2020, doi: [10.1109/TPAMI.2020.3001940](https://doi.org/10.1109/TPAMI.2020.3001940).
- [138] T. Wang, Y. Zhu, C. Zhao, X. Zhao, J. Wang, and M. Tang, "Attention-guided knowledge distillation for efficient single-stage detector," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.
- [139] Y. Li, Y. Gong, and Z. Zhang, "Few-shot object detection based on self-knowledge distillation," *IEEE Intell. Syst.*, early access, Sep. 12, 2022, doi: [10.1109/MIS.2022.3205686](https://doi.org/10.1109/MIS.2022.3205686).
- [140] M. Gao et al., "An embarrassingly simple approach for knowledge distillation," 2018, *arXiv: 1812.01819*.
- [141] D. Li, S. Tasci, S. Ghosh, J. Zhu, J. Zhang, and L. Heck, "RILOD: Near real-time incremental learning for object detection at the edge," in *Proc. 4th ACM/IEEE Symp. Edge Comput.*, 2019, pp. 113–126.
- [142] M. He et al., "Cross domain object detection by target-perceived dual branch distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9570–9580.
- [143] G. Li, X. Li, Y. Wang, S. Zhang, Y. Wu, and D. Liang, "Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 1306–1313.
- [144] Y. Chu, P. Li, Y. Bai, Z. Hu, Y. Chen, and J. Lu, "Group channel pruning and spatial attention distilling for object detection," *Appl. Intell.*, vol. 52, pp. 16246–16264, 2022.
- [145] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1921–1930.
- [146] L. Zhang and K. Ma, "Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–20.
- [147] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," 2021, *arXiv:2104.13921*.
- [148] C. Li et al., "YOLOv6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.
- [149] M. He et al., "Cross domain object detection by target-perceived dual branch distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9570–9580.
- [150] Q. Lan and Q. Tian, "Adaptive instance distillation for object detection in autonomous driving," 2022, *arXiv:2201.11097*.
- [151] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 21–37.
- [152] S. Miao and R. Feng, "Object-oriented relational distillation for object detection," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 1510–1514.
- [153] V. VS, P. Oza, and V. M. Patel, "Instance relation graph guided source-free domain adaptive object detection," 2022, *arXiv:2203.15793*.
- [154] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2014, pp. 740–755.
- [155] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.
- [156] Y. Zhang et al., "Prime-aware adaptive distillation," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 658–674.
- [157] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3967–3976.
- [158] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.



**Zhihui Li** received the PhD degree from the School of Computer Science and Engineering, University of New South Wales (UNSW). She is a research professor with Shandong Artificial Institute, Qilu University of Technology. Her research interests are machine learning and its applications to computer vision and multimedia. She has published more than 40 high-quality papers in top-tier venues, three of which are identified as ESI highly cited papers.



**Pengfei Xu** received the PhD degree from Xidian University in 2014. He is an associate professor with the School of Information Science and Technology, Northwest University, Xi'an, China. He has spent most of his time working on computer vision and pattern recognition and has published more than 50 papers in international journals and conferences, including *IEEE Transactions on Pattern Analysis and Machine Intelligence*/*IEEE Transactions on Image Processing*, *IEEE Transactions on Geoscience and Remote Sensing* and *IJCAI* etc. He won the first prize of Shaanxi Provincial Natural Science Award in 2020. His research on facial recognition and behavior analysis of Golden Monkey was interviewed by CCTV-13.



**Xiaojun Chang** (Senior Member, IEEE) is a professor with the Australian Artificial Intelligence Institute, University of Technology Sydney. He is also an honorary professor with the School of Computing Technologies, RMIT University. Before joining UTS, he was an associate professor with the School of Computing Technologies, RMIT University, Australia. He has spent most of his time exploring multiple signals (visual, acoustic, textual) for automatic content analysis in unconstrained or surveillance videos. He achieved top performance in various international competitions, such as TRECVID MED, TRECVID SIN, and TRECVID AVS.



**Luyao Yang** is currently working towards the master's degree with the School of Information Science and Technology, Northwest University, Xi'an, China. Her main research interests include deep learning and computer vision. Currently, she is focusing on animal skeleton detection.



**Yuanyuan Zhang** is currently working towards the master's degree with the School of Information Science and Technology, Northwest University, Xi'an, China. Her main research interests include deep learning and computer vision. Currently, she is focusing on object detection and knowledge distillation.



**Lina Yao** (Senior Member, IEEE) is a senior principal research scientist and a Science Lead for Translational Machine Learning @ CSIRO's Data61. Her current research interests include data mining and machine learning with applications to Internet of Things, information filtering and recommending, human activity recognition, and brain-computer interface.



**Xiaojiang Chen** received the PhD degree in computer software and theory from Northwest University, Xi'an, China, in 2010. He is currently a professor with the School of Information Science and Technology, Northwest University. His mainly works have been published in international journals and conferences, including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE/ACM Transactions on Networking*, AAAI, etc, and were nominated for the ACM CCS 2018 Best Paper Award and Sensys2019 Best Paper Award. His current research interests include localization and performance issues in wireless ad hoc, mesh, sensor networks and machine learning.