IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

*Digital Object Identifier 10.1109/ACCESS.2023.0322000*

# A Review of Knowledge Distillation in Object Detection

**SHENGJIECHENG[1], QIUXIAZHAO[2], XINYUNZHANG[3], NURBIYA YADIKAR[1], and KURBAN UBUL[1,3,*], (Member, IEEE)**

[1]School of Information Science and Engineering, Xinjiang University, Ürümqi 830046, China
[2]School of Software, Xinjiang University, Ürümqi 830046, China
[3]Key Laboratory of Xinjiang Multilingual Information Technology, Xinjiang University, Ürümqi 830046, China

Corresponding author: Kurban Ubul (kurbanu@xju.edu.cn)

**ABSTRACT** Target detection is a revolutionary advancement in computer vision that provides the ability to identify specific targets in images for a wide variety of applications, including but not limited to video surveillance, face recognition, and autonomous driving. Although target detection has been developed to a high level and can be deployed for applications in several fields, there are still some problems in practice, such as the two-phase detection algorithm has high detection accuracy but slow detection speed, while the one-phase detection algorithm is fast, but its accuracy is poor. We need to combine their respective advantages further for related algorithm research, and we need to reach a balance between detection speed and detection accuracy so that the algorithms can be deployed on edge devices with limited computational power. Knowledge Refinement, as one of the common means of model compression, can solve the above problems, and it reduces the deployment difficulty of the algorithms. In this paper, we summarize the use of knowledge compression on target detection. We conclude the methods mentioned in the paper from an objective and unbiased perspective and suggest possible improvement directions, and we provide an outlook on the future trend of combining distillation learning and target detection. This paper provides a clear overview of the field of target detection and provides an idea of future trends.

**INDEX TERMS** target detection, knowledge distillation, deep learning.

## I. INTRODUCTION

LATELY , with the rapid development of deep learning, the performance of object detection has been greatly improved. Target detection can now be divided into three branches: two phases [1], [2], [3], [4], one phase [5], [6], [7], and transformer-based [8]. A two-stage detector using RPN extracts regions of interest, which are then sent to the prediction network for localization and classification of instances. Single-stage detectors perform localization and classification of instances directly on the obtained feature maps. The transformer-based detector segments feature maps into small blocks of images (patches) and then encodes each segmented block of images (patches) to form a position embedding, which is then sent to the transformer module for training, and finally localizes and classifies instances in the prediction layer. Nevertheless, every detector is not perfect, for example, the detection speed of two phases algorithm is slow, the detection accuracy of one phase algorithm is poor, and the transformer-based algorithm needs to consume gigantic arithmetic power, which brings strict constraints for the use of the algorithms. Recently, with the wide-spread use of mobile devices and embedded, model light-weighting has become one of the hottest directions in the development of target detection. Knowledge compression uses transfer learning, which distills the knowledge learned from sophisticated models to tiny models, which will greatly reduce the calculation expenses and the convergence speed of the models, which can alleviate the difficulty of deploying models on mobile devices and help realize real-time and low-power computer vision applications. Knowledge distillation is of great importance for light-weighting target detection algorithms.

## II. BACKGROUND

The improvement in accuracy against the detector comes at the cost of huge computational overhead and reduced inference speed of the detector. It is not a simple task to deploy a high-accuracy detector on an arithmetic-constrained device. How to reduce the complexity of the detector and speed up the inference of the detector has attracted many scholars to study this problem. The general solution is Neu-

ral Architecture Search (NAS) [9], [10], [11], [12] which automatically searches for the optimal architecture, pruning [13], [14], [15] which reduces the complexity by removing redundant parameters from the network, quantization [16], [17], [18], [19] which is a strategy of sharing weights, there is also knowledge distillation, which uses migration learning as the easiest and easiest way to compress models. NAS is to find the best one among many architectures, and it can only use the existing methods. Pruning requires many experiments to remove redundant branches from the network. Quantification requires specific software and hardware support. By contrast, knowledge distillation is the easiest way to compress models.

In 2006, knowledge distillation was first proposed by Bucilu˘ et al [20]. Hinton et al [21] used the method for compressed neural networks, using soft labels from a teacher's network to guide students' learning. The idea of knowledge distillation is very simple. Training a large model with good performance as a network of teacher is used to guide a small model of learning, hoping to transfer the knowledge of teacher to student. Firstly, knowledge distillation was applied to image classification, but it is much more complicated than classification, and the effect of using the distillation method on the classification task directly on the detection task is not very obvious, so the distillation method needs to be redesigned. It raises the question, what kind of knowledge can better help students learn? FitNets [22] pointed out that the teacher's intermediate layer knowledge could be used to guide students in their learning, and since then more and more feature-based distillation algorithms [23], [24] have been proposed. Through experiments, we also found that learning response-based [25] and relationship-based [26], [27], [28] knowledge can also help students to achieve good grades. With the help of knowledge distillation, we have greatly improved the performance of small models without a great modification of the model structure [29]. It is to reduce the scale of the model with minimum cost [29]. Knowledge distillation has developed rapidly in recent years and has played an important role in compressing models.

## III. IMPLEMENTATION OF KNOWLEDGE DISTILLATION ON TARGET DETECTION

This subsection focuses on the distillation of knowledge of isomorphism and heterogeneity.

### A. ISOMORPHIC KNOWLEDGE DISTILLATION

This subsection describes isomorphic distillation in which the teacher-student pairs have the same network structure, the difference being that they have different network depths. [30], [31] found that the structural knowledge present in distillation can be better transferred only if the structures of the teacher-student pairs are similar.

LEKD [32] is the first time that knowledge distillation is used for target detection with multiple classifications. It uses the Faster R-CNN framework and performs knowledge distillation in three parts backbone, RPN, and RCN, respectively. The following challenges exist in target detection.

a. The target detection task label is more informative, and the model learned based on the label is more complex and has more loss after compression.

b. There is a problem of category imbalance, with more background categories

c. The task is more complex, with both category classification and location regression.

d. The existing knowledge distillation is mainly for data in the same domain, and for the task of cross-domain target detection, there is a higher requirement for knowledge distillation.

To cope with these issues, the authors propose LEKD, an end-to-end detection framework based on knowledge migration. For the backbone network, hint learning in FitNet is used for distillation, and adaptation layers are added to make the dimensionality of feature maps match. Combining the knowledge distillation approach with RCN and RPN networks, a cross-entropy loss with weights is used to solve the problem that the object is not in balance with the target object. Finally, we use the output of the teacher network regression as an upper bound, and if the student network's regression output performs better than the teacher network, then it will not trigger the loss function to penalize it. The structure of LEKD is shown in Figure 1.

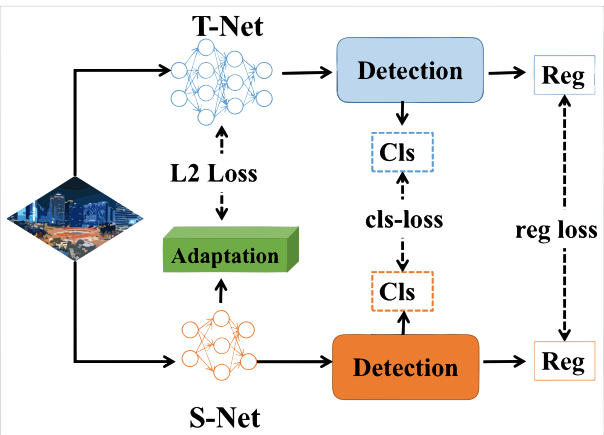



**FIGURE 1.** It represents the distillation detection process of LEKD [32].

The backbone networks for detection models are mostly from the classification side, although these backbone networks are good at classification tasks and may not be the best choice for detection tasks. If we want to design and find a more effective detection network, we need to train these models on ImageNet, and then refine them for detection tasks, which will extend the training time of the models and the authors also found that small networks trained on ImageNet have poor results, which will be on lead to worse detection accuracy of small models. Therefore, the authors proposed the Mimic [33] feature-based distillation method, which uses a trained large network to guide the learning of the small detection network. It uses the region of interest generated by the large network to guide the small network to generate its region of interest, and if the sizes of their feature maps do

not match, a conversion layer needs to be added to the small network to tune its size to match the large network. If you want to get better results, you can also add a large network of supervision on the classification and regression head. The architecture of Mimic is shown in Figure 2.

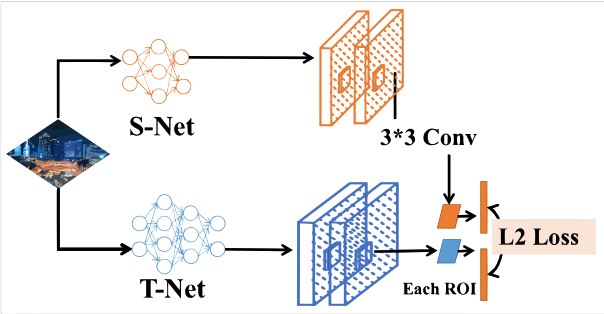



**FIGURE 2.** It represents the overall network structure of Mimic [33].

The previous approach of PAD [34] focused on studying that part of knowledge that should be transferred in distillation, treating all samples equally during training, such as hard samples that were assigned larger weights in [35], [36], hard samples that were not easy to learn, for which students struggled to learn in distillation and did not significantly improve student accuracy. To address the above deficiencies, PAD is proposed to explore which sample is more important, and in distillation will assign large weight values to simple samples and small weight values to hard samples. By modeling the data uncertainty knowledge, each of the samples is measured by a Gaussian distribution N($\mu$ , $\sigma$). Students' predictions are expressed in terms of the mean $\mu$ , and the uncertainty in extracting knowledge from the teachers' network is estimated using the variance $\sigma$ , adding a variance branch in the student branch compared to the original algorithm to provide the network with the ability to assess uncertainty. The difference between PAD and the previous method is shown in Figure 3.

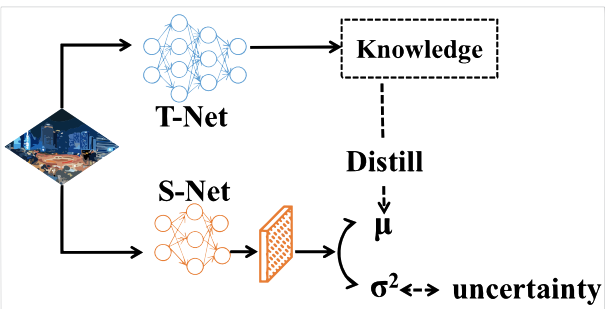



**FIGURE 3.** It represents the general architecture of the PAD network [34].

If students were to imitate all the features on the teacher's feature map equally, it would lead to a decrease in student performance. The authors believe that this is due to the large number and complexity of features in the detection model and the excessive noise in the background. Therefore, the authors propose FGFI [37] to consider fine-grained distillation of only the anchors near the GT. These knowledge-intensive locations are first located, and the student model is allowed to mimic the teacher's features at these locations, directly ignoring the background features. The schematic diagram of FGFI is shown in Figure 4.

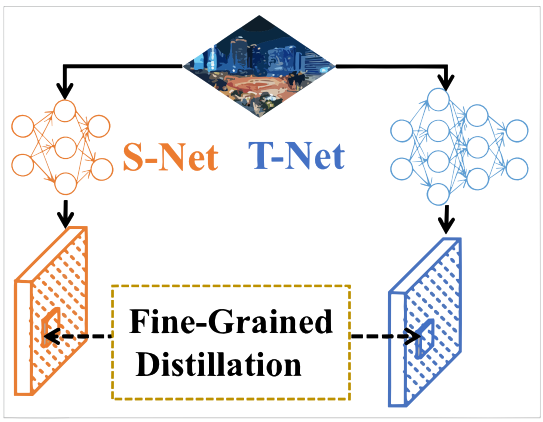



**FIGURE 4.** It demonstrates the principles of FGFI [37] and emphasizes the importance of features in the vicinity of GT.

Networks for target detection are more complex, especially for two-stage networks with many combinations of modules. The authors proposed a target detection distillation method(TARD) for task-level adaptive regularization [38], distilling the model from three modules: backbone network, classification head, and regression head, respectively. The architecture of TARD is shown in Figure 5.

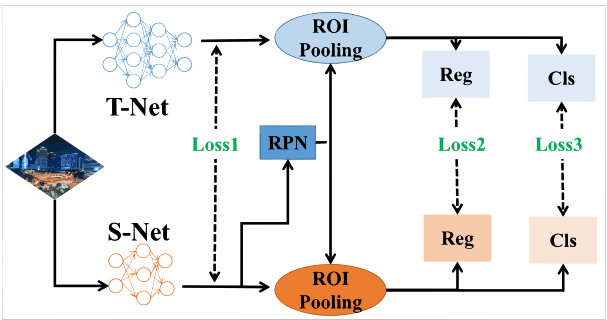



**FIGURE 5.** It represents the general architecture diagram of TARD [38], and shows the location of the distillation.

For the backbone network, the features of the region where the foreground target is located in the feature maps are first extracted using the gaussian mask for feature distillation. The gaussian mask is used to emphasize the central area of the target and to reduce the noise introduced by the background.

The boxes predicted by the teacher and student RPNs are different because of the large number of negative samples in the classification prediction, which introduces noise in distillation, and the authors propose to share the region of interest extracted by the student and teacher RPNs on the prediction classification branch. Because the teacher network does not have a priori knowledge of the prediction frames provided by the students, its classification results have some generalization ability. Using only the classification knowledge of

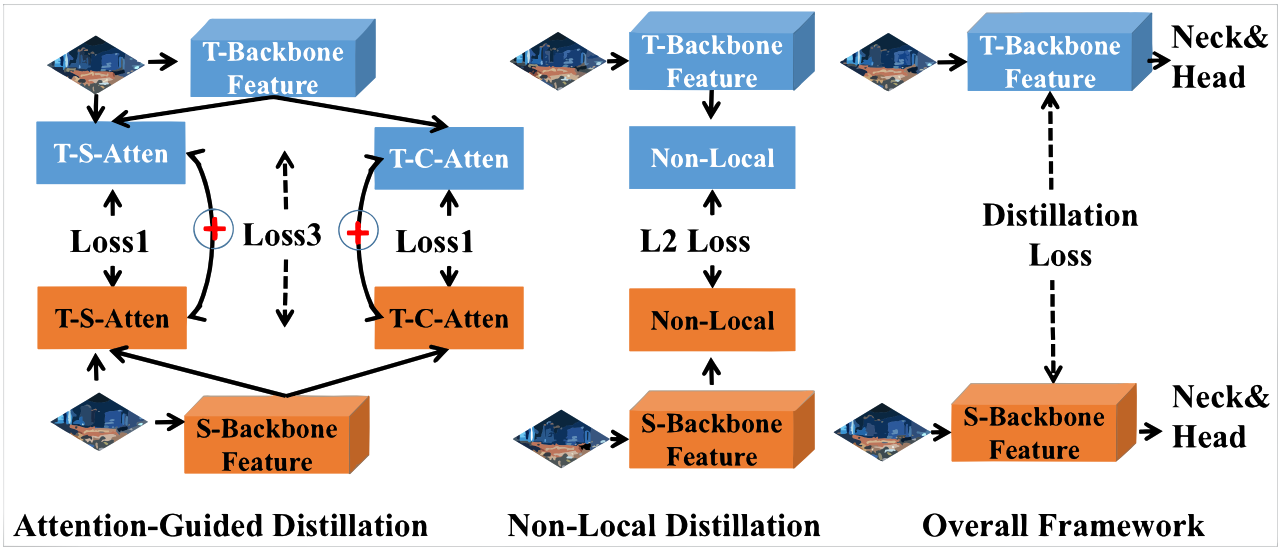



**FIGURE 6.** It represents the distillation principle of TAED [39], which is composed of local distillation and attention-directed distillation.

the positive sample to guide students' learning improves the distillation effect.

Due to the unbounded nature of regression prediction, the teacher's predicted knowledge may be wrong or even contrary to GT; therefore, the authors propose adaptive distillation on the predicted regression branch. If the prediction of the teacher's network is correct, it is used as knowledge to guide students' learning, otherwise, it is not distilled. The correctness of the teacher's prediction is determined by calculating the value of the IOU.

To address the lack of balance among foreground and background pixels and the lack of not extract the relationship among different pixels in target detection, the authors proposed TAED [39], which introduces attention-directed distillation and non-local distillation to solve the problem. Attention-directed distillation is the equivalent of drawing focus for students by using attention mechanisms to discover key pixels of foreground objects. The non-local distillation method teaches students that the features of not only individual pixels, but as well as the relational features of the global context are worth learning. The attention-led distillation method yields spatial and channel attention by averaging pooling in the channels and space dimensions, respectfully. The non-local distillation method catches the pixel relationships in an image with a non-local module. The individual parts of TAED and the overall architecture are shown in Figure 6.

FGFI believes that the background information includes too much noise and that imitating all features of the teacher's ground equally would harm the student's performance, so only features near the GT are distilled. TADF uses the instances in the feature map masked with Gaussian and the positive samples in the detection head as the student network learning objects. The authors experimentally verified that feature information in the background is also essential to the learning of student models and that it is undesirable to discard background features arbitrarily. Since the target and background regions contain different information, the authors proposed a new decoupled feature-based distillation algorithm DeFeat [40], which assigns different weights to different regions.

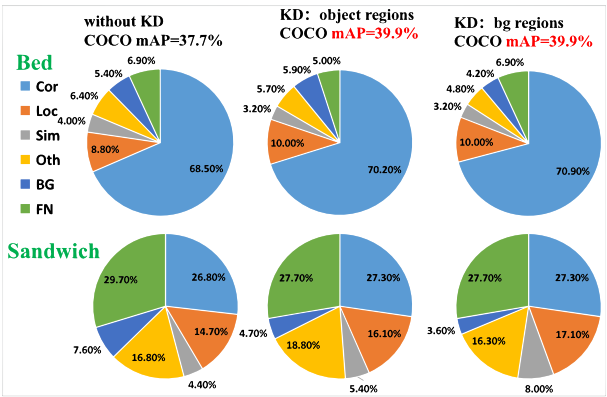



**FIGURE 7.** It proves that it was wrong to discard the background arbitrarily before, and that background knowledge is also important for detection during distillation.

Through experiments, the authors found that distillation through the background region can obtain comparable results to distillation through the object region only, so the background region in feature maps can complement the target region, and the two are complementary. As shown in Figure 7, it can be found that the positive sample information helps to enhance the precision of results, while the information on the background helps to reduce false detections.

In feature maps, a mask is used to divide the features into a target region and a background area, and the knowledge of these two parts is distilled to students separately. When distilling the classification loss of positive and negative proposal samples obtained from RPN separately, the loss of a

positively proposing sample is usually greater than the loss of a negatively proposing sample, and if the two losses are not balanced, the loss of negative samples will be largely ineffective, and thus the effect is worse than no distillation. The method used by the authors is to separate the positive and negative proposals from each other and calculate the losses separately. The architecture of DeFeat is shown in Figure 8.

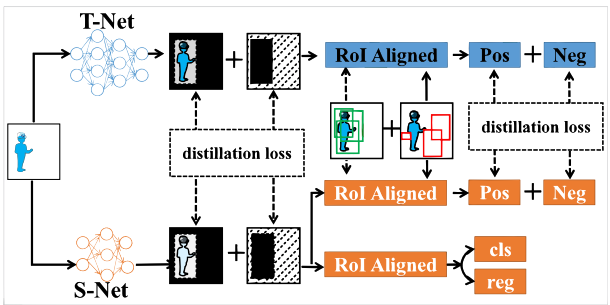



**FIGURE 8.** **The overall network structure of DeFeat [40], which separates foreground and background features, is then distilled separately.**

Before the DSIG [41] method, distillation methods in target detection were pixel-based. Distillation in classification tasks was based on point-to-point and got good results. However, in the objective detection task, we are more concerned with the concept of instance, and the relationship between these instances is critical when we do detection. Due to the neglected relationship between instances and the imbalance of characteristics. It is more suitable than the previous mask based on the foreground area designed by hand because it naturally represents a focus of the network. The meaning of an edge inside the graph structure is a similarity between the two instances. In this paper, a simple and efficient cosine similarity is chosen to measure the similarity between instances. The final distillation is performed between the knowledge graphs of the teacher pattern and the student pattern. In other words, DSIG transforms the feature graph knowledge of the teacher model target detection into a knowledge graph and then distills the knowledge in the graph to the students. The architecture of DSIG is shown in Figure 9.

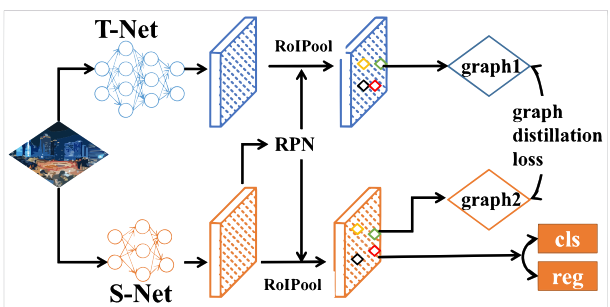



**FIGURE 9.** **It represents the overall network structure of DSIG [41], which transforms feature distillation into graph distillation.**

For unbalanced foreground and background regions with many irrelevant redundant pixels [42], previous distillation detection methods all need to carefully design the proportion of the sample that is positive or nonpositive, and distillation of only GT-related regions may ignore potential information regions in the background, previous distillation detection methods also neglect valuable relational information between instances, and current detection distillation methods are weak in generalizing to different detection frameworks and do not work well in multiple detection frameworks simultaneously. To address the above shortcomings, the authors proposed the generic instance distillation GID [25] method, which does not consider distinguishing between positive and negative samples and exploits deep teacher features, relationships between instances, and response-based knowledge that can be used in various detection frameworks. GID involves a generic instance selection module GISM, which first quantifies the teacher model prediction instances and student model prediction instances, and then select distinguishable instances for distillation. In order to make greater utilization of the knowledge provided by the teacher model, we also abstract and use knowledge based on features, relationships, and responses. The architecture of GID is shown in Figure 10.

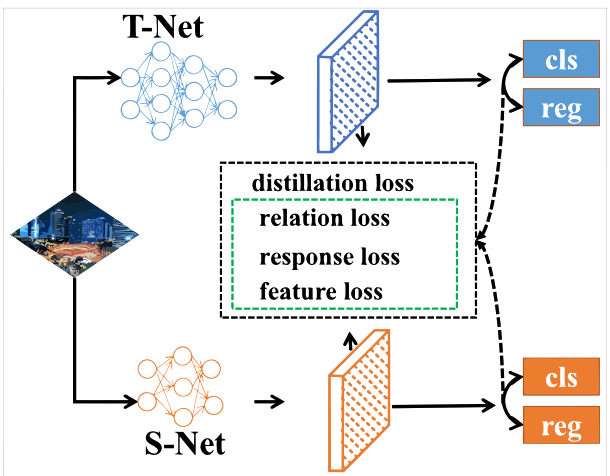



**FIGURE 10.** **It represents the workflow of GID [25]. The distillation loss of the network has three components.**

Most of the distillation before FPS [43] mimics the features near GT, which has the following problems. First, the selected foreground features from GT can only comprise the categories that already exist in the dataset, while the non-object features are ignored, so some characteristics are lost. Second, extracting only the features near GT disregard the deficiencies of the tutors' detectors, which will affect the performance of students if a category is incorrectly identified by the teacher. To solve the above problems, a novel Feature Richness Scoring (FRS) is presented to select the key characteristics of the distillation process. This approach is effective in retrieving important features outside the bounding box, and eliminating the harmful features in the bounding box. In this method, the classification score of the feature map is used as the mask for the richness of the features. Firstly, the mask is pixel-level and fine-grained, which allows us to extract student detectors in a more refined manner to enhance the performance of the

detector. Moreover, it is more applicable to detectors with FPN modules since our method is able to generate correlated feature richness masks per FPN layer of the student detectors, which are based on features retrieved from the respective FPN layers of the teacher detectors. The architecture of FRS is shown in Figure 11.

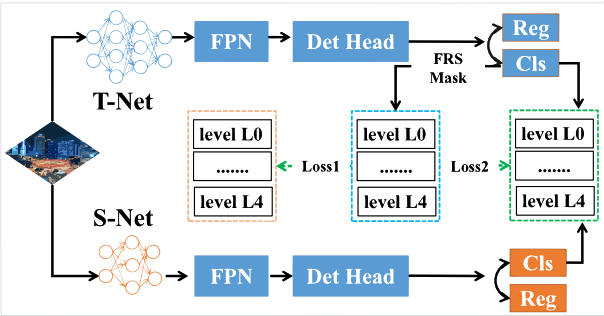



**FIGURE 11.** The overall architecture of FRS [43], where a distillation mask representing the feature richness is generated at each fpn layer.

Reducing input image resolution can increase model reasoning speed, but decreasing input image resolution will result in lower model performance. To resolve this topic, we proposed the MSRLR [44] approach, which uses a model being trained with high-resolution images as a tutor to instruct a model being trained with low-resolution images in order to enhance the capability of the small model to perform comparably to the large model for high-resolution input images, the recognition accuracy of the former is better than that of the latter for low-resolution images. The authors want to train a teacher model with multi-resolution knowledge and then distill the knowledge to the student model after the training. The teacher model is trained using high resolution images and low resolution images as input data, and Aligned Multi-Scale Training is proposed for the different sizes of feature maps so that the corresponding FPN layers get the same feature size; it is not difficult to understand that the high resolution model is better at detecting small objects, and the low resolution model is better for the detection of large models. Crossing feature-level fusion is used in MSRLR to fuse feature maps with different resolutions to give the model better compatibility with inputs of different input resolutions. The final multi-resolution teacher model obtained can transfer knowledge to students through distillation. The structure diagram of Aligned Multi-Scale Training and Crossing feature-level fusion is shown in Figure 12,13.

For network architecture and network, parameters are closely linked [45] and together affect the Network performance. In knowledge distillation, the network architectures of teacher-student pairs need to be matched [30], [31] for better transfer of dark knowledge. [46] proposed a framework Joint-DetNas for the joint optimization of NAS, pruning, and distillation. For the first time, NSA, pruning, and distillation are combined to compress a detection model, and the framework consists of student morphism and dynamic distillation. Student morphism is an approach that aims to optimize the

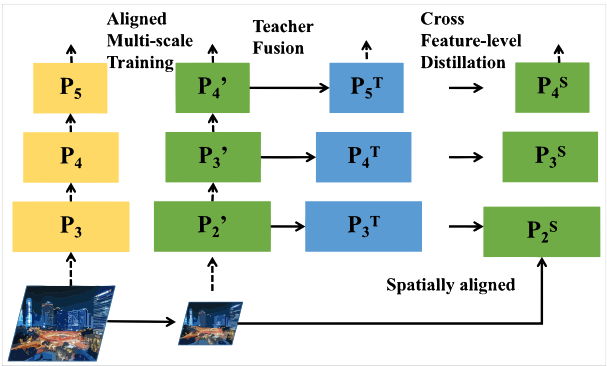



**FIGURE 12.** It shows the distillation principle of MSRLR [44], which improve the detection performance of low resolution images.

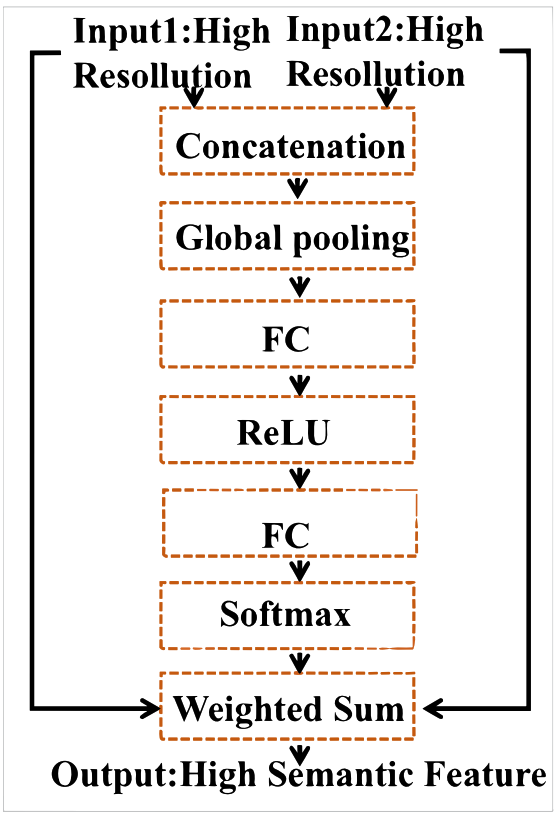



**FIGURE 13.** It shows the structure of the CFF [44] module, which dynamically fuses high-resolution images and low-resolution features.

structure of the student and remove redundant parameters. An action space and a weight succession training of strategies are elaborately designed to eliminate the prerequisites for ImageNet pre-training of the backbone network, allowing students to take full advantage of the trained weights and flexibly adapt their network architecture. Dynamic distillation constructs a pool of elastic teachers to provide a sufficiently large number of powerful detectors designed to find the best matching teachers for students and to transfer their knowledge to students. Due to the presence of the elastic teacher pool, which facilitates the search and reduces the number of training sessions for the teacher network, the de-

tector trains only one oversized network to obtain a collection of all sub-networks. During the search process, the framework employs a concise hill-climbing strategy to optimize the teacher-student pairs. Because of weight inheritance and flexible teacher pool, the evaluation time of each student pair is shorter, and no extra training is needed for the final student detector. The Joint-DetNas architecture is illustrated in Figure 14.

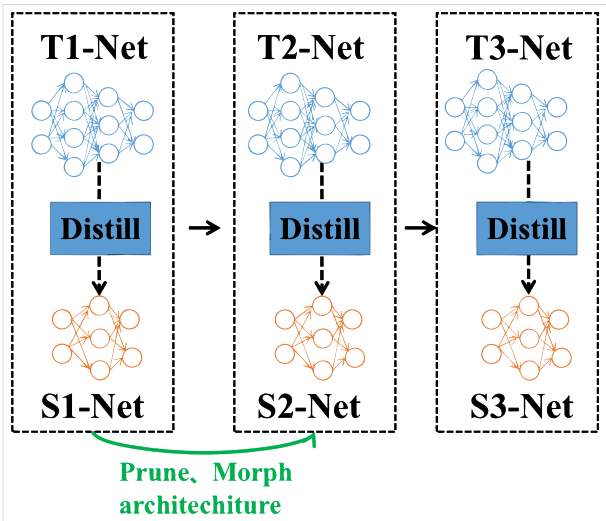



**FIGURE 14.** The overall network structure of Joint-DetNas [46], which consists of Student morphism and Dynamic distillation.

The authors found that the previous distillation used only the same level of information in the teacher-student pairs to instruct the students without fully utilizing what the teacher had learned. By analyzing the network structure, the authors found that students' high-level features can learn helpful information from teachers' low-level features, which is similar to the human learning curve [47]. To make better use of the teacher's knowledge this paper proposes a knowledge review with a ABF module and a HCL module. Knowledge review [48] allows the use of previous knowledge to guide current knowledge. In this review mechanism only the teacher's shallow features are used to supervise students' deeper features; if deeper and more abstract features are used to supervise students' learning, it is like asking a secondary school student to attend a university, this is not conducive to improving students' performance. The ABF module exploits the ideas of [49], [50], where the ABF module is able to generate different attention maps according to the input characteristics and finally fuse the features that pass through the ABF module to improve the semantic information of the features. we propose HCL based on [51], which uses a spatial pyramid pool to separate knowledge transfer into different levels of contextual information. The architecture diagram of knowledge review, ABF, and HCL are shown in Figures 15,16.

Aiming at the problem of unbalanced foreground and background in object detection, there is still no unified approach on how to select distillation regions, although there have been many research methods. Through experiments, the authors

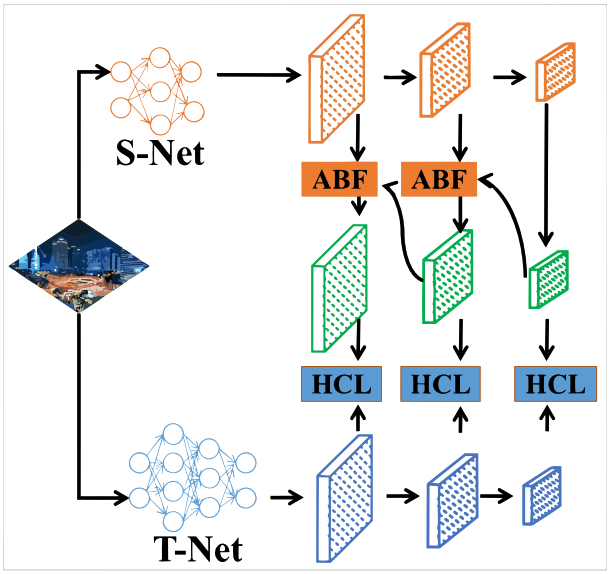



**FIGURE 15.** It represents the working process of knowledge review in the distillation process, and adds the knowledge review mechanism.

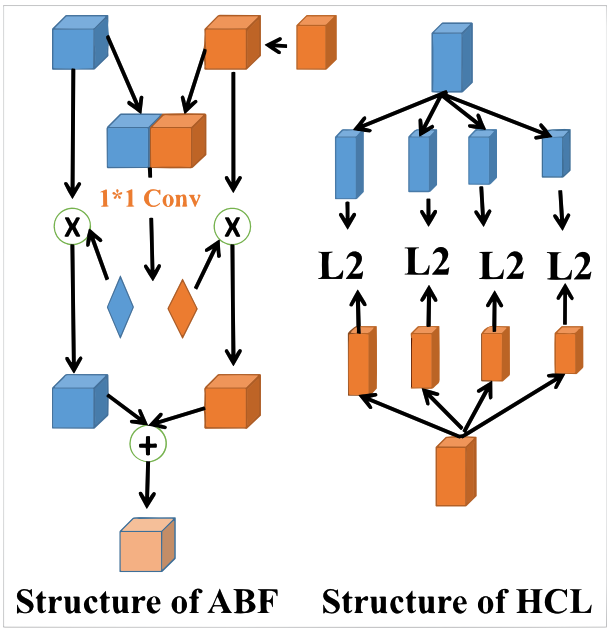



**FIGURE 16.** It represents the architecture diagram of ABF and HCL [48].

found that teachers and students focus on quite different pixels and channels. In this paper, the authors propose local and global distillation(FGD) [52]. Local distillation method splits the foreground and background, that force students to pay attention to the tutors' key pixels and important channels. Global distillation method uses GcBlock [53] to reconstruct the relationships between pixels and transmit the relational information from the teacher to the student to compensate for the global information lost in the local distillation method. Because it calculates the loss only feature-wise, FGD can be used for various detectors. The architecture of FGD is shown

in Figure 17.

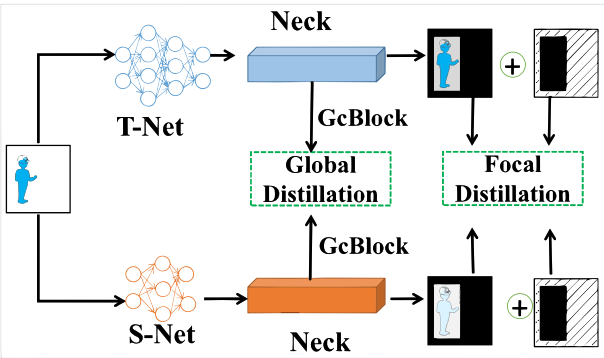



**FIGURE 17.** In the overall architecture of FGD [52], the distillation is divided into local distillation and global distillation.

Target detection is a two-stage task that first requires locating where the target is located and then, secondly, classifying the target. Among the distillation methods before LD [54], many of them only focus on how to imitate the deep features of the teacher instead of imitating the classification logit because of its inefficiency in extracting localization information. The detection efficiency is extremely poor when the edges of the objects have no clear boundaries, especially for lightweight models. The feature maps extracted by previous methods are a mixture of semantic and localization knowledge, and it is difficult to determine whether transferring mixed knowledge at each location is beneficial to performance and which regions benefit from the transfer of certain types of knowledge. For the above problem, the authors propose LD solve the edge blurring problem. Based on logit simulation, it is the first time that a new method is proposed to extract the local knowledge from the feature map, which conveys semantic and localization knowledge separately. In addition, it establishes valuable Local Areas (VLR), which help to identify effectively which areas benefit from classification or local education. First, the representation of the bounding box is transformed into a probability distribution. Then, the extraction site was determined by the area weighting of the primary distillation area and the valuable local area. Finally, the LD loss of the two models was calculated. The structure of LD is shown in Figure 18.

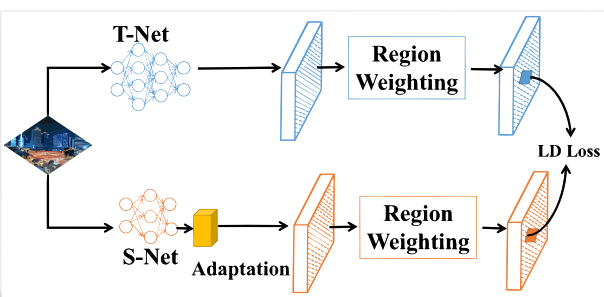



**FIGURE 18.** It represents the network structure of the LD [54], which can effectively improve the detection performance of edge-blurred objects.

During the study of teacher-student pairs of distillation, the authors found that for the same instance, teachers and students ranked the detected candidate frames significantly differently, which explains the difference in teacher-student performance. The large gap between the difference in teacher and student feature responses and the difference in predictions suggests that it is not optimal for students to imitate all of the teacher's features. Based on existing research, the authors proposed

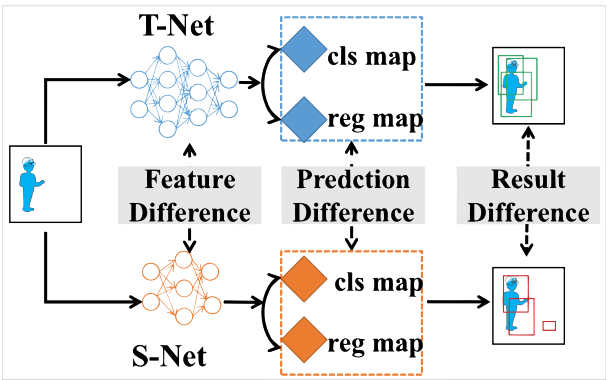



**FIGURE 19.** The overall architecture of ICD [55], it will extract the teacher's knowledge about the ordering of the prediction frames.

two methods, Rank mimics (RM) [55] and predictive-guided Feature Imitation (PFI) [55], and used in a single-stage detector, RM [55] extracts the ranking of the teacher's candidate boxes as a piece of new knowledge and passes it to the student, and then uses softmax and KL divergence to align the distribution of the teacher student's top N anchors. It was found that differences in features at locations with large prediction differences may be small, while differences in features at locations with small prediction differences may be large. Treating all features equally can lead to poor distillation. PFI [55] distilled knowledge of locations with large prediction discrepancies for teacher students. The overall architect of the network is shown in Figure 19.

In the target detection task, the local knowledge learned by teachers and students is different for instance, surrounded by the same GT, and even more so for small targets and ambiguous instances because the characteristics of these examples are sensitive to small disturbances in the feature space of teachers and students. Considering knowledge as a vector in the feature space, for this, the authors proposed the prototype generation module (PGM) [56] and robust distilling module (RDM) [56]. PGM to find a set of common basis vectors as prototypes in TSspace. Inspired by dictionary learning [57], [58], [59], prototypes are found in TSspace for each category in the dataset by the similarity between vectors. RDM is then used for robust knowledge construction and delivery of prototype-based global knowledge. Because the knowledge learned by the teacher may contain noise, RDM gives students a voice in identifying noise. On the basis of such an observation, the credible knowledge should have analogous phenotypes under the measurements of both models, and the validity of knowledge can be evaluated by the

differences in the phenotypes. The soundness of knowledge can be assessed by differences in representation in the two models. The soundness of the knowledge can be assessed by the difference in representation in the two models TSpace.The global knowledge formed by RDM has a smaller gap in TSspace, which can effectively mitigate the impact of noise on local knowledge. The structure diagram of PGM and RDM is shown in Figure 20.

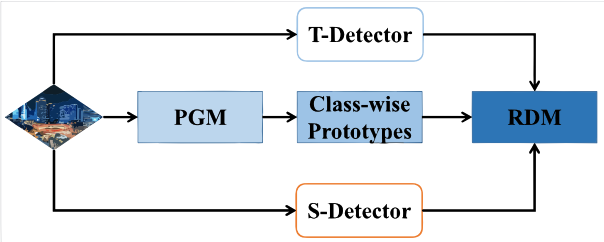



**FIGURE 20. The overall architecture of ODGK [56], which consists of two modules with different functions, PGM and RDM.**

The simplest and easiest distillation method to implement is to extract deep features from the middle layer, but it ignores the importance of logit. Compared to feature-based distillation on logit, it has the following advantages, firstly feature-based distillation method needs a more complex structure to align features, and secondly, the semantic information of logits is high-level. In this paper, we once again focus on logit distillation and decouple the knowledge distillation method proposed by Hinton [20] to analyze the shortcomings of the original method and then propose decoupled knowledge distillation(DKD) [60]. Original distillation formula: KD=TCKD+(PtT)NCKD.Through the experiments, we found that TCKD has an important role in transferring knowledge related to the difficulty of the sample and that using NCKD alone may be more effective than using KD. The relationship between teacher network confidence and NCKD is negatively correlated in the original formulation; however, it is shown through experiments that higher confidence in the teacher network's predictions is beneficial to the training process of distillation and therefore should not be suppressed. In the original formulation, TCKD and NCKD are coupled, but they convey different knowledge, and such coupling results in no way for them to flexibly adjust their respective flexibility. Therefore, the authors propose a new formula: DKD=$\alpha$TCKD+$\beta$NCKD, DKD can solve the two problems just mentioned very well: on the one hand, TCKD and NCKD are decoupled, and their respective importance can be adjusted independently, and on the other hand, NCKD, which is more important for distillation, is no longer inhibited by the high confidence level generated by the teacher, greatly improving the flexibility and effectiveness of distillation. The difference between KD and DKD is shown in Figure 21.

The feature-based knowledge distillation approach before MGD [61] allows students to imitate the teacher's features, which improves student performance but with limited performance gain. The features extracted by the backbone network

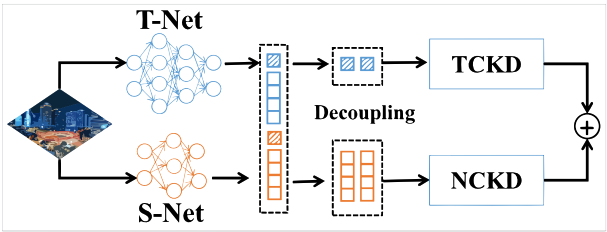



**FIGURE 21. The overall architecture of DKD [60], which decouples distillation into target category and non-target category distillation.**

have high-level semantic information, and somehow each feature pixel contains the information of the neighboring pixels [62], so we can use some of the pixels to restore the whole feature map. MGD is fundamentally different from the previous method, which is to generate teachers' feature maps rather than imitate teachers' feature maps. It performs a random mask on the student feature map to attain a partial feature map, and then a generative block is used to force students to generate the teacher's features using only their partial features to enhance the students' representational ability. The differences in attention span among the student-generated feature maps and the teacher's feature maps are still relatively large, and it proves that MGD is fundamentally different from the previous approach. The architecture of MGD is shown in Figure 22.

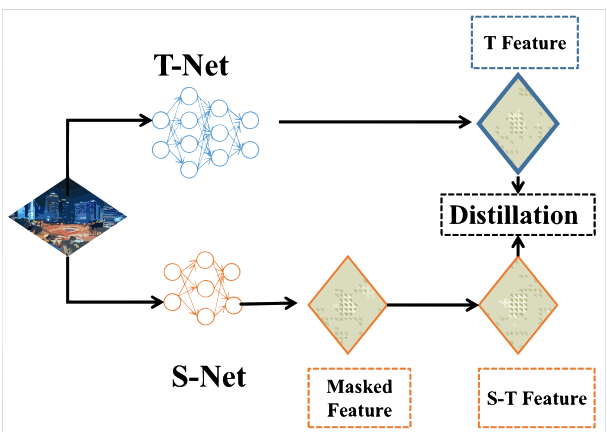



**FIGURE 22. It represents the distillation principle of MGD [61], which uses the student's feature map to generate the teacher's feature map.**

The paper is based on the distillation of intermediate features, distillation using only response-based knowledge [25], [32], [40] has been shown to be ineffective, the authors found that using the Lp parametric to calculate the distance between feature maps, does not capture the spatial relationship between features, which can lead to ignoring some information in the feature maps. For instance, the spatial relationship between features, the correlation between teacher-student features, and the importance of individual features, previous distillation methods need to design complex sampling schemes and sometimes the sampling schemes introduce some defects, based on which the authors proposed SSIM-KD [63],

which is an important metric for the similarity of two images. Overall the SSIM-KD method has three components, first, the backbone network extracts the feature maps, then feature fusion is carried out at the neck of the network, and then distance calculation is performed on the output features. In

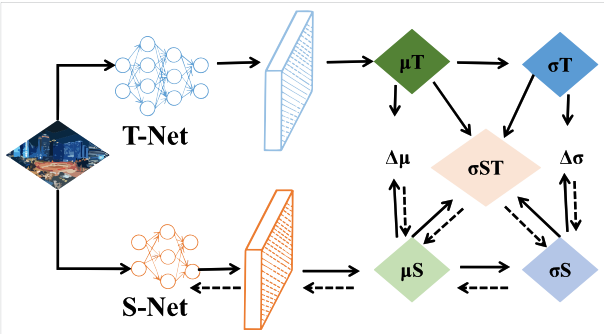



**FIGURE 23.** **It represents the framework of SSIM-KD [63], which differs from previous elaborate methods of selecting positive sample areas.**

the process of calculating distances, we no longer compare point-by-point by pixel but divide the feature map into small blocks, and then decompose the feature space into brightness (mean), contrast (variance), and structure (intercorrelation) components within each block, gradually minimizing the local differences in mean and variance and maximizing the intercorrelation between teacher features and student features during the training process, and finally the detection head for category prediction and position regression. The architecture of SSIM-KD is shown in 23.

In this paper, the authors analyze that the poor distillation results on the detection task are caused by two factors. On the one hand, an extreme imbalance between foreground and background features is one of the most frequent problems in the sphere of target detection. On the other hand, small target objects lose a lot of features during feature extraction, which leads to unfriendly detection of small targets and, therefore, poor detection performance of small targets. For the detection dilemma encountered above, the authors of this paper proposed dual relational knowledge extraction DRKD [64]. Pixel-based relationship extraction can improve the drawback of convolutional networks that only focus on local features, and student networks can learn the relationship between foreground and background features, and we extract multi-scale feature maps from teacher-student networks, and feed the feature maps to different GIoRe modules to capture global pixel relationships, which can solve the problem of feature imbalance. The authors found that small-sized target instances are closely related to other-sized instances, so they designed instance-based relationship extraction to obtain a richer feature representation for small-sized target instances. The method also uses a Gaussian function to calculate the similarity between strength features, a relationship filtering module to highlight valuable relationships and uses these relationships to complete the instance-based distillation. Distillation based on strong relationships can also alleviate the imbalance of foreground and background characteristics, and it can make the student network more focused on foreground features. The architecture of DRKD is shown in Figure 24.

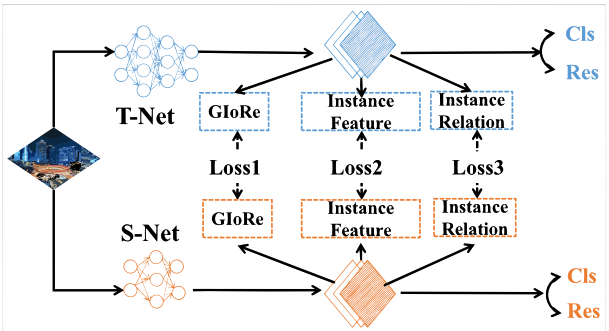



**FIGURE 24.** **The overall architecture of DRKD [64], it will perform distillation of pixel-level relation, instance and instance relations.**

### *B. HETEROGENEOUS KNOWLEDGE DISTILLATION*

This subsection introduces heterogeneous knowledge distillation on detection tasks. The network architectures of heterogeneous knowledge distillation teacher-student pairs are very different. Previous knowledge distillation on target detection has rarely studied heterogeneous knowledge distillation, yet the lightweight detectors used for deployment are usually not the same as the best detectors due to environmental resource constraints. To improve this problem, scholars have started to focus their attention on heterogeneous distillation methods.

In distillation methods based on the mask to extract information-rich foreground regions, background noise is automatically excluded. However, there is a certain accuracy gap between teachers and students trained by those methods. We have found that these approaches do not take full advantage of the semantic information of each layer of the FPN, resulting in inefficient knowledge extraction among FPN-based detectors. To address the shortcomings mentioned above, the authors propose the GDetKD [65] framework that allows distillation for both heterogeneous and homogeneous networks, and GDetKD contains within it a new semantic-guided feature simulation (SGFI [65]) technique that automatically performs soft matching between all pyramid-level feature pairs to provide optimal guidance to students, improving the consistency between teacher-student features. To go a step further, we introduce contrast extraction (CKD [65]) to efficiently capture the information encoded in the relationships among different feature areas and to achieve maximum consistency in the features representation of teacher-student. The structure diagram of SGFI and the architecture diagram of GDetKD are shown in Figures 25 and 26.

Inspired by [66], an intermediate assistant is added to the heterogeneous distillation to transfer the teacher's dark knowledge to the students. The authors proposed the HEAD [67] framework by adding an extra branch of detection head after the student backbone network, which acts as an assistant between the teacher and the student, and the assistant

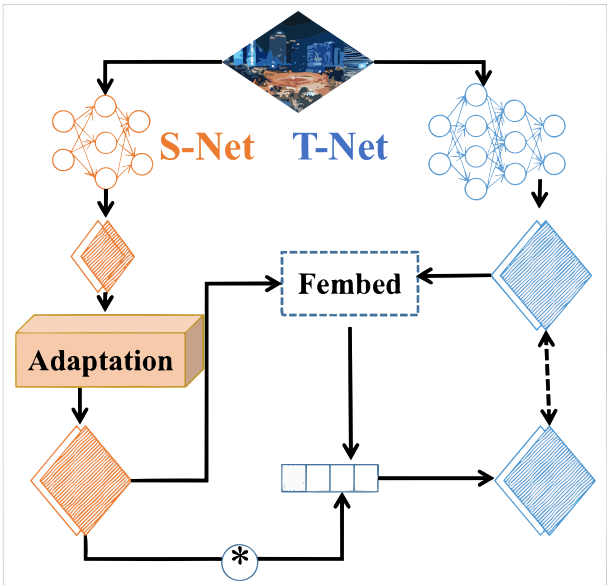



**FIGURE 25.** It represents the network structure of SGFI [65], which can automatically perform soft matching of features in the fpn layer.

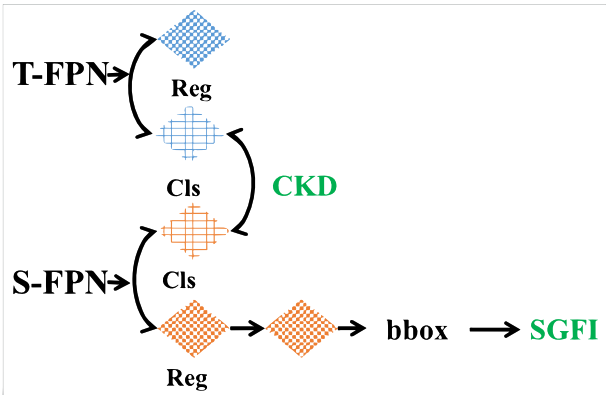



**FIGURE 26.** The general architecture of GDetKD [65], which consists of SGFI and CKD.

detection head is the same as the teacher's detection head, and also converts heterogeneous distillation into homogeneous distillation, using the proposed AKD to estimate the quality of the teacher's knowledge transfer to the assistant, which then passes it on to the student through the backpropagation in the network. As an assistant function, CKD is used to observe the students' learning results. In addition, when a good teacher detector is not available, HEAD can be extended to a teacher-free mode [68], [69], [70], [71] TF-HEAD, where several different branches of assistant detectors can be connected behind the student backbone network, and each detector head can use backpropagation to replicate the learned knowledge to the student during training, TH-HEAD can also bring significant performance gains. The architecture of HEAD and TF-HEAD are shown in Figures 27 and 28.

By examining existing methods, the authors found that there is less research on target detection for heterogeneous knowledge distillation and that previous feature-based knowl-

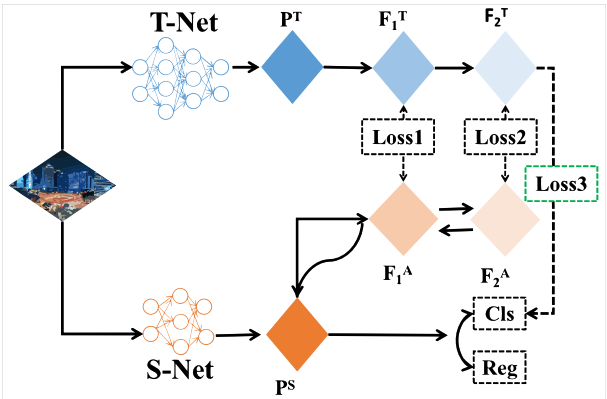



**FIGURE 27.** It represents the structural diagram of HEAD [67], which converts heterogeneous distillation into homogeneous distillation.

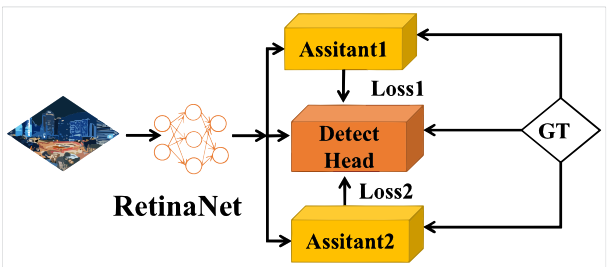



**FIGURE 28.** It represents the architecture diagram of TF-HEAD [67], where distillation can be done without a teacher model.

edge distillation methods are imperfect. Most of the previous approaches add a feature adaptation layer among the teacher-student network to align the features of the student network and the teacher network, which causes two problems and leads to suboptimal results. Firstly if the semantic information discrepancy among the teacher network and student network features is too large, direct alignment pairs will put an extra burden on the student network. Secondly, in each feature layer and FPN layer, if the activation value of that layer is large, it will easily control the convergence direction of the gradient and introduce noise with high probability. So the authors

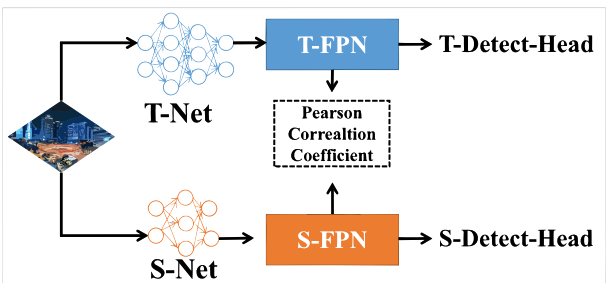



**FIGURE 29.** It represents the network structure of PKD [72], which uses PPMCC to calculate the fpn layer feature resemblance.

proposed PKD [72], which can solve the problems mentioned above. PKD uses Pearson Correlation Coefficient(PPMCC) for distillation, which is equally helpful for isomorphic distillation and heterogeneous distillation, and can accelerate

**TABLE 1. It summarizes the performance gains of the knowledge distillation technique on FasterRCNN with the dataset COCO2017, ResNet101 for the teachers' backbone network and ResNet50 for the students' backbone network.**

| Method | Distillation | mAP | AP50 | AP75 | APS | APM | APL |
|---|---|---|---|---|---|---|---|
| teacher | R152-FPN | 41.3 | | | 24.4 | 45.3 | 54.0 |
| Teacher | R101 | 39.8 | | | 22.5 | 43.6 | 52.8 |
| teacher | R101-FPN | 42.04 | - | - | 25.22 | 45.55 | 54.6 |
| student | R50 | 38.4 | | | 21.5 | 42.1 | 50.3 |
| student | R50-FPN | 37.4 | | | 21.8 | 41.0 | 47.8 |
| FGFI [37] | R152-R50-FPN | 39.9 | - | - | 22.9 | 43.6 | 52.8 |
| TADF [38] | R152-R50-FPN | 40.1 | - | - | 23.0 | 43.6 | 53.0 |
| DeFeat [40] | R152-R50-FPN | 40.9 | - | - | 23.6 | 44.8 | 53.5 |
| FBKD [39] | R101-R50 | 40.2 | 60.4 | 43.6 | 22.8 | 43.8 | 53.2 |
| GID [25] | R101-R50 | 40.2 | 60.8 | 43.6 | 23.6 | 43.9 | 53.0 |
| ODGK [56] | R101-R50 | 40.6 | 61.0 | 44.0 | 23.4 | 44.4 | 53.3 |
| FitNet [22] | R101-R50-FPN | 38.76 | 59.62 | 41.80 | 22.32 | 42.20 | 50.70 |
| ReviewKD [48] | R101-R50-FPN | 40.36 | 60.97 | 44.08 | 23.60 | 43.81 | 52.87 |
| FRS [43] | R101-R50 | 39.5 | 60.1 | 43.3 | 22.3 | 43.6 | 51.7 |
| CKD [65] | R152-R50-FPN | 40.3 | - | - | 17.2 | 34.8 | 43.0 |
| G-DetKD [65] | R152-R50-FPN | 41.0 | - | - | 18.1 | 36.6 | 44.5 |
| DKD [60] | R101-R50-FPN | 39.25 | 60.90 | 42.73 | - | - | - |
| DKD+ReviewKD [60] | R101-R50-FPN | 40.65 | 61.51 | 44.44 | - | - | - |
| LD [54] | R101-R50 | 42.1 | 60.3 | 45.6 | - | - | - |
| DSIG [41] | R101-R50 | 41.55 | 62.15 | 45.27 | 24.44 | 45.34 | 53.95 |
| TAED [39] | R101-R50 | 43.9 | 64.2 | 48.1 | 25.3 | 48.0 | 58.7 |
| SSIM-KD [63] | R101-R50 | 40.9 | 61.0 | 44.9 | 23.7 | 44.5 | 53.5 |

**TABLE 2. It summarizes the performance gains of the knowledge distillation technique on FasterRCNN with the dataset COCO2017, ResNeXt101 for the teachers' backbone network, and ResNeXt50 for the students' backbone network.**

| Method | Distillation | mAP | AP50 | AP75 | APS | APM | APL |
|---|---|---|---|---|---|---|---|
| Teacher | R152-RetinaNet | 40.1 | | | 24.1 | 43.9 | 52.8 |
| Teacher | R101-RetinaNet | 38.5 | | | 20.6 | 42.6 | 51.6 |
| Student | R50-RetinaNet | 36.9 | - | - | 19.9 | 40.7 | 49.0 |
| Teacher | RX101-RetinaNet | 41.0 | | | - | - | - |
| student | RX50-RetinaNet | 37.4 | | | - | - | - |
| ICD [55] | R101-50-Retina | 39.6 | - | - | 21.4 | 44.0 | 52.5 |
| FGD [52] | R101-50-Retina | 39.7 | - | - | 22.0 | 43.7 | 53.6 |
| MGD [61] | RX101-50-Retina | 41.0 | - | - | 23.4 | 45.3 | 55.7 |
| PKD [71] | R101-50-Retina | 40.8 | 60.3 | 43.4 | 23.0 | 45.1 | 54.7 |
| DRKD [64] | R101-50-Retina | 40.3 | 59.7 | 42.9 | 23.4 | 44.2 | 53.4 |

network convergence, PKD improves the stability of training. The method works only on the FPN layer, first normalizing the teacher-student ground features, then using PCC to determine the correlation between features and selecting the most useful features for distillation. Based on previous experimental experiences, features with larger activation values can be more helpful in distillation, and features with smaller activation values play a small role. The network structure of PKD is shown in Figure 29.

## IV. RESULT COMPARISON

Here, we count some results obtained by knowledge distillation in the field of target detection. The datasets VOC07+12 and MSCOCO. VOC and COCO are commonly used for target detection and image segmentation.VOC 07+12 has 20 object categories, COCO contains 164K images, and COCO has 80 object categories, which are challenging and authoritative. The detection network is the classical two-stage algorithm FasterRCNN, where the teacher-student backbone network is either a resnet101-resnet50 combination or a RetinaNet101-RetinaNet50 combination. We summarize the results of knowledge distillation for target detection over the past 5 years, and we compare the advantages and disadvantages of each distillation method from a fair perspective. From the table below, we can clearly see that knowledge distillation can significantly raise the performance of mini-models without changing the structure of the network, and knowledge distillation can make a great contribution to the deployment of future projects.

From the data in the tables, we can find that there are three types of knowledge that can be used for knowledge distillation, feature-based, response-based, and relationship-based knowledge.

In our perception, the images are background except for the detected objects, and the computation of background features wastes a lot of computational resources and seriously affects the detection performance, but [40], [43] found that background features are also helpful for detection performance, and in addition, current knowledge distillation methods ignore the role of categories outside the dataset for target detection, and arbitrarily discarding background features is an unwise choice. The intermediate layer features

**TABLE 3. It summarizes the performance gains of the knowledge distillation technique on FasterRCNN with a VOC dataset, ResNet101 for the teacher backbone network and ResNet50 for the student backbone network.**

| Method | Distillation | mAP(%) |
|---|---|---|
| Teacher | R101-FPN | 82.13 |
| Student | R50-FPN | 80.53 |
| Student | ResNet50(ROI Align) | 71.7 |
| Teacher | R101 | 74.30 |
| student | R50 | 70.00 |
| student | R50(ROI Align) | 71.7 |
| FitNet [22] | R101-R50-FPN | 82.2 |
| Mimic [33] | R101-R50-FPN | 80.90 |
| FGFI [37] | R101-R50-FPN | 82.10 |
| FBKD [39] | R101-R50-FPN | 82.00 |
| ICD [55] | R101-R50-FPN | 82.40 |
| GID [25] | R101-R50-FPN | 82.60 |
| Mimic + PAD [34] | R101-R50-FPN | 81.11 |
| FGFI + PAD [34] | R101-R50-FPN | 81.25 |
| DeFeat [40] | R101-R50-FPN | 81.47 |
| ODGK [56] | R101-R50-FPN | 82.90 |
| LD [54] | R101-R18-FPN | 75.90 |
| TADF [38] | R101-R50 | 74.50 |

utilized in knowledge distillation want to prefer the student feature maps to be as similar as possible to the teacher feature maps by calculating a distance between the feature maps and adding the distance values to the loss function and then using backpropagation to improve the degree of similarity between teacher feature maps and student feature maps. Response-based knowledge is for students to imitate the soft output of the teacher's network, using additional information about the similarities within and between classes, which is the simplest and most efficient way to handle this, but this would be missing the supervisory information in the middle layer, which is not a very significant performance gain for the student network. The most commonly used in current knowledge distillation is the combination of feature-based and response-based knowledge [25], [38], [40], [50], [53], [73]. Relationship-based knowledge [41], [43], [74], [75], [76], [77]is an extension of response-based knowledge and feature-based knowledge, and it studies the relationship between different layers and samples more comprehensively and deeply.

## V. CONCLUSION

The development of target detection algorithms can be broadly divided into 3 time periods: conventional target detection cycle, deep learning target detection cycle, and lightweight cycle of target detection model. Conventional detection algorithms rely heavily on the design of manual features and the robustness of the network is not good, so the application is very limited. Deep learning-based target detection algorithms have qualitatively improved performance compared to traditional algorithms, but they suffer from large network sizes and too many parameters, which makes the deployment of the model extremely difficult. For the deployment of detection models, model lightweight has become a trend. Experimental data indicate that there are many redundant neurons in neural networks, and the neurons that are helpful for the final detection only account for 5-10% of the overall neurons. We have reason to believe that we can achieve the same performance as big models if we make the best use of small models.

The current problems of knowledge distillation are as follows. In the first place, the positive and nonpositive samples in the image are extremely unbalanced. If students are allowed to fully imitate the teacher's feature map, it may bring down the performance or get suboptimal results. Secondly, the majority of detection networks are designed based on deep learning, and the features after convolutional neural networks will focus more on local features and destroy the global contextual relationships. Finally, in knowledge distillation, the performance of the student network rises relatively small when the teacher's network depth gap is large. For the problem mentioned above, we are ready to use the attention mechanism and weight assignment strategy to solve the problem of positive and negative sample imbalance. We use the attention mechanism to discover important spatial features and channel features, and then fuse the features before and after the attention mechanism and assign different weights to them so that the important features are retained without losing much background information. Inspired by Transformer, for the problem that convolutional networks corrupt global contextual relationships, we put the convolutional network and Transformer in one block so that global-local contextual relationships can be learned to enhance the characterization of features. For the problem of performance increase degradation caused by too large a gap between the depth of the networks of teacher-student pairs, we intend to add a helper intermediate layer among the teacher-student networks to transform the knowledge extracted by the teacher and solve the problem that the distance between teacher features and student features is too great to cause the student networks to fail to understand the knowledge distilled from the teacher.

## VI. FUTURE OUTLOOK

Knowledge distillation on detection tasks has achieved a lot in recent years, and it has significantly reduced the cost of deploying detection models. I believe that knowledge distillation on detection tasks may have the following aspects.

In practical applications, the application of heterogeneous knowledge distillation is larger than that of homogeneous distillation due to the limitation of the environment. Can we make a unified definition of the transferred knowledge to fundamentally solve the dilemma that heterogeneous networks cannot perform direct distillation?

Nowadays, the most commonly used distillation methods are offline distillation and online distillation, and the performance of students trained for offline distillation still has a gap with teachers. How can we make the best use of the knowledge passed by teachers to enable students to achieve the results of online distillation training?

The existing knowledge distillation methods rarely make use of the attention mechanism, which is the same as the

original intention of knowledge distillation, both of which are the teacher network pointing out the key knowledge for the student network, so that the student network takes less or no detours. In addition, the attention mechanism can also address global contextual relationships that are corrupted by convolutional networks, and experimental results have shown that utilizing relationship-based knowledge in the distillation process is beneficial to the performance improvement of student networks.

With the computing advantages of network platforms and computers, many new detection algorithms are end-to-end based so that users do not need to be concerned with how to extract features, how to calculate candidate frames, how to classify and localize targets but simply feed the data into the network and get detection results. Yet, the current process of detective distillation frameworks generally involves training the teacher network first and then using it to guide the learning of the student network. This is not quite consistent with the original intention of designing end-to-end detection algorithms, and it also increases the workload. Therefore, we would like to design end-to-end knowledge distillation algorithms, where the user only needs to provide data.

At the moment, the development of detection algorithms have faced bottlenecks, and it is necessary to exploit ways to improve the detection accuracy of existing models, lead to an expansion of its generalization ability, and enhance their deployment efficiency in real-world applications. In this context, knowledge refinement becomes increasingly important because it has the potential to enhance the performance and applicability of existing models. Therefore, it has become a promising research topic for scholars in this field in recent years.

## ACKNOWLEDGMENT

This work was supported by National Key Research and Development Program of China (No.2021YFB2802100) and National Science Foundation of China under Grant (62266044, 62061045, 61862061).

## REFERENCES

[1] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," *ArXiv*, vol. abs/1605.06409, 2016.

[2] R. B. Girshick, "Fast r-cnn," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, 2015.

[3] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask r-cnn," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[4] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.

[5] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, "You only look one-level feature," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13034–13043, 2021.

[6] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2020.

[7] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017.

[8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *ArXiv*, vol. abs/2005.12872, 2020.

[9] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "Detnet: A backbone network for object detection," *ArXiv*, vol. abs/1804.06215, 2018.

[10] G. Ghiasi, T.-Y. Lin, R. Pang, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7029–7038, 2019.

[11] C. Jiang, H. Xu, W. Zhang, X. Liang, and Z. Li, "Sp-nas: Serial-to-parallel backbone search for object detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11860–11869, 2020.

[12] L. Yao, H. Xu, W. Zhang, X. Liang, and Z. Li, "Sm-nas: Structural-to-modular neural architecture search for object detection," 2019.

[13] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," *ArXiv*, vol. abs/1608.08710, 2016.

[14] X. Li, Y. Zhou, Z. Pan, and J. Feng, "Partial order pruning: for best speed/accuracy trade-off in neural architecture search," *IEEE*, 2019.

[15] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient transfer learning," 2016.

[16] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding," *arXiv: Computer Vision and Pattern Recognition*, 2015.

[17] R. Li, Y. Wang, F. Liang, H. Qin, J. Yan, and R. Fan, "Fully quantized network for object detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] Y. Wei, X. Pan, H. Qin, and J. Yan, "Quantization mimic: Towards very tiny cnn for object detection," *ArXiv*, vol. abs/1805.02152, 2018.

[19] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng, "Quantized convolutional neural networks for mobile devices," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4820–4828, 2015.

[20] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

[21] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Computer Science*, vol. 14, no. 7, pp. 38–39, 2015.

[22] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *CoRR*, vol. abs/1412.6550, 2014.

[23] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge transfer via distillation of activation boundaries formed by hidden neurons," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 3779–3787, 2019.

[24] Z. Huang and N. Wang, "Like what you like: Knowledge distill via neuron selectivity transfer," *arXiv preprint arXiv:1707.01219*, 2017.

[25] X. Dai, Z. Jiang, Z. Wu, Y. Bao, Z. Wang, S. Liu, and E. Zhou, "General instance distillation for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7842–7851, 2021.

[26] Y. Liu, J. Cao, B. Li, C. Yuan, W. Hu, Y. Li, and Y. Duan, "Knowledge distillation via instance relationship graph," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7096–7104, 2019.

[27] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

[28] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1365–1374, 2019.

[29] J. Tang, R. Shivanna, Z. Zhao, D. Lin, A. Singh, E. H. Chi, and S. Jain, "Understanding and improving knowledge distillation," *arXiv preprint arXiv:2002.03532*, 2020.

[30] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4794–4802, 2019.

[31] Y. Liu, X. Jia, M. Tan, R. Vemulapalli, Y. Zhu, B. Green, and X. Wang, "Search to distill: Pearls are everywhere but not the eyes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7539–7548, 2020.

[32] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," *Advances in neural information processing systems*, vol. 30, 2017.

[33] Q. Li, S. Jin, and J. Yan, "Mimicking very efficient network for object detection," in *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 6356–6364, 2017.

[34] Y. Zhang, Z. Lan, Y. Dai, F. Zeng, Y. Bai, J. Chang, and Y. Wei, "Prime-aware adaptive distillation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pp. 658–674, Springer, 2020.

[35] B. Li, Y. Liu, and X. Wang, "Gradient harmonized single-stage detector," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 8577–8584, 2019.

[36] M. Zhao, H. Wu, D. Niu, and X. Wang, "Reinforced curriculum learning on pre-trained neural machine translation models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 9652–9659, 2020.

[37] T. Wang, L. Yuan, X. Zhang, and J. Feng, "Distilling object detectors with fine-grained feature imitation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4933–4942, 2019.

[38] R. Sun, F. Tang, X. Zhang, H. Xiong, and Q. Tian, "Distilling object detectors with task adaptive regularization," *arXiv preprint arXiv:2006.13108*, 2020.

[39] L. Zhang and K. Ma, "Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors," in *International Conference on Learning Representations*, 2021.

[40] J. Guo, K. Han, Y. Wang, H. Wu, X. Chen, C. Xu, and C. Xu, "Distilling object detectors via decoupled features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2154–2164, 2021.

[41] Y. Chen, P. Chen, S. Liu, L. Wang, and J. Jia, "Deep structured instance graph for distilling object detectors," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4359–4368, 2021.

[42] C. Tung, A. Goel, X. Hu, N. Eliopoulos, E. S. Amobi, G. K. Thiruvathukal, V. Chaudhary, and Y.-H. Lu, "Irrelevant pixels are everywhere: Find and exclude them for more efficient computer vision," in *2022 IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pp. 340–343, IEEE, 2022.

[43] D. Zhixing, R. Zhang, M. Chang, S. Liu, T. Chen, Y. Chen, *et al.*, "Distilling object detectors with feature richness," *Advances in Neural Information Processing Systems*, vol. 34, pp. 5213–5224, 2021.

[44] L. Qi, J. Kuen, J. Gu, Z. Lin, Y. Wang, Y. Chen, Y. Li, and J. Jia, "Multi-scale aligned distillation for low-resolution detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14443–14453, 2021.

[45] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Training pruned neural networks," 2018.

[46] L. Yao, R. Pi, H. Xu, W. Zhang, Z. Li, and T. Zhang, "Joint-detnas: upgrade your detector with nas, pruning and dynamic distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10175–10184, 2021.

[47] L. E. Yelle, "The learning curve: Historical review and comprehensive survey," *Decision sciences*, vol. 10, no. 2, pp. 302–328, 1979.

[48] P. Chen, S. Liu, H. Zhao, and J. Jia, "Distilling knowledge via knowledge review," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5008–5017, 2021.

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

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[51] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.

[52] Z. Yang, Z. Li, X. Jiang, Y. Gong, Z. Yuan, D. Zhao, and C. Yuan, "Focal and global knowledge distillation for detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4643–4652, 2022.

[53] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pp. 0–0, 2019.

[54] Z. Zheng, R. Ye, Q. Hou, D. Ren, P. Wang, W. Zuo, and M.-M. Cheng, "Localization distillation for object detection," *arXiv preprint arXiv:2204.05957*, 2022.

[55] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 1306–1313, 2022.

[56] S. Tang, Z. Zhang, Z. Cheng, J. Lu, Y. Xu, Y. Niu, and F. He, "Distilling object detectors with global knowledge," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pp. 422–438, Springer, 2022.

[57] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural computation*, vol. 15, no. 2, pp. 349–396, 2003.

[58] I. Tošić and P. Frossard, "Dictionary learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27–38, 2011.

[59] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on signal processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[60] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11953–11962, 2022.

[61] Z. Yang, Z. Li, M. Shao, D. Shi, Z. Yuan, and C. Yuan, "Masked generative distillation," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XI*, pp. 53–69, Springer, 2022.

[62] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

[63] P. De Rijk, L. Schneider, M. Cordts, and D. Gavrila, "Structural knowledge distillation for object detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3858–3870, 2022.

[64] Z. Ni, F. Yang, S. Wen, and G. Zhang, "Dual relation knowledge distillation for object detection," *arXiv preprint arXiv:2302.05637*, 2023.

[65] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3591–3600, 2021.

[66] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 5191–5198, 2020.

[67] L. Wang, X. Li, Y. Liao, Z. Jiang, J. Wu, F. Wang, C. Qian, and S. Liu, "Head: Hetero-assists distillation for heterogeneous object detectors," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pp. 314–331, Springer, 2022.

[68] M. Ji, S. Shin, S. Hwang, G. Park, and I.-C. Moon, "Refine myself by teaching myself: Feature refinement via self-knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10664–10673, 2021.

[69] K. Kim, B. Ji, D. Yoon, and S. Hwang, "Self-knowledge distillation with progressive refinement of targets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6567–6576, 2021.

[70] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3903–3911, 2020.

[71] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3713–3722, 2019.

[72] W. Cao, Y. Zhang, J. Gao, A. Cheng, K. Cheng, and J. Cheng, "Pkd: General distillation framework for object detectors via pearson correlation coefficient," *arXiv preprint arXiv:2207.02039*, 2022.

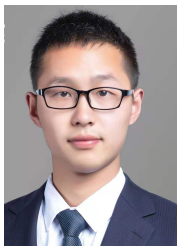
[73] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

[74] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, 2018.

[75] J. Li, G. Tan, X. Ke, H. Si, and Y. Peng, "Object detection based on knowledge graph network," *Applied Intelligence*, pp. 1–22, 2022.

[76] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3588–3597, 2018.

[77] K. Ramakrishnan, R. Panda, Q. Fan, J. Henning, A. Oliva, and R. Feris, "Relationship matters: Relation guided knowledge transfer for incremental learning of object detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 250–251, 2020.



**SHENGJIE CHENG** was born in Xuchang ,Henan, China, in 1998. He obtained the bachelor's degree in Computer Science and Technology from the School of Computer Science (Software) of Pingdingshan University, China, in 2020. He is currently pursuing a master's degree in Computer Science and Technology at the School of Information Science and Engineering, Xinjiang University, China. His main research interests include object detection, image processing and knowledge distillation.

**QIUXIA ZHAO** was born in Urumqi, Xinjiang, China, in 1995. She obtained a Bachelor's degree in Computer Science and Technology from the School of Information Management, Xinjiang University of Finance and Economics, China in 2019. She is currently pursuing a master's degree in software at the School of Software, Xinjiang University, China. Her main research interests include object detection and image processing.

**XINYUN ZHANG** was born in 2000, is an undergraduate in Xinjiang University. Her current main research interests include knowledge distillation and object detection. She applied the College Students' Innovative Entrepreneurial Training Plan Program in 2020 and successfully completed it. As a collaborator, she published a Chinese core paper and an IE retrieval paper. As the student first inventor, she has been granted a utility model patent and a software copyright.

**NURBIYA YADIKAR** was born in Urumqi, Xinjiang, China, in 1970. She received the B.S. degree in electronic information technology from Xinjiang University, China, in 1993, and the M.S. degree in advanced cyber technology from the Kyoto Institute of Technology, Japan, in 2004. She is currently an Assistant Professor (Lecturer) with the School of Information Science and Engineering, Xinjiang University. She has authored more than 30 articles. Her research interests include image processing, pattern recognition, and signal processing. She is a member of the Pattern Recognition Committee for China Artificial Intelligence Association (CAAI-PR), and the China Computer Federation (CCF).

**KURBAN UBUL** (Member, IEEE), professor, doctoral supervisor, director of the China Computer Federation (CCF) and China Society of Graphic Graphics (CSIG), vice chairman of CCF Urumqi, chairman of NOI Xinjiang Competition Committee, deputy secretary-general of the Pattern Recognition and Machine Intelligence Professional Committee (CAA-PRMI) of the Chinese Association of Automation, member of CCF Computer Vision Professional Committee (CCF-CV) and Artificial Intelligence and Pattern Recognition Special Committee (CCF-AI) Member of the CSIG Document Analysis and Recognition Professional Committee (CSIG-DIAR), Member of the Pattern Recognition Professional Committee of the Chinese Artificial Intelligence Society (CAAI-PR) and the Multilingual Intelligent Information Processing Professional Committee (CAAI-IMLIP), Senior Member of CCF, Member of IEEE, IAPR, ACM, IAENG, Director of the Electronic Department of the School of Information Science and Engineering at Xinjiang University, and "Double First Class" Department of the Computer Science Group at Xinjiang University, Research platform - responsible for real-time 3D reconstruction and understanding of large-scale scenes, and responsible for the national first-class undergraduate program "Communication Engineering". Reviewer for journals such as IEEETPAMI, Neurocomputing, and IEEETHMS. Served as the Chairman of the CCBR2018 Program Committee, PRCV2019/IJCB2021 Domain Chairman, Local Chairman of NCIG2020/CPCC2022/CCFAI2023, and Conference Chairman of NLPAI2020/2021/2022; Served as a procedural committee member or reviewer for multiple meetings such as CVPR, ICPR, ICDAR, CCFAI, CCBR, etc; Led one national key research and development project, four national natural science foundation projects, and more than 10 provincial and ministerial level and other projects. Published more than 160 papers in domestic and foreign journals and conferences, published three academic monographs, more than 10 patents, more than 40 computer software copyrights, and won 8 provincial and ministerial level or above awards.