Expert Systems With Applications 240 (2024) 122553

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Research on knowledge distillation algorithm based on Yolov5 attention mechanism☆

ShengjieCheng [a], Peiyong Zhou [a], YuLiu [a], HongjiMa [a], Alimjan Aysa [b,c], Kurban Ubul [a,c,*]

[a] School of Computer Science and Technology, Xinjiang University, Urumqi 830046, China
[b] School of Software, Xinjiang University, Urumqi 830046, China
[c] Key Laboratory of Xinjiang Multilingual Information Technology, Xinjiang University, Urumqi 830046, China

ARTICLE INFO

Keywords:
Target detection
Knowledge distillation
Deep learning
Feature acquisition
Migration learning
Model compression

ABSTRACT

The current most advanced CNN-based detection models are nearly not deployable on mobile devices with limited arithmetic power due to problems such as too many redundant parameters and excessive arithmetic power required, and knowledge distillation as a potentially practical model compression approach can alleviate this limitation. In the past, feature-based knowledge distillation algorithms focused more on transferring the local features customized by people and reduced the full grasp of global information in images. To address the shortcomings of traditional feature distillation algorithms, we first improve GAMAttention to learn the global feature representation in images, and the improved attention mechanism can minimize the information loss caused by processing features. Secondly, feature shifting no longer defines manually which features should be shifted, a more interpretable approach is proposed where the student network learns to emulate the high-response feature regions predicted by the teacher network, which increases the end-to-end properties of the model, and feature shifting allows the student network to simulate the teacher network in generating semantically strong feature maps to improve the detection performance of the small model. To avoid learning too many noisy features when learning background features, these two parts of feature distillation are assigned different weights. Finally, logical distillation is performed on the prediction heads of the student and teacher networks. In this experiment, we chose Yolov5 as the base network structure for teacher–student pairs. We improved Yolov5s through attention and knowledge distillation, ultimately achieving a 1.3% performance gain on VOC and a 1.8% performance gain on KITTI.

1. Introduction

Target detection algorithms have entered a rapid development phase with great success in the past decades. Especially as deep learning techniques have been widely applied and continuously improved in recent years, detection algorithm development is at an unprecedented level. Detection algorithms based on deep learning can be broadly characterized into uni-stage detection algorithms (Lin, Goyal, Girshick, He, & Dollár, 2017; Liu et al., 2016; Redmon, Divvala, Girshick, & Farhadi, 2016), bi-stage detection algorithms (Girshick, Donahue, Darrell, & Malik, 2014; He, Gkioxari, Dollár, & Girshick, 2017; Ren, He, Girshick, & Sun, 2015), and Transformer-based detection algorithms (Carion et al., 2020; Zhang et al., 2022), which are favored by enterprises and many scholars and have become the mainstream branch of algorithms in detection tasks.

The performance of deep-learning based detector algorithms is qualitatively improved compared to the detection performance of traditional detector algorithms that are dependent on artificial feature extractions. Deep learning models have a strong nonlinear fitting capability, which allows them to sustain competitive detection results even when working on detection tasks in complex environments. Some detection algorithms have achieved impressive results in unique detection environments, achieving real-time detection. Examples include autonomous driving, real-time industrial defect detection, and real-time drug analytics, although these often leverage high computational resources and exceptional equipment support, making it challenging to deploy such sophisticated detection algorithms on cameras or other computationally constrained devices. Although there have been significant advances in the detection accuracy of current detector algorithms,

☆ This document is the results of the research project funded by the National Key Research and Development Program of China (No. 2021YFB2802100), National Science Foundation of China under Grant (62266044, 62061045, 61862061) and 2022 Department Linkage project of Xinjiang Uyghur Autonomous Region, China (2022B03035).

* Corresponding author at: School of Computer Science and Technology, Xinjiang University, Urumqi 830046, China.
E-mail addresses: shengjie@stu.xju.edu.cn (ShengjieCheng), kurbanu@xju.edu.cn (K. Ubul).

https://doi.org/10.1016/j.eswa.2023.122553
Received 15 August 2023; Received in revised form 15 October 2023; Accepted 10 November 2023
Available online 15 November 2023

their development has still encountered obstacles. Recent innovations in detection algorithms are mainly focused on two branches: on the one hand, stacking models add various modules to the detection network, which can lead to improved detection results of the model, nevertheless, the drawback of this strategy is that the model becomes larger and requires more computing resources, which in turn increases the difficulty of project deployment; on the other hand, scholars have proposed light-weight networks, such as SqueezeNet (Iandola et al., 2016) and ESPNet (Mehta, Rastegari, Caspi, Shapiro, & Hajishirzi, 2018), to supplant the backbone network of detection algorithms to speed up detection. Such light-weight networks often work at the expense of inferior detection quality. Neither of these two innovative ideas can simultaneously combine detection speed and detection accuracy. Increasing performance of algorithms for practical applications, the design of detection algorithm innovation should focus on the relationship between detection speed and detection accuracy, which attracts many scholars' research, and they proposed pruning (Ghosh, Srinivasa, Amon, Hutter, & Kaup, 2019; Tang et al., 2021; Xie et al., 2020), weight quantization (Li, Wang, Liang et al., 2019; Wei, Pan, Qin, Ouyang, & Yan, 2018; Yang et al., 2020), weight sharing (Dai, Zhuang et al., 2021; Zhang, Wang, Figueiredo, & Balzano, 2018), and knowledge distillation (Guo et al., 2021; Wang, Yuan, Zhang, & Feng, 2019) to solve the above problems. Pruning can directly shrink the scale of the models, though the pruning algorithm is poorly interpretable and requires different pruning schemes for different models, in addition to increasing the risk of model overfitting; weight quantization has limited improvement in detection performance and requires special hardware; weight sharing reduces the size of the weight file but does not fundamentally scale down the model; By comparison, knowledge distillation has the unique advantage of using the idea of migration learning to enhance detection accuracy without changing the original network architecture. This approach enables small models to balance the detection accuracy of large models with higher detection speed, and distilling knowledge that is an accurate method of model compression improves the generalization ability and robustness of the model.

CNN-based networks have good features extraction and classification capabilities, which make excellent contributions to the iterative upgrading of detection algorithms, such networks focus more on local features extraction, and the convolution process is performed pixel by pixel, which is not conducive to capturing the global contextual relationship features in the image and may lead to detection that does not achieve optimal results. Some researchers use large convolution kernels or dilated convolutions (Yu & Koltun, 2015) to obtain more global information, but it can significantly burden the model. To cope with this problem, we modified the existing channel-space attention mechanism to model the global characteristics of the image with a slight increase in calculation overhead. In global relationship modeling, some information would be lost in channel compression and conversion recovery. To compensate for the lost information, we allocate varied proportional weight factors to the features before and after the attention module for feature fusion to obtain more expressive feature maps.

Knowledge distillation keeps refreshing the best detection results on the goal detection task, scholars also prefer bi-stage detection algorithms as teacher–student pair networks, few people use knowledge distillation techniques in the Yolo model. Throughout this article, we utilize Yolov5s as the student model and Yolov5l as the teacher model with attention improvement and knowledge distillation operations. Feature-based approaches occupy the majority of traditional knowledge distillation detection algorithms, where the transferred features are those that one specifies within the Ground Truth(GT), Wang et al. (2019) pointed out that features near the GT are also beneficial for detection, and for this reason, we proposed to limit the features to a specific place no longer, instead deciding which features to transfer based on the location of the teacher's response features and the location of the GT. Guo et al. (2021) and Zhixing et al. (2021) affirmed the background features' importance, but Guo et al. (2021) directly separating the instances within the GT from the background would destroy the relationship between pixels; We also consider that images are not only a collection of discrete pixel points but that relationships exist between individual pixel points, making full use of the features present in the image. In summary, we propose Yolov5-AKD, which first uses attention to model global information, and then divides the feature distillation process into two steps; a first step uses the Response Feature Transfer(RFT) module to shift knowledge of the high response characteristics of the teacher's network, followed by evaluating the degree of similarity of the two feature graphs from an overall viewpoint so that as many useful features as possible can be extracted. Finally, a logical distillation operation is done in the prediction phase of the student and teacher model. The teacher's soft prediction category scores are employed to instruct the student network's training, which can help the student network better discriminate between types of objects with similar characteristics.

Contribution:

We propose Yolov5-AKD, which uses feature-based and logic-based distillation for the Yolov5s algorithm.

We present attentional feature fusion, which compensates for the global information corrupted by Convolutional Neural Network(CNN) and reduces the feature loss after attention.

A new feature transfer paradigm RFT proposed in the distillation method can effectively transfer features useful for detection while considering the importance of background features.

2. Related work

2.1. Attention mechanism

Attention mechanisms are neural networks designed to mimic the attention mechanisms of the human visual system, which helps neural network models to focus on essential features when performing feature extraction. Attention mechanisms are commonly practiced for computer vision tasks such as target detection, image classification, and semantic segmentation. Liu, Shao, and Hoffmann (2021) and Woo, Park, Lee, and Kweon (2018) presented attention that can focus on overall features and detailed features on the feature maps, adaptively learn important geometric and color characteristics on the feature graphs, increasing the interaction ability between specific information. Wang, Girshick, Gupta, and He (2018) computes the distance between each pixel point and the remaining pixel points to establish long-range dependencies so that each pixel captures the global range of information and compensates for the global contextual relationships corrupted by convolutional networks. Hu, Shen, and Sun (2018) extracted different features based on different convolutional kernels and evaluated the coefficients of each convolutional kernel. Finally, they suppressed the features learned by the convolutional kernels with small coefficients to filter out the important features. Li, Wang, Hu and Yang (2019) proposed a dynamic selection mechanism of convolutional kernels to create branches of convolutional kernels with different sizes according to the relationship between the perceptual field and the target size and then fuse the features extracted by different convolutional kernels to obtain semantically richer feature maps and further improve the detection effect. Wang et al. (2018) add a large computational overhead, and Hu et al. (2018), Li, Wang, Hu et al. (2019), Liu et al. (2021) and Woo et al. (2018) involve channel conversion when performing feature processing, which can lead to information loss. The impact of this lost information on the final detection results needs to be considered when the attention network is designed.

2.2. Object detection

Detection algorithms have been iterating rapidly over the past few years. Detection algorithms can be categorized into two groups accord-

ing to the calculation approach: traditional detection algorithms and deep learning based detection algorithms. HOG (Pang, Yuan, Li, & Pan, 2011) and SURF (Bay, Tuytelaars, & Van Gool, 2006) are conventional algorithms that are insensitive to the scale and rotation of the target and have terrible detection results for detection tasks with intricate backgrounds. The preeminent detection algorithms are currently rooted in deep learning, and each of these algorithms has distinct advantages and disadvantages, and they all have good results for specific tasks. Uni-stage detection algorithm (Lin et al., 2017; Liu et al., 2016; Redmon et al., 2016) can obtain a good inference speed because it does not need to generate and filter candidate frames, can save considerable computational expenses, and directly perform the regression of goal category and location coordinate values, so it is suitable for scenarios requiring high real-time performance, such as UAV tracking and real-time traffic flow analysis; Bi-stage detection algorithms are popular among researchers, and with the aid of sophisticated network structures and enormous computation resources, the detection accuracy of bi-stage detection algorithms (Girshick et al., 2014; He et al., 2017; Ren et al., 2015) is usually higher than that of uni-stage detection. It can be applied to face recognition and license plate recognition, which do not need high real-time capability. The RCNN series (Girshick et al., 2014; Ren et al., 2015) algorithms are representative of bi-stage detection algorithms, which can obtain more accurate target class and location information. However, since specialized algorithms (SS Uijlings, Van De Sande, Gevers, & Smeulders, 2013, RPN Ren et al., 2015) are required to generate and filter candidate frames, this will invariably augment the computational overhead of the model and also lower the inference speed of the model, will further limit the application scenarios of the model; Transformer-based detection algorithms (Carion et al., 2020; Zhang et al., 2022) have achieved good results on publicly available datasets and still have great potential to be unexplored, which treats an image as a sequence and harnesses a multi-headed attention mechanism to acquire characteristics from the whole image directly. It uses a linear layer to conduct goal classification and position prediction on the output feature sequence. The computational efficiency of Transformer-based algorithms is poor, and plenty of computational resources and storage devices are required for model inference, which is currently not suitable for project deployment and implementation in resource-constrained practical application environments. Improving the running speed of existing models while maintaining good performance is a critical issue to be addressed.

2.3. Knowledge distillation

In 2015, Hinton, Vinyals, and Dean (2015) suggested that applying knowledge distillation to the domain of classification tasks can notably raise the classification precision of minor models. Knowledge distillation is a process in which the knowledge acquired from well-trained and convergent complex models to models with fewer parameters and faster inference, which subsequently improves the performance of small models, it opens up a brand-new direction for the development of target detection. Since classification and detection tasks require different amounts of information to solve the problem, if the distillation methods in classification tasks are directly implemented in detection tasks, this often fails to achieve the expected results. Therefore, the distillation loss function needs to be redesigned for the detection task to achieve knowledge transfer. Chen, Choi, Yu, Han, and Chandraker (2017) first used knowledge distillation for detection tasks, and Chen et al. (2017) and Sun, Tang, Zhang, Xiong, and Tian (2020) proposed to do knowledge transfer on the backbone network, classification head and coordinate regression head. Wang et al. (2019) argued that background features containing too much noise would affect the detection performance gain, so proposed to use the mask to transfer features within and near the GT to the student model. Guo et al. (2021) and Zhixing et al. (2021) demonstrated the importance of background features experimentally and that the image background contains more than just noise that is harmful to detection, so that the distillation process no longer only transfers features within the GT, but also distills some of the background features. Zhang et al. (2020) modeled uncertainty knowledge for each training sample, assigning different weighting coefficients based on the difficulty of the sample, as a way to decide which samples the model should focus on learning more about to be more beneficial to the student detection accuracy. Zhang and Ma (2020) used the attention-directed distillation to solve the foreground-context feature imbalance problem and also used NonLocal (Wang et al., 2018) to reconstruct the relationship between different pixels to compensate for the global contextual features that were ruined by convolution. Yang, Li, Jiang et al. (2022) used local distillation to separate foreground and background features for students to learn key features and then used global distillation to offset the global context missing from the local distillation process. Li et al. (2022) noted that teacher–student networks learn different feature maps and differ significantly in the ranking of candidate boxes in the prediction phase of the network. Hence, the authors proposed aligning teacher–student prediction distributions to transfer knowledge of teachers' prediction box rankings to student networks. Chen, Chen, Liu, Wang, and Jia (2021) transformed instance distillation into graph distillation by converting instances in the feature maps into graph nodes and using the similarity between instances as edges of the graph. De Rijk, Schneider, Cordts, and Gavrila (2022) pointed out that the previous use of L_p parametric to calculate feature similarity ignored the spatial relationship between features and the importance of individual features. So, Structural Similarity Index(SSIM) was applied to compute the similarity of feature maps instead of the L_p function, which could better reflect the distance between feature graphs. To strengthen the representational ability of the student feature maps, Yang, Li, Shao et al. (2022) abandoned the past practice of having the student models directly mimic the feature maps of the teacher models and replaced it with using the student's partial feature maps to generate the teacher's entire feature maps. The distillation methods mentioned above require human guidance when transferring features. We cannot verify whether this part of the features is compatible with the network. We should prioritize transferring the features learned by the network to minimize the human subjective design.

3. Implementation details

Under this section, the specific implementation of Yolov5-AKD is described in detail. It consists of GAMF to extract global and partial features, then joint RFT and SSIM to transfer the response features and background features of the network to help the student network to generate semantically strong feature maps, and finally logical distillation in the prediction stage of the network by using the relationship between the categories in the dataset, and the structure of the network of Yolov5-AKD is shown in Fig. 1. The overall loss of Yolov5-AKD consists of three components: intermediate feature loss $\ell_{feature}$, logical distillation loss ℓ_{logit}, the category loss and the positional regression loss often found in frame-based detectors ℓ_{det}. The overall loss ℓ_{loss} is composed as shown in Eq. (1):

$$\ell_{loss} = \ell_{feature} + \ell_{logit} + \ell_{det} \tag{1}$$

3.1. Attention mechanism improvement

In CNN-based detection algorithms, channel and spatial features directly affect detection performance. Channel features help the model identify different target classes and extract relevant semantic information. In contrast, spatial features help the model to accurately predict the location of bounding boxes and understand the structural information in the feature maps. However, the CNN-based detection algorithm cannot grasp the global information in the image well, so the GAMAttention module is introduced for Yolov5s. The GAMAttention (Liu et al., 2021) module will lose some features when extracting global features,

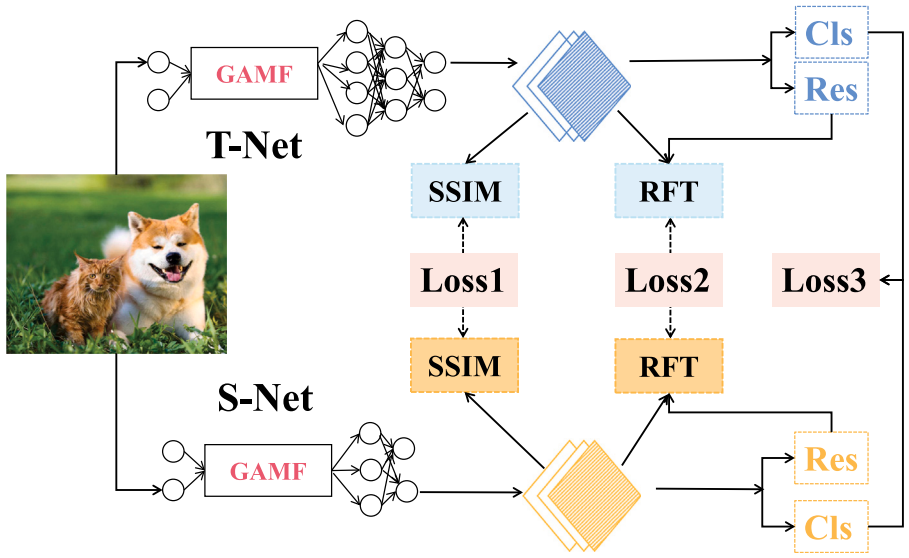


Fig. 1. It represents the Yolov5-AKD overall network structure. Improved GAMAttention is being applied to the backbone network to grab the global and local messages in the image. RFT efficiently transfers the high response features of the teacher model, joint SSIM by mimicking the feature maps of generative teacher network and logical distillation is done at the prediction head of the student–teacher networks.

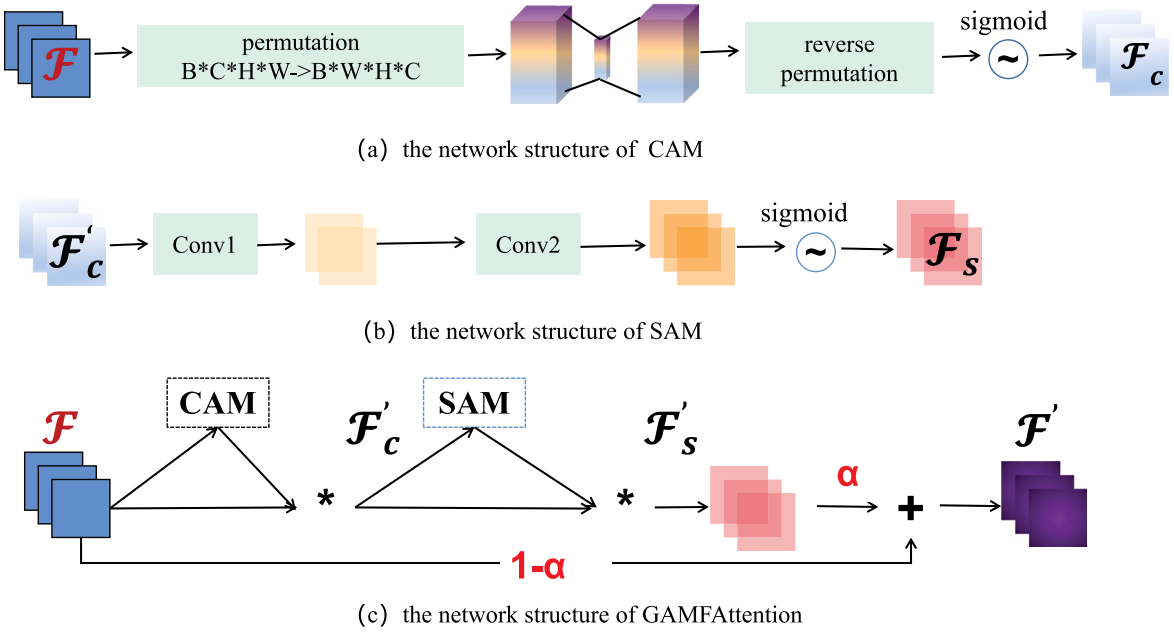


Fig. 2. It represents the network structure of GAMFAtention as well as the network structure of the individual components and highlights the differences with GAMAttention. Fig. 2.a represents the network structure of CAM, Fig. 2.b represents the network structure of SAM, and Fig. 2.c represents the network structure of GAMFAttention.

so we modify this attention network. The improved GAMAttention introduces residual connections to pass the bottom features directly to the top features, which can retain more knowledge and upgrade the characterization capability of the feature maps, naming the improved module as GAMFAttention, it is comprised of two modules serially connected: CAM (channel attention module) and SAM (spatial attention module), which will learn image features comprehensively from 3 perspectives: channel, space and channel-space, which will further deepen the network's understanding of the feature maps. CAM will learn the significance of each channel, for the input feature maps $\mathcal{F} \in \mathbb{R}^{BxCxHxW}$, it will first adjust the dimensionality of $\mathcal{F}$ and send the transformed feature maps to the MLP to calculate the channel attention score of each feature, the feature maps are converted to its raw dimensions and then sigmoid processed to get the channel attention feature map $\mathcal{F}_c$, $\mathcal{F}$ and $\mathcal{F}_c$ will do a dot product to get $\mathcal{F}_c'$ can highlight which channels are important, the channel attention is calculated as shown in Eqs. (2) and (3):

$$\mathcal{F}_c = CAM(\mathcal{F}) \tag{2}$$

$$\mathcal{F}_c' = \mathcal{F}_c * \mathcal{F} \tag{3}$$

SAM will learn the important spatial features on the feature maps, and it employs two convolutional layers for spatial feature integration, here to avoid the computational explosion growth, introduce a reduction factor r, for the convolutional processing of the features and then sigmoid processing, you can get the spatial attention feature maps $\mathcal{F}_s$, $\mathcal{F}_c'$ and $\mathcal{F}_s$ do a dot product to get $\mathcal{F}_s'$ can highlight the spatial features that should be focused on in detection, the calculation of spatial attention is shown in Eqs. (4) and (5):

$$\mathcal{F}_s = SAM(\mathcal{F}_c') \tag{4}$$

$$\mathcal{F}'_s = \mathcal{F}_s * \mathcal{F}'_c \tag{5}$$

The attention mechanism will lose some information after processing the feature maps, to compensate for the lost information, we assign different weight coefficients to the bottom feature $\mathcal{F}$ and $\mathcal{F}'_s$ the top feature, and sum the two to obtain a richer and more abstract semantic feature $\mathcal{F}'$ to remedy the information damage caused by the channel conversion, and the mathematical expression of the fused features is shown in Eq. (6), GAMF's structure is shown in Fig. 2.

$$\mathcal{F}' = \alpha\mathcal{F}'_s + (1-\alpha)\mathcal{F} \tag{6}$$

3.2. Features distillation

On the targeted detection task, the essence of feature knowledge distillation to maximize the similarity score between feature maps of teacher–student pairs, a similarity calculation function is applied to measure the degree of similarity between the feature maps as loss values and backpropagation is leveraged to update the student model parameters to refine the small model detection performance. Most scholars implicitly agree that the image features within the annotation box are the most important for the detection task. The imbalance of front-ground and background features is an essential factor that affects further performance improvement, so in previous studies of distillation algorithms (Sun et al., 2020; Wang et al., 2019; Yang, Li, Jiang et al., 2022), scholars elaborate a variety of masks, such as binary masks, Gaussian masks, etc., for refining the features within the GT to avoid the effect of background noise. However, according to the feature heat maps, certain inconsistencies may arise. The background features can be used to ameliorate the generalization and robustness of models, which is not meaningless for the detection task, and the importance of the background features needs to be considered when distilling. Guo et al. (2021) and Zhixing et al. (2021) carefully analyzed background features and found that distillation can result in significant performance gains for the model if the background features are handled well. Some people have questioned why there is still significant performance variation among networks even after learning this crucial knowledge. We share our view here that the knowledge transferred by these methods is artificially defined, and the main regions of the image are extracted using a mask, such rigid separation can break the connection between instances, and outside the mask may also contain some knowledge to be focused on. According to the shortcomings of the preceding algorithms, a new paradigm RFT is proposed to transfer the main features according to the feature response regions. In Yolov5-AKD, we divide the feature distillation into two phases, firstly, we select the high response area of the features according to the teacher's prediction box and GT box, use RFT to complete the main knowledge transfer, and secondly, we use SSIM to compare the similarity of the two graphs and align the teacher–student feature graphs as much as possible, which can make up for the shortage of RFT.

3.2.1. RFT module

We combine feature-based and response-based knowledge to propose the RFT transfer module and then transfer the response features efficiently. The specific process is as follows: Yolo will predict nine prediction frames of different sizes for each target in the feature maps $\mathcal{F}_1$. First, we evaluate the IOU value of each prediction frame and GT frame, and merge the prediction frame and GT frame corresponding to the maximum IOU value into a region, and set the pixel value of the merged region to 1 to indicate the region where the main features are located, and the other regions to 0 to indicate the background region, to obtain the mask map $\mathcal{M}$. The $\mathcal{M}$ cage region is the high response feature region of the target. Because the scale of the teacher network and student network is different, the size of the obtained feature maps are not the same, Φ_{trans} layer converts the shape of the feature maps to the same size to ensure that the later operations can be carried out,

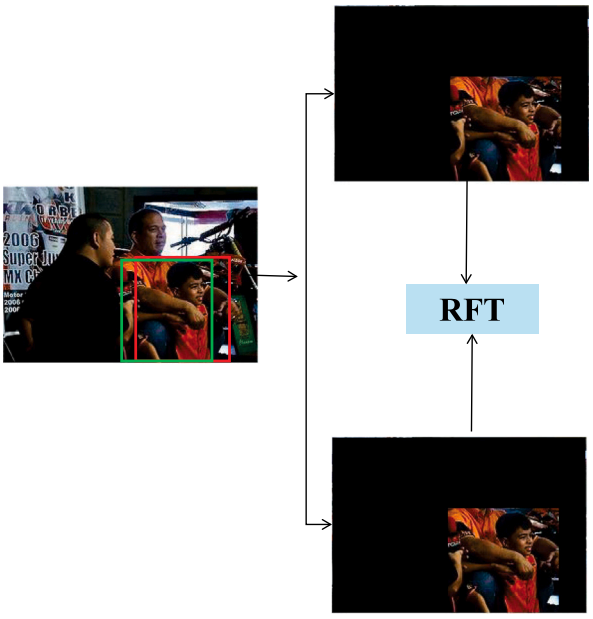


Fig. 3. It represents the process of transferring the response features by the RFT module, the red box indicates the real annotation position of the object, the green box indicates the predicted position, and $\mathcal{M}$ is produced based on the positional relationship between the two, thus accomplishing the response features transfer.

the mask $\mathcal{M}$ extracts the high response regions on the feature maps, and Mean Squared Error(MSE) calculates the resemblance scores of the high response regions of teacher features T and student features S to complete the knowledge transfer. N_{obj} = hm* wm* cm, N_{obj} refers the number of pixel points enclosed in the mask, and hm, wm, cm denotes the length, width, and number of channels of the mask, correspondingly. In an extreme case, if none of the nine prediction frames intersect with the GT frame, we only transfer the features within the GT and no longer merge the prediction frames with the GT frame. The method involves the mathematical equation shown in Eq. (7), Fig. 3 illustrates the production process of $\mathcal{M}$ and reflects the principle of RFT distillation.

$$\ell_{main-fea} = \frac{1}{2N_{obj}} \sum_{h=1}^{hm} \sum_{w=1}^{wm} \sum_{c=1}^{cm} \mathcal{M}_{h,w}(\Phi_{trans}(S_{h,w,c}) - T_{h,w,c}) \tag{7}$$

3.2.2. SSIM module

Having the student and teacher networks generate identical feature maps for the detection task is challenging. In Yolov5-AKD, we do our best to align the feature maps, and extracting semantically rich feature maps is crucial for accurate classification and location regression later. The relationship between targets is broken when transferring features using RFT, and here we use SSIM-aligned feature maps to compensate for the broken relationship features. In the past feature distillation used MSE, Peak Signal-to-Noise Ratio(PSNR), and Cosine to calculate the similarity between feature maps to complete knowledge transfer, which can also achieve the purpose of improving performance, but these functions evaluate the feature maps differences by calculating the pixel values at the corresponding positions, without considering the image as a whole. When SSIM calculates the feature maps' similarity, it no longer calculates only the corresponding position similarity scores and then accumulates them to take the average value, it will first divide the feature maps into uniform blocks, calculate the brightness, contrast, and structural similarity for each block, and finally perform a weighted average of the similarity of all blocks to get the overall SSIM index, which considers the influence of all pixels in the block on the point when calculating the difference score, establishes the

relationship that pixel points are interrelated, and can capture global features, in addition to capturing the image structure knowledge that was ignored by previous knowledge distillation. SSIM (De Rijk et al., 2022; Wang, Bovik, Sheikh, & Simoncelli, 2004) takes into account the difference between pixel values (luminance), the spatial distribution of pixels (contrast) and statistical features of the image (structure), and it can evaluate the similarity between two feature maps more comprehensively. The similarity calculated by this method is more reliable than MSE, PSNR, and Cosine, more feature knowledge can be transferred when distilling. Fig. 4 depicts the working principle of the SSIM module, the luminance L, contrast C, and structure Str of the image are calculated as shown in Eq. (8):

$$L(S,T)=\frac{2\mu_S\mu_T+C_1}{\mu_S^2+\mu_T^2+C_1}$$
$$C(S,T)=\frac{\sigma_S\sigma_T+C_2}{\sigma_S^2+\sigma_T^2+C_2} \tag{8}$$
$$Str(S,T)=\frac{\sigma_{ST}+C_3}{\sigma_S\sigma_T+C_3}$$

where μ_S, μ_T denote the mean of pixel values within the block, σ_S, σ_T denote the distribution of pixel values within the block in the dynamic range, σ_{ST} denotes the covariance within the corresponding block of the image, S, T represent the student model feature graphs and the teacher model feature graphs, C1, C2, C3 are to avoid the denominator to be 0 and ensure the division can be calculated properly. In feature distillation, we need to reduce the distance between the mean and variance and maximize the covariance to complete the knowledge transfer. In standard image processing, the larger the calculation result of SSIM indicates that the images are more similar. However, we want to use the calculation result of SSIM as part of the loss. Hence, we need to change its calculation and use the calculation result of SSIM_Loss as the loss value to complete the next step of backpropagation, and the calculation process of SSIM and SSIM_Loss is shown in Eqs. (9) and (10):

$$SSIM(S,T)=\frac{(2\mu_S\mu_T+C_1)(2\sigma_{ST}+C_2)}{(\mu_S^2+\mu_T^2+C_1)(\sigma_S^2+\sigma_T^2+C_2)} \tag{9}$$

$$SSIM\_Loss(S,T)=\frac{1-SSIM(S,T)}{2} \tag{10}$$

Using SSIM_Loss to have the student network mimic the teacher network to generate semantically richer feature maps can compensate for the connections between targets that are overlooked when transferring response features, and experiments found that $\lambda = 0.8$ and $\beta = 0.2$ worked best, with the overall loss formula for distillation of features shown in Eq. (11) (see Fig. 4):

$$\ell_{feature}=\lambda\ell_{main-fea}+\beta\ell_{SSIM\_Loss} \tag{11}$$

3.3. Logical distillation

Logical distillation refers to narrowing the distance between soft labels, representing the predicted probability distribution of categories in the dataset, compared to hard labels, which explore the association of the categories in the dataset. During distillation training, loss functions such as cross-entropy and MSE can be used to compare the gap between the output of the student model's prediction categories and the teacher's soft labels. Through soft labels, the student network can be allowed to learn how to distinguish between categories with similar features, thus increasing the forecasting precision of the student network. The super-reference distillation temperature t controls the degree of smoothing of the soft labels, which is used to control the importance between each category and enhance the model's ability to handle category-uncertain knowledge, Net represents the network, Z^S represents the actual logic of the student network for the object being category c, Z^T represents the actual logic of the teacher network

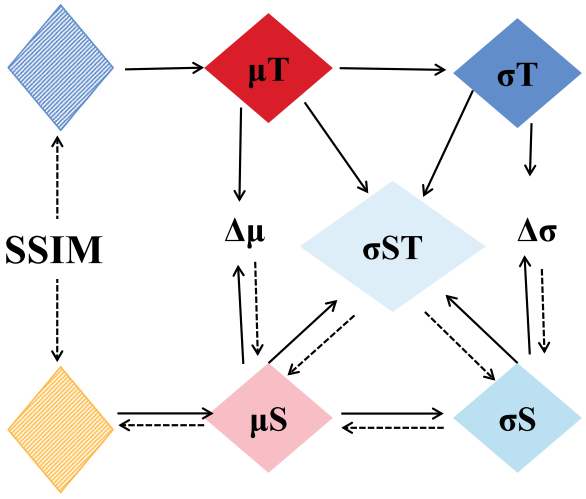


Fig. 4. It represents the network structure of the SSIM module, which accomplishes the transfer of background features, where the blue diamond box represents the feature maps of the teacher's network and the orange diamond box represents the feature maps of the student's network.

for the object being category c, P_c^S means the probability that the student network predicts the category of an object to be c when the distillation temperature is t, and P_c^T denotes the probability that the teacher's network predicts the category of an object to be c at a distillation temperature of t, C stands for the number of categories in the dataset. We use KL dispersion to measure the relative entropy of the distance between the student's logical output and the teacher's soft label, completing the logical knowledge distillation. The mathematical formulas involved in logic distillation are shown in Eqs. (12), (13), and (14):

$$P^{S,t}(c|\mathrm{Net}^S)=\frac{exp(Z_c^S/t_{obj})}{\sum_{j=1}^{C}exp(Z_c^S/t_{obj})} \tag{12}$$

$$P^{T,t}(c|\mathrm{Net}^T)=\frac{exp(Z_c^T/t_{obj})}{\sum_{j=1}^{C}exp(Z_c^T/t_{obj})} \tag{13}$$

$$\ell_{logit}=L_{kl}(P^{S,t},P^{T,t}) \tag{14}$$

4. Experiment part

4.1. Datasets

To prove this theory's validity and facilitate comparisons with other methods. VOC2007,[1] VOC2012[2] and KITTI[3] were selected to train the Yolov5-AKD model in the experiments.

VOC is one of the most authoritative datasets in target detection, which contains 20 target categories and is commonly used for tasks such as multi-categorical target detection and semantic segmentation. The trainval of VOC2007 contains 5011 images and it contains 12 608 instances, the trainval of VOC2012 contains 11 540 images and it contains 27 450 instances, and the test set of VOC2007 contains 4952 images and it contains 12 032 instances. In the experiment, VOC2007 and VOC2012 were merged, and the merged trainval was divided into training set and validation set according to the ratio of 8:2, and the test set of VOC2007 was chosen to evaluate the performance of Yolov5-AKD.

[1] http://host.robots.ox.ac.uk/pascal/VOC/voc2007.
[2] http://host.robots.ox.ac.uk/pascal/VOC/voc2012.
[3] http://www.cvlibs.net/datasets/kitti.

KITTI is commonly used for research on autonomous driving and scene understanding, which offers rich scene and annotation information, KITTI contains a total of 7481 training images and 7581 test images. For the experiments, labeled training images were used as the dataset and unlabeled test data was ignored. The training data contains a total of 7481 images and 39597 instances. Owing to the relatively small training data to minimize the data imbalance problem, we merge the target categories of KITTI, the specific changes are Truck, Van, and Tram are merged into the Car category, Person_sitting is merged into Pedestrian, and DontCare and Misc are set as the background category. The merged KITTI contains only Car, Pedestria, and Cyclist, and the data is separated into training set, validation set, and test set according to the ratio of 7:2:1, and the computation of the final evaluation indexes are done on the test set.

4.2. Evaluation indicators

We comprehensively evaluated the performances of Yolov5-AKD with respect to Precision, Recall, mAP and mAP0.5:0.95.

Precision refers to how many of the samples predicted by the model to be positive turned out to be actual positive samples. The higher the value of precision, the better the model's predictions. TP stands for the amount of samples correctly predicted as being in a positive category by the model, and FP states that it is the amount of samples incorrectly predicted as being in a positive category by the model, which is calculated as shown in Eq. (15):

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

Recall is a measure of the proportion of positive samples that the model can correctly predict, with higher Recall proving more comprehensive in finding positive samples. TP means the amount of positively categorized samples correctly predicted by the model and FN measures the amount of negatively categorized samples incorrectly predicted by the model, which is calculated as shown in Eq. (16):

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

The mAP is one of the evaluation metrics for the multi-classification detection task, where the average prediction accuracy is averaged for each category, and a higher mAP value represents a better detection performance of the model. R denotes the recall of the model, with P representing the precision of the model, c indicating the number of dataset categories, which is calculated as shown in Eq. (17):

$$AP = \sum_{i=0}^{n-1} (R_{i+1} - R_i) * P(R_{i+1})$$
$$mAP = \frac{\sum_{j=1}^{c} AP_j}{c} \tag{17}$$

mAP0.5:0.95, which expresses the mean value of the mAP corresponding to each threshold on the iou threshold from 0.5 in steps of 0.05 to 0.95, for a more comprehensive and precise assessment of model performance, which is calculated as shown in Eq. (18):

$$mAP0.50:0.95 = \sum_{i=0}^{9} mAP[0.5 + 0.05 * i] \tag{18}$$

4.3. Experimental parameter settings

All experiments were implemented on two 2080ti GPUs, with the SGD optimizer used in the methodology, the batch-size is set to 64. The iou threshold is set to 0.5 during the training and validation phases. 200 epochs were trained on the VOC dataset, and 300 epochs were trained on the KITTI dataset. There are two reasons for using these parameters in the experiments, one is to take into account the computational efficiency and the other is to facilitate the comparison with the previous detection algorithms, which used such settings in the previous methods.

Table 1
Performance comparison of Yolov5-AKD with other best detection algorithms on VOC dataset.

Method	Distillation	mAP (%)
Teacher	R101-FPN	82.13
Student	R50-FPN	80.53
Teacher	Yolov5l-GAMF	91.00
Student	Yolov5s-GAMF	86.30
DeFeat (Guo et al., 2021)	R101-R50-FPN	82.28
FGFI (Wang et al., 2019)	R101-R50-FPN	82.10
FBKD (Zhang & Ma, 2020)	R101-R50-FPN	82.00
ICD (Li et al., 2022)	R101-R50-FPN	82.40
GID (Dai, Jiang et al., 2021)	R101-R50-FPN	82.60
ODGK (Tang et al., 2022)	R101-R50-FPN	82.90
FitNet (Romero et al., 2014)	R101-R50-FPN	82.20
Yolov8s	–	84.70
Yolov5s	–	85.90
Ours	Yolov5l-s-GAMF	87.20(1.3↑)

Table 2
Performance comparison of Yolov5-AKD with other best detection algorithms on KITTI dataset.

Method	Distillation	mAP (%)
Teacher	Yolov5l-GAMF	93.30
Student	Yolov5s-GAMF	90.60
Teacher	R101-FPN	86.87
Student	R50-FPN	84.40
DeFeat (Guo et al., 2021)	R101-R50-FPN	85.08
FGFI (Wang et al., 2019)	R101-R50-FPN	87.91
SSD (Girshick et al., 2014)	–	81.48
Yolov3	–	85.68
Yolov8s	–	90.70
Yolov8m	–	92.10
Yolov5s	–	89.50
Ours	Yolov5l-s-GAMF	91.30(1.8↑)

Table 3
Study of the effect of α on the performance of Yolov5-AKD on VOC dataset.

α	Precision (%)	Recall (%)	mAP (%)	AP0.5:0.95 (%)
0	80.80	81.30	85.90	61.40
0.8	74.20	74.40	78.20	50.70
0.5	80.80	80.10	85.40	60.30
0.2	83.80	79.40	86.30	61.50
0.1	83.30	79.60	86.20	61.30
0.3	83.00	80.60	86.30	61.50
1	81.90	80.10	85.70	60.80

4.4. Comparison with others methods

Yolov5s was chosen as the benchmark in the experiment, and Yolov5-AKD was compared with the popular uni-stage detection network and bi-stage distillation detection network, and the bi-stage detection network was selected to be the FasterRCNN network, Tables 1 and 2 represent the comparison results on VOC and KITTI datasets, respectively, and it can be found that the detection performance of Yolov5-AKD has a significant advantage. - indicates that no distillation operation is performed, R101 indicates that the backbone of the teacher network is resnet101, R50 indicates that the backbone of the student network is resnet50, and FPN indicates that the model uses the FPN structure for multi-scale feature fusion.

5. Ablation experiments

5.1. Coefficient analysis of attentional feature fusion

Extensive experiments investigated the effect of the coefficient α in Eq. (6) on the experimental performance, the experimental results are shown in Tables 3 and 4.

Table 4
Study of the effect of α on the performance of Yolov5-AKD on KITTI dataset.

α	Precision (%)	Recall (%)	mAP (%)	AP0.5:0.95 (%)
0	90.50	83.40	89.50	61.00
0.3	95.00	83.00	90.20	63.20
0.8	94.50	83.30	90.50	63.40
0.9	90.90	85.60	90.60	63.90
1	94.20	83.90	90.30	63.20

Table 5
Performance gains obtained for different modules on VOC dataset, ✓ indicates that the corresponding module has been added to the method.

Yolov5s	GAMF	MSE+Logit	SSIM+Logit	RFT	mAP (%)
✓					85.90
✓	✓				86.30
✓	✓	✓			86.60
✓	✓		✓		86.80
✓	✓		✓	✓	87.20

Table 6
Performance gains obtained for different modules on KITTI dataset, ✓ indicates that the corresponding module has been added to the method.

Yolov5s	GAMF	MSE+Logit	SSIM+Logit	RFT	mAP (%)
✓					89.50
✓	✓				90.60
✓	✓	✓			90.70
✓	✓		✓		90.90
✓	✓		✓	✓	91.30

Table 7
Results of GAMF compared to other attention mechanisms on VOC dataset, – indicates that the attention module is not added.

Method	Attention module	mAP (%)
Yolov5s	–	85.90
	CBAM (Woo et al., 2018)	86.20
	GAM (Liu et al., 2021)	85.70
	SEA (Hu et al., 2018)	86.00
	SKA (Li, Wang, Hu et al., 2019)	86.20
	ECA (Wang et al., 2020)	86.20
	GAMF	86.30

Table 8
Results of GAMF compared to other attention mechanisms on KITTI dataset, – indicates that the attention module is not added.

Method	Attention module	mAP (%)
Yolov5s	–	89.50
	CBAM (Woo et al., 2018)	90.40
	GAM (Liu et al., 2021)	90.30
	SEA (Hu et al., 2018)	90.60
	SKA (Li, Wang, Hu et al., 2019)	90.40
	ECA (Wang et al., 2020)	90.20
	GAMF	90.60

Fig. 5. It represents the characteristic heat map of Yolov5s.

Fig. 6. It represents the characteristic heat map of Yolov5s after adding GAMAttention.

The experimental data in Tables 3 and 4 show that the attention fusion mechanism is able to help the model better capture the object features and improve the prediction accuracy and robustness of the detection model. The experimental data show that the best results are achieved with $\alpha = 0.3$ on VOC and $\alpha = 0.9$ on KITTI. We observed a strange phenomenon in the experiments: adding attention leads to a decrease in the performance of VOC, which also indicates that attention processing will lead to the local information loss in the feature maps, eventually affecting the detection effect.

5.2. Effect of different modules on detection performance

Tables 5 and 6 summarize the performance of different modules on VOC and KITTI, experimental results show that each module is compatible and combining them works best. In addition, the data in the tables demonstrate that the similarity error between feature maps outperforms MSE when using SSIM for feature refinement, which also implies that individual pixel points in an image are not isolated. Yolov5-AKD ultimately achieves a 1.3% and 1.8% improvement in performance on VOC and KITTI, respectively.

5.3. Comparing the GAMF module with other attention mechanisms

Adding an attention module to extract key features is one of the most commonly used and simplest methods in detection tasks. Through numerous experiments, it can be found that GAMAttention loses some features when extracting key features, which can be demonstrated in Fig. 6, thus affecting the improvement of the detection effect. Aiming at this drawback, this paper proposes an attention feature fusion mechanism to compensate for the lost features. GAMFAttention was optimized by comparing it with other attention mechanisms. Tables 7 and 8 present the results of GAMFAttention compared with other attention methods on VOC and KITTI, respectively.

5.4. Comparison of heat map results

We visualize the feature heat maps and the target prediction scores, and it is clear that the model learns more and more features after a series of improvements, and the model's prediction accuracy is further improved. Fig. 5 represents the feature heat maps of the original model Yolov5s. Fig. 6 represents the feature heat maps with the addition of the GAMAttention module. Fig. 7 represents the feature heat maps with the addition of the GAMFAttetion module, it can be seen that the model focuses on both global and local features. Fig. 8 represents the feature heat maps with the addition of GAMFAttention and the distillation operation. Comparing Fig. 8 with Fig. 5, it can be noticed that the performance of the model is substantially improved after adding the improved attention mechanism and performing knowledge distillation.

5.5. Confusion matrix comparison

From Figs. 9 and 10, it can be noticed that after the improvement of Yolov5s by attention and knowledge distillation, the performance of

Fig. 7. It represents the characteristic heat map of Yolov5s after adding GAMFAttention.

Fig. 8. It represents the characteristic heat map of Yolov5-AKD.

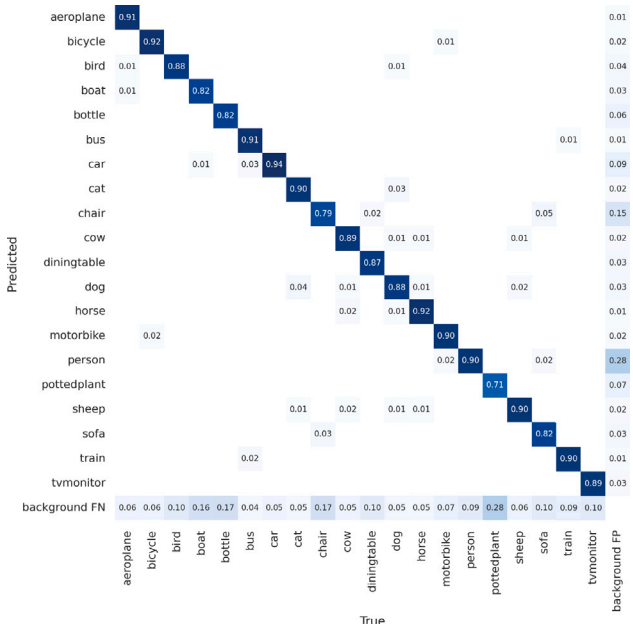


Fig. 9. It is the confusion matrix of Yolov5-AKD for VOC.

the model has achieved further improvement. After comparing the two confusion matrix plots, it can be clearly observed that the prediction precision of most of the classes has been increased, and the prediction of correct classes as incorrect and background classes has been significantly improved.

5.6. Comparison of other indicators

Fig. 11 compares Yolov5s and Yolov5-AKD in terms of Precision, Recall, and box_loss. It is evident that Yolov5-AKD demonstrates remarkable progress in these metrics. By improving the attention mechanism and implementing knowledge distillation, the model achieves a

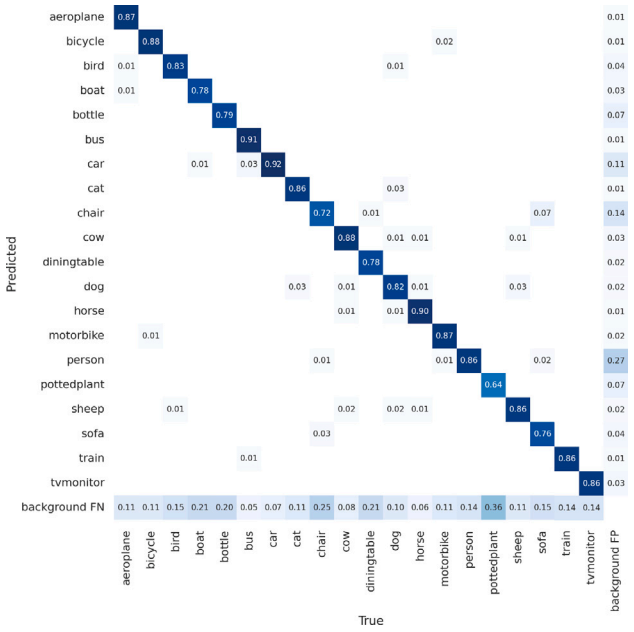


Fig. 10. It is the confusion matrix of Yolov5s for VOC.

smaller box_loss value while further enhancing Precision and Recall. In comparison to Yolov5s, Yolov5-AKD incurs only a slight computational overhead. We analyze that the GAMFAttention model not only focuses on local features but also strengthens the understanding of global features. Furthermore, through intermediate feature distillation using RFT and SSIM, the student model obtains a semantically richer feature map. Lastly, logical distillation effectively utilizes feature similarities between categories, contributing to the significant performance gains of Yolov5-AKD.

6. Conclusion

In this paper, we use an improved attention mechanism to capture the global features present in the image and use feature fusion to make up for the information loss due to attention processing features. The shortcomings of previous knowledge distillation are analyzed. An efficient knowledge transfer RFT module is proposed to transfer the main features using the relationship between response features and GT frames. For the relationship between instances not attended by RFT, SSIM is engaged to guide the student network to generate feature maps containing more instance relationships, assigning weight coefficients to these two parts to avoid extreme cases. Finally, logical distillation was introduced for the model. Yolov5-AKD updated the VOC and KITTI detection results, and numerous trials have also shown that Yolov5-AKD is effective.

Knowledge distillation can further optimize the existing detection algorithms in the coming years. The current application scenario of uni-stage detection algorithms is broader than that of bi-stage detection algorithms, it has relatively low deployment conditions, and future research on knowledge distillation can focus on uni-stage detection algorithms. Due to the internal complexity of neural networks, manually defined features may not be optimal, and there may be differences in understanding features by different people. The experiments in this paper show that transferring the knowledge of high response features can avoid the above problems, and this part of the knowledge is fixed for the model. Future research on goal detection distillation needs to consider the image as a whole and further mine the response knowledge of the model, which is the closest to the model cognition. Knowledge distillation, one of the most commonly used model compression

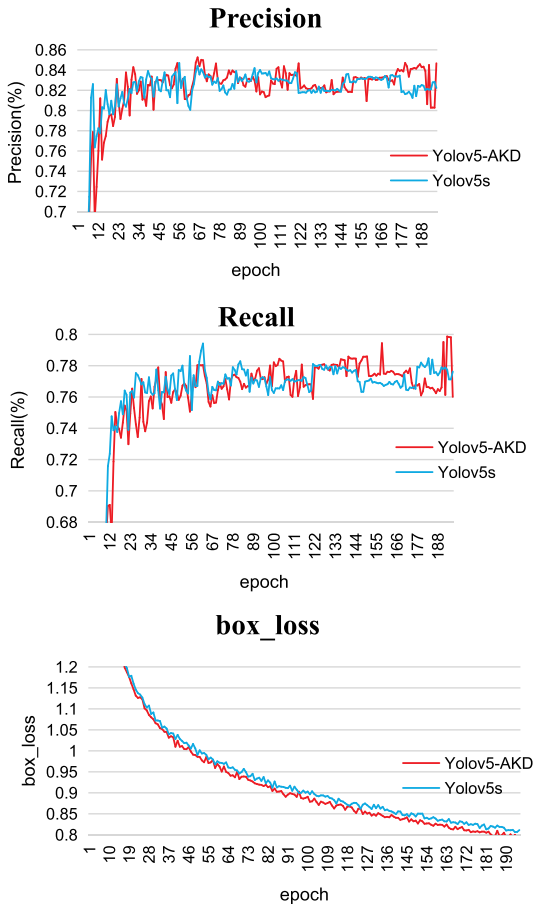


Fig. 11. It compares the Yolov5s with the Yolov5-AKD using the three evaluation metrics of box_loss, Precision, and Recall.

methods, does not directly reduce the number of participants in a small model, and relevant experiments have shown that the number of neurons contributing to the detection is 5%–10% of the total number of neurons, with significant neuron redundancy. Knowledge distillation and model pruning can be combined in future research to achieve improved model performance and reduced detection model size, upgrade the model's real-time performance, and ease the pressure of model deployment and implementation.

CRediT authorship contribution statement

ShengjieCheng: Conceptualization, Methodology, Writing – original draft, Software. **Peiyong Zhou:** Supervision. **YuLiu:** Data curation, Validation. **HongjiMa:** Investigation, Validation. **Alimjan Aysa:** Project administration, Supervision, Writing – review & editing. **Kurban Ubul:** Project administration, Supervision, Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research was subsidized by National Key Research and Development Program of China (No. 2021YFB2802100), National Science Foundation of China under Grant (62266044, 62061045, 61862061) and 2022 Department Linkage project of Xinjiang Uyghur Autonomous Region, China (2022B03035), We hereby express our deepest gratitude to the above-mentioned organizations. Thanks are also due to colleagues in the laboratory and other members of the research team whose cooperation and joint efforts made this study possible. We gratefully acknowledge the time and effort that the reviewers have invested in this study, and your professional reviews have provided the accuracy and credibility that have made our paper better.

References

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). Surf: Speeded up robust features. In *Computer vision–ECCV 2006: 9th European conference on computer vision, Graz, Austria, May 7-13, 2006. Proceedings, part I 9* (pp. 404–417). Springer.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213–229). Springer.

Chen, Y., Chen, P., Liu, S., Wang, L., & Jia, J. (2021). Deep structured instance graph for distilling object detectors. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4359–4368).

Chen, G., Choi, W., Yu, X., Han, T., & Chandraker, M. (2017). Learning efficient object detection models with knowledge distillation. *Advances in Neural Information Processing Systems*, 30.

Dai, X., Jiang, Z., Wu, Z., Bao, Y., Wang, Z., Liu, S., et al. (2021). General instance distillation for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7842–7851).

Dai, D., Zhuang, Z., Wei, J., Xia, S., Li, Y., & Zhu, H. (2021). Random sharing parameters in the global region of convolutional neural network. *IEEE Transactions on Artificial Intelligence*, *3*, 738–748.

De Rijk, P., Schneider, L., Cordts, M., & Gavrila, D. (2022). Structural knowledge distillation for object detection. *Advances in Neural Information Processing Systems*, *35*, 3858–3870.

Ghosh, S., Srinivasa, S. K., Amon, P., Hutter, A., & Kaup, A. (2019). Deep network pruning for object detection. In *2019 IEEE international conference on image processing (ICIP)* (pp. 3915–3919). IEEE.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580–587).

Guo, J., Han, K., Wang, Y., Wu, H., Chen, X., Xu, C., et al. (2021). Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2154–2164).

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961–2969).

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. arXiv preprint arXiv:1602.07360.

Li, G., Li, X., Wang, Y., Zhang, S., Wu, Y., & Liang, D. (2022). 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).

Li, X., Wang, W., Hu, X., & Yang, J. (2019). Selective kernel networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 510–519).

Li, R., Wang, Y., Liang, F., Qin, H., Yan, J., & Fan, R. (2019). Fully quantized network for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2810–2819).

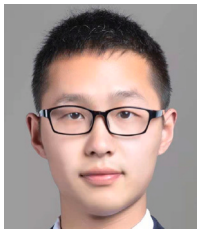
Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988).

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., et al. (2016). Ssd: Single shot multibox detector. In *Computer vision–ECCV 2016: 14th European conference, Amsterdam, the Netherlands, October 11–14, 2016, proceedings, part I 14* (pp. 21–37). Springer.

Liu, Y., Shao, Z., & Hoffmann, N. (2021). Global attention mechanism: Retain information to enhance channel-spatial interactions. arXiv preprint arXiv:2112.05561.

Mehta, S., Rastegari, M., Caspi, A., Shapiro, L., & Hajishirzi, H. (2018). Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 552–568).

Pang, Y., Yuan, Y., Li, X., & Pan, J. (2011). Efficient hog human detection. *Signal Processing*, *91*, 773–781.
Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788).
Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems, Vol. 28*.
Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2014). Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550.
Sun, R., Tang, F., Zhang, X., Xiong, H., & Tian, Q. (2020). Distilling object detectors with task adaptive regularization. arXiv preprint arXiv:2006.13108.
Tang, Y., Wang, Y., Xu, Y., Deng, Y., Xu, C., Tao, D., et al. (2021). Manifold regularized dynamic network pruning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5018–5028).
Tang, S., Zhang, Z., Cheng, Z., Lu, J., Xu, Y., Niu, Y., et al. (2022). Distilling object detectors with global knowledge. In *European conference on computer vision* (pp. 422–438). Springer.
Uijlings, J. R., Van De Sande, K. E., Gevers, T., & Smeulders, A. W. (2013). Selective search for object recognition. *International Journal of Computer Vision*, *104*, 154–171.
Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*, 600–612.
Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7794–7803).
Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11534–11542).
Wang, T., Yuan, L., Zhang, X., & Feng, J. (2019). Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4933–4942).
Wei, Y., Pan, X., Qin, H., Ouyang, W., & Yan, J. (2018). Quantization mimic: Towards very tiny cnn for object detection. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 267–283).
Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3–19).
Xie, Z., Zhu, L., Zhao, L., Tao, B., Liu, L., & Tao, W. (2020). Localization-aware channel pruning for object detection. *Neurocomputing*, *403*, 400–408.
Yang, Z., Li, Z., Jiang, X., Gong, Y., Yuan, Z., Zhao, D., et al. (2022). Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4643–4652).
Yang, Z., Li, Z., Shao, M., Shi, D., Yuan, Z., & Yuan, C. (2022). Masked generative distillation. In *European conference on computer vision* (pp. 53–69). Springer.
Yang, Z., Wang, Y., Han, K., Xu, C., Xu, C., Tao, D., et al. (2020). Searching for low-bit weights in quantized neural networks. *Advances in Neural Information Processing Systems*, *33*, 4091–4102.
Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.
Zhang, Y., Lan, Z., Dai, Y., Zeng, F., Bai, Y., Chang, J., et al. (2020). 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.
Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., et al. (2022). Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605.
Zhang, L., & Ma, K. (2020). Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International conference on learning representations*.
Zhang, D., Wang, H., Figueiredo, M., & Balzano, L. (2018). Learning to share: Simultaneous parameter tying and sparsification in deep learning. In *International conference on learning representations*.
Zhixing, D., Zhang, R., Chang, M., Liu, S., Chen, T., Chen, Y., et al. (2021). Distilling object detectors with feature richness. *Advances in Neural Information Processing Systems*, *34*, 5213–5224.



ShengjieCheng 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 Computer Science and Technology, Xinjiang University, China. His main research interests include object detection, image processing and knowledge distillation.

Peiyong Zhou, master, lecturer, member of the China Computer Federation, graduated from Xinjiang University in 1997 with a bachelor's degree and Xinjiang University with a master's degree in 2000. Currently, he is engaged in machine learning and image processing, and has published more than 10 papers.

YuLiu was born in ShiYan, Hubei, China, in 1999. He obtained the bachelor's degree in HUBEI POLYTECHNIC UNIVERSITY, China, in 2021. He is currently pursuing a master's degree in Computer Science and Technology at the School of Computer Science and Technology, Xinjiang University, China. His main research interests include video object tracking.

HongjiMa was born in Jining, Shandong, China, in 1998. He obtained the bachelor's degree in Software Engineering from the College of Computer Science & Technology of Qingdao University, China, in 2021. He is currently pursuing a master's degree in Computer Technology at the School of Computer Science and Technology, Xinjiang University, China. His main research interests include object detection and single object tracking.

Alimjan Aysa is currently an Professor with the School of Computer Science and Technology, Xinjiang University. He has also been as a Researcher with the Xinjiang Laboratory of Multi-language Information Technology since 2002. His research interests include natural language processing, pattern recognition, and digital signal processing. He has authored 2 books and over 60 papers. His work has appeared in many Chinese core journals such as the Computer Engineering and Application, the Chinese Information Processing, and Computer Application. He is a member of the Image and Graphics Society of China, and the China Computer Federation.

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.