Pattern Recognition 149 (2024) 110235

Contents lists available at ScienceDirect

Pattern Recognition

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

Closed-loop unified knowledge distillation for dense object detection

Yaoye Song [a,b], Peng Zhang [a,b,*], Wei Huang [c], Yufei Zha [a,b], Tao You [a], Yanning Zhang [a]

[a] School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi, The People's Republic of China
[b] Ningbo Institute of Northwestern Polytechnical University, The People's Republic of China
[c] School of Mathematics and Computer Sciences, Nanchang University, Nanchang, The People's Republic of China

ARTICLE INFO

Keywords:
Triple parallel distillation
Hierarchical re-weighting attention distillation
Dense object detection
Closed-loop unified

ABSTRACT

Most of knowledge distillation methods for object detection are feature-based and have achieved competitive results. However, only distillating in feature imitation part does not take full advantage of more sophisticated detection head design for object detection, especially dense object detection. In this paper, a triple parallel distillation (TPD) is proposed which can efficiently transfer all the output response in detection head from teacher to student. Moreover, to overcome the drawback of simply combining the feature-based with the response-based distillation with limited effect enhancement. A hierarchical re-weighting attention distillation (HRAD) is proposed to make student learn more than the teacher in feature information, as well as reciprocal feedback between the classification-IoU joint representation of detection head and the attention-based feature. By jointly interacting the benefits of TPD and HRAD, a closed-loop unified knowledge distillation for dense object detection is proposed, which makes the feature-based and response-based distillation unified and complementary. Experiments on different benchmark datasets have shown that the proposed work is able to outperform other state-of-the-art distillation methods for dense object detection on both accuracy and robustness.

1. Introduction

Knowledge distillation for object detection has been regarded as a promising solution [1–4], to maintain the inference speed of small-sized student network as well as the accuracy of the large-sized teacher network in more recent studies. Those work mainly focused on the feature-based schemes [5,6] with the differences of distilled specific regions in the feature map. This kind of methods can relatively promise the student to distill the teacher's knowledge from the feature imitation, and can be conveniently applied to popular object detectors [7–10]. However, this kind of distillation for object detection does not fully consider the important information in the detection head, which loses the advantage of more sophisticated detection head design for object detection, especially dense object detection.

As shown in Table 1, for the dense object detectors GFL [11] GFLV2 [12], the AP value of the teacher models with ResNet101 2× training is more than 4.5 higher than that of the student models with ResNet-50 1× training. Comparatively, the gap between some non-dense object detectors [8,13,14] with ResNet-101 and ResNet-50 is only about 2.4 AP. This means that a better design of detection head would determine the final performance difference of dense object detection based on the same backbone network, and eventually obtain more benefits than non-dense target detection so far. Therefore, how to make the best use of the output response in detection head during knowledge distillation becomes crucial problem.

Following the way of response-based distillation to transfer the output response in detection head, as shown in the left part of Fig. 1, LD [1] proposes localization distillation by utilizing the general distribution of localization branch. To enhance both semantic and localization information, this paper proposes separately using LD loss and KD loss functions for the positive region (PR) and valuable localization region (VLR). However, ablation experiments reveal that incorporating KD loss leads to only marginal improvements in overall performance, and incorporating KD loss from VLR results in decreased overall performance. Furthermore, the contribution of LD in dense object detection is significantly inferior compared to feature-based KD methods. Therefore, we contend that distilling knowledge through the localization branch and incorporating a separate classification KD loss within the localization region has a negligible impact.

In this study, As shown in the right part of Fig. 1, from the perspectives of overall parallel structure and the equilibrium of each branch in the dense object detection head, we fully utilize the output response in parallel detection head respectively. Besides, we merge the PR and VLR of the localization branch as combined VLR (CVLR) to make the distillation process more sufficient and efficient. This whole process is

* Corresponding author at: School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi, The People's Republic of China.
E-mail address: zh0036ng@nwpu.edu.cn (P. Zhang).

https://doi.org/10.1016/j.patcog.2023.110235
Received 26 September 2023; Received in revised form 8 December 2023; Accepted 26 December 2023
Available online 28 December 2023

Table 1
Comparisons of different object detectors. These detectors are classified into two types: dense and non-dense. The last two rows show the effect of the detector with ResNet-50 (1×) and ResNet-101 (2×) backbone networks. 1× and 2× means the training schedule. 1×: single-scale training 12 epochs. 2×: multi-scale training 24 epochs. All results are trained on COCO 2017 train set and tested on COCO 2017 val set.

Backbone	Dense		Non-dense	
	GFL [11]	GFLV2 [12]	Faster R-CNN [8]	Cascade R-CNN [13]
ResNet-50(1×)	40.2	41.0	37.4	40.3
ResNet-101(2×)	**44.7(+4.5)**	**45.8(+4.8)**	39.8(+2.4)	42.8(+2.5)

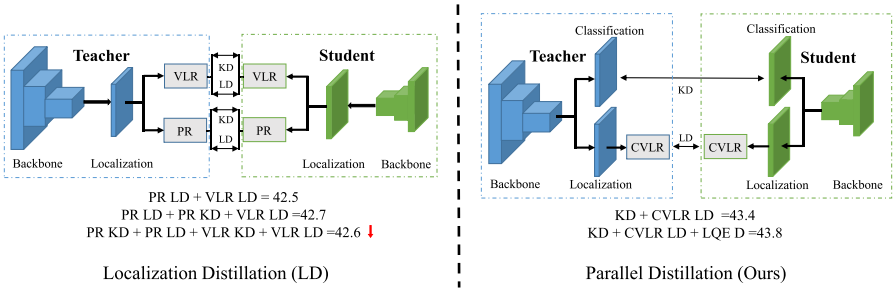


Fig. 1. Illustration of existing LD [1] and our parallel distillation. Localization: localization branch in detection head, Classification: classification branch in detection head, PR: positive region, VLR: valuable localization region, CVLR: combined valuable localization region, LD: Kullback–Leibler divergence loss for localization information KD: Kullback–Leibler divergence loss for semantic information.

commonly referred to as parallel distillation. Our method surpasses the results achieved by LD [1] while employing a reduced distillation loss function, showcasing one of our innovative contributions.

According to the above analysis, in this paper, a novel **T**riple **P**arallel **D**istillation (TPD) is proposed by adapting the detection head structure of GFLV2 [12]. Based on this competitive TPD, the response-based distillation can be jointly incorporated into the parallel detection framework with the feature-based distillation to achieve better results. However, the current mainstream feature-based knowledge distillation methods have very limited improvement in overall performance when combined with response-based TPD. Considering that the complementary effect between TPD and feature-based distillation enables the student to learn more beyond the knowledge limitation of teacher, a **H**ierarchical **R**e-weighting **A**ttention **D**istillation (HRAD) is also proposed by re-designing the attention mechanism to use the spatial, channel and relation attention collaboratively.

More importantly, the classification-IoU joint representation, which is the fusion of localization quality estimation and classification, from the teacher's detection head is feedback to each layers of attention feature. By this way, not only the feature can become more discrimination in each layer of FPN, but a reciprocal feedback can also be achieved between the Classification-IoU joint representation of detection head and the attention-based feature. By taking advantage of TPD and HRAD as closed-up unified distillation, a **C**losed-loop **U**nified Knowledge **D**istillation (CUD) is proposed for dense object detection. The main contributions in this work can be summarized as follows:

- A triple parallel distillation (TPD) is proposed to substantially exploit the output response of all branches, which is able to achieve a superior response-based distillation for dense object detection.
- A hierarchical re-weighting attention distillation (HRAD) is proposed to enable student network to learn more than the teacher network in feature imitation. With a reciprocal feedback between the classification-IoU joint representation of detection head and the attention-based feature, the detection performance has been further improved.
- A closed-loop unified distillation is proposed by jointly interacting the benefits of TPD and HRAD, which makes the feature-based and response-based distillation unified and complementary. An extensive experiments on the COCO, PASCAL VOC datasets have been conducted to evaluate the effectiveness for a state-of-the-art performance.

2. Related works

In this section, we briefly review the related works including dense object detection and knowledge distillation.

2.1. Dense object detection

To overcome the limitation of non-dense object detector [8,13,14], an increasing amount of work has been devoted to the study of dense object detection [15] more recently. It integrates classification and regression branch into a feed forward network and reduces the RPN module at the same time. This makes the overall network structure more concise with less computation, and has a great advantage in speed compared to the non-dense detectors in more widely practices. For the dense detectors with faster inference, such as SSD [16] and YOLO series [17–19], the dense proposal made them extremely difficult to distinguish the positive and negative samples during training, which limits the accuracy to be further improved. To solve this problem, RetinaNet [9] employs the focal loss, FCOS [20], CenterNet [21] and Reppoints [22] utilize the keypoints instead of setting anchors to generate more positive samples during sampling, which is known as the anchor-free detection. ATSS [23] proposes an Adaptive Training Sample Selection to automatically select positive and negative samples according to statistical characteristics of object. It also proves that the essential difference between anchor-based and anchor-free detectors is actually the definition of positive and negative training samples. On the basis of [11], an enhanced Localization Quality Estimation (LQE) named GFLV2 [12] is proposed to merge the localization quality estimation into the classification prediction. This method is also used as a baseline detection framework the proposed knowledge distillation strategy because of its simple structure of achieving high accuracy with relatively low computational overhead.

2.2. Knowledge distillation

Knowledge distillation [24,25] is a method for model compression and acceleration, and to enhance the performance of student models with the guidance of a teacher's deeper and larger models. Usually, it has two categories according to the classification from the position of distillation. One category is the response-based distillation [24], which utilizes the soft targets response of output layer. The other is

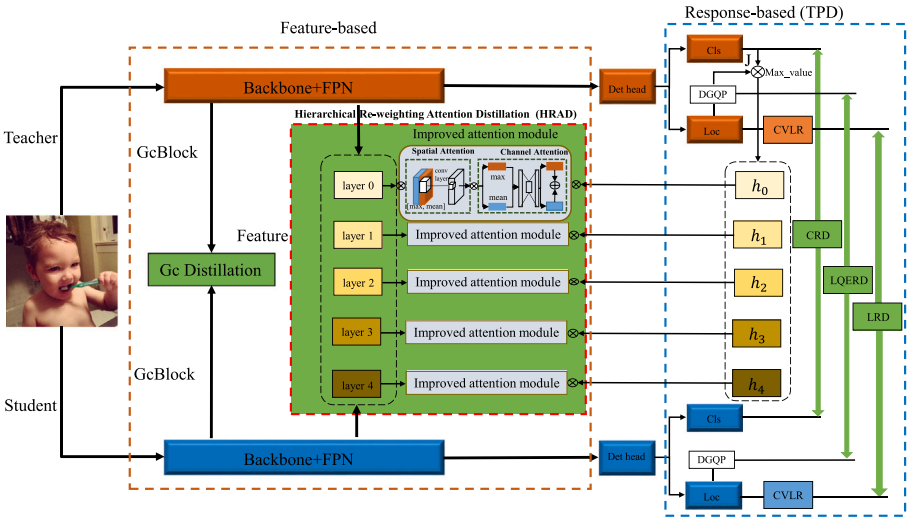


Fig. 2. Illustration of closed-loop unified knowledge distillation (CUD). The red dotted box indicates our hierarchical re-weighting attention distillation (HRAD), the brown dotted box represents the feature-based distillation, and the blue dotted box is our response-based distillation TPD. The gray rectangular box represents our proposed improved attention module, and the green rectangles represent the locations of knowledge distillation in CUD. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the feature-based distillation, which utilizes the semantic information of feature. Unlike the response-based distillation only in the output layer, feature-based distillation can be performed in the complex intermediate layer. With an increasing number of studies on the knowledge distillation, it has been widely used in object detection more recently.

2.3. Knowledge distillation for object detection

Feature-based: Unlike the classification task [26,27], the feature-based distillation losses in the detection task encounter extreme imbalances between positive and negative instances. To address this problem, Li et al. [5] attempt distillation in the region proposal network (RPN) module of Faster R-CNN, which distills a certain proportion of positive and negative instances sampled by RPN. Subsequently, Wang et al. [28] believe that selecting the entire RPN module for distillation is not conducive to the definition of positive and negative samples, and then propose the fine-grained mask to distill the regions calculated by the ground-truth bounding boxes. Guo et al. [6] distill the foreground and background separately to enable more beneficial to the student networks.

For another type of object detection using attention-based knowledge distillation [29,30], Zhang et al. [29] propose an attention-guided distillation and a non-local distillation to balance the pixels between foreground and background, and further to avoid the lack of distillation on the relation between different pixels. Yang et al. [30] propose a Focal distillation and a and Global distillation. The former is to force the student into paying attention to the teacher's critical pixels and channels, and the later is to rebuild the relation between different pixels. Those distillation are categorized as a branch of feature-based distillation because its core operation still mainly focuses on the feature extraction.

Response-based: In the limited studies of response-based knowledge distillation for object detection, LD et al. [1] proposes a localization distillation (LD) to efficiently transfer the localization knowledge from teacher to the student. However, this method does not fully consider all the output responses of the dense object detector. When compared with the most effective feature-based methods, the result is relatively uncompetitive.

Different from using the feature-based and the response-based distillation independently as above, in this study, both of the distillation are jointly employed with the interactive feedback mechanism. For the proposed CUD, there are two challenges to be solved: (1) Response-based distillation fail to distill the useful output responses sufficiently in the detection head. (2) The performance improvement obtained by simply combination of feature-based and response-based distillation is very limited.

3. Proposed work

Different works have validated that distilling the localization response distribution can efficiently transfer the localization knowledge from teacher to student [1]. However, it has also been found that the meaningful knowledge not only exists in the output response of localization branch, but also in the classification branch and localization quality estimation process. This inspired us to design a triple parallel distillation (TPD) as shown in Fig. 2 (blue dotted box), which takes full use of all model branches from the overall structural perspective of detector head.

In order to allow student learn information more from teacher's feature information, a hierarchical re-weighting attention distillation (HRAD) is also proposed. As shown in the Fig. 2 (red dotted box), the proposed distillation is able to extract effective feature attention, as well as to re-weight it with the classification-IoU joint representation of detection head in each layer.

The proposed TPD and HRAD are integrated into an end-to-end closed-loop unified knowledge distillation (CUD) for dense object detection. Rather than a simple combination of response-based and feature-based distillation, a mutual feedback mechanism has been incorporated in the proposed CUD to achieve remarkable performance. The details of our work is introduced as below:

3.1. Triple parallel distillation for dense object detection

The most popular response-based knowledge distillation for image classification is known as soft targets [24], which is an operation to soften the distribution of output. Specifically, this operation also means that the probabilities of the input belonging to different categories and can be estimated by a Softmax function.

$$p(z_i, T) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{1}$$

where z_i is the logit for the ith class, and the temperature factor T is introduced to control the importance of each soft targets. It is worth

noting that when $T = 1$, the function is equivalent to the original Softmax function, when $T \to 0$, it tends to be a Dirac delta distribution. When $T \to \infty$, it is approximated as a uniform distribution. To make the probability distribution contain more information, T is generally set to $T > 1$ to soften the distribution. The distillation loss to measure the similarity between the probability distributions of teacher and student can written as:

$$L_{kd}(p_t, p_s) = L_{kl}((p_t, T), (p_s, T)) \tag{2}$$

where the $L_{kl}((p_t, T), (p_s, T))$ employs Kullback–Leibler divergence loss.

3.1.1. Localization response distillation

To directly mimic the final distributed prediction of localization branch, inspired by [11,12], we utilized the 'general distribution' $P(x)$ which represents the arbitrary distribution of bounding box to reflect the uncertainty of localization. In this distribution, each edge of the bounding box can be formulated as: $\hat{b} = \int_{-\infty}^{+\infty} P(x)x\mathrm{d}x = \int_{b_0}^{b_n} P(x)x\mathrm{d}x$, which is within a predefined output range of $[b_0, b_n]$. Meanwhile, the continuous range of domain is converted into a discrete list $[b_0, b_1, \dots, b_n]$, and the estimated bounding box value $\hat{b}$ can be written as:

$$\hat{b} = \sum_{i=0}^{n} P(b_i)b_i \tag{3}$$

Then, the knowledge distillation in localization branch $L_{loc}^{kd}(B_t, B_s)$ for all the four edges of bounding box is formulated as:

$$L_{loc}^{kd}(B_t, B_s) = \sum L_{kl}^{b}((p_t^b, T), (p_s^b, T)) \tag{4}$$

Unlike LD [1] which distills the localization information in the Valuable Localization Region (VLR) and positive sample region respectively. This study exploit the simplicity and performance by combining VLR with positive sample region. Experiments prove that this method can achieve better results, and we will also discuss it in detail in the experiments section.

3.1.2. Classification response distillation

Compared to distillation of the localization branch, distillation of the classification branch is simpler because the output of the classification is originally in the form of a distribution. And the expression of the loss function for classification distillation L_{cls}^{kd} is:

$$L_{cls}^{kd}(C_t, C_s) = L_{kl}^{c}((p_t^c, T), (p_s^c, T)) \tag{5}$$

3.1.3. Localization quality estimation response distillation

The gap of classification and localization branches is further bridged by distilling the reliable localization quality estimation (LQE) score **Q**, which is obtained by the distribution-guided quality predictor (DGQP) to guide the classification **C**. Considering that the LQE score is a discrete value, we adapt $L2$ loss to make the student network closer to the teacher network. The loss function for distillation of localization quality L_{lqe} is defined as:

$$L_{lqe}(Q_t, Q_s) = \|Q_t - Q_s\|_2 \tag{6}$$

where Q_t and Q_s represent the localization quality estimation score which is obtained by and the DGQP sub-network of teacher and student in GFLV2.

3.2. Triple parallel distillation loss function

The total loss for training the student model can be represented as:

$$\begin{aligned} L_{TPD} = {} & \lambda_0 L_{cls}(C_s, C^{gt}) + \lambda_1 L_{loc}^{kd}(B_s, B^{gt}) + \lambda_2 L_{dfl}(B_s) \\ & + \lambda_3 L_{cls}^{kd}(C_t, C_s) + \lambda_4 L_{loc}^{kd}(B_t, B_s) + \lambda_5 L_{lqe}(Q_t, Q_s) \end{aligned} \tag{7}$$

where the fist three loss functions are proposed in [11,12], they are the classification loss L_{cls}, bounding box regression loss L_{loc} and the distribution focal loss L_{dfl} of student detector, respectively. The L_{cls}^{kd}, L_{loc}^{kd}, L_{lqe} represents the distillation loss in the TPD.

3.3. Hierarchical re-weighting attention distillation for dense object detection

3.3.1. Hierarchical re-weighting attention distillation

Usually, the attention-based distillation tends to extract simple mean of spatial and channel attentions to enhance detection performance. However, it is hard to obtain better results from the direct distillation with more advanced attention mechanisms, such as SENet [31] and CBAM [32]. In our work, an improved attention mechanism based on CBAM is improved with the re-designed modules for the characteristics of knowledge distillation. Specifically, since feature-based distillation generally sets the temperature T to be less than 1, this is to make the output distribution of features steeper. To limit the value domain within $[0, 1]$ for outputs, the CBAM employs the sigmoid function after extracting the average-pooled and max-pooled features simultaneously, which obviously contradicts with the setting of distillation temperature. By discarding the sigmoid function, we define the spatial attention map M^S and channel attention map M^C as:

$$M^S(F) = \frac{1}{C}\sum_{c=1}^{C}(f^{7\times7}(AvgPool(F_c); MaxPool(F_c))) \tag{8}$$

$$\begin{aligned} M^C(F) = {} & \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(MLP(AvgPool(F_{i,j})) \\ & + MLP(MaxPool(F_{i,j}))) \end{aligned} \tag{9}$$

where H, W, C denotes the feature's height, width, and channel. And the spatial attention mask A^S and channel attention mask A^C can be formulated as:

$$A^S(F) = H \cdot W \cdot softmax(M^S(F)/T) \tag{10}$$

$$A^C(F) = H \cdot W \cdot softmax(M^C(F)/T) \tag{11}$$

where T is the temperature hyper-parameter to adjust the output distribution.

After obtaining a valid attentional feature map, a hierarchical re-weighting factor H, which is obtained from the output of teacher's detection head, is introduced to allow the **R**esponse feedback the **F**eature in each FPN layer. Not only does this enable the distillation to incorporate the localization quality estimation in detection head, but also to make each layer of FPN more discriminative from each other. As shown in Fig. 2 (green dotted box), the maximum output value of the classification-IoU joint representation is then utilized to re-weight the feature in each FPN layer:

$$H_l = \max_{1\le c'\le C} y_{c'l}^T \tag{12}$$

where l represents the lth FPN layer, $c' \in [1, C]$ means the output category of classification-IoU joint representation, C is the number of categories in the dataset, and y^t is the probability of teacher, the proposed mixed attention feature loss L_{maf} is defined as follows:

$$L_{maf} = \sum_{l=1}^{M}\frac{1}{N_l}\sum_{i=1}^{H}\sum_{j=1}^{W}H_{lij}\sum_{k=1}^{C}(F_{k,i,j}^T - f_{adapt}(F_{k,i,j}^S))^2 A_{i,j}^S A_k^C \tag{13}$$

where l represents the lth FPN layer, M is the number of FPN layers, i, j, k represent the width W, height H and channel C of feature map, respectively. $F_{k,i,j}^T$ and $F_{k,i,j}^S$ denote the feature maps of teacher detector and student detector, and A^S and A^C denote the improved spatial and channel attention mask of the teacher detector.

To mimic the spatial and channel attention mask of teacher detector, the improved channel and spatial attention feature losses are also proposed as:

$$L_{caf} = \sum_{l=1}^{M}\frac{1}{N_l}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{C}(F_{k,i,j}^T - f_{adapt}(F_{k,i,j}^S))^2 A_k^C \tag{14}$$

$$L_{saf} = \sum_{l=1}^{M} \frac{1}{N_l} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{k=1}^{C} (F^T_{k,i,j} - f_{adapt}(F^S_{k,i,j}))^2 A^S_{i,j} \quad (15)$$

The hierarchical re-weighting attention loss L_{HRAD} is the sum of L_{maf}, L_{caf} and L_{saf}:

$$L_{HRAD} = \alpha L_{maf} + \beta L_{caf} + \gamma L_{saf} \quad (16)$$

3.3.2. Gc distillation

To capture the relation between pixels in an image, GcBlock [33] is used with the Gc loss L_{Gc} is as:

$$L_{Gc} = \delta \sum (R(F^T) - R(F^S))^2 \quad (17)$$

The losses above are then combined as feature loss L_{Feat}:

$$L_{Feat} = L_{HRAD} + L_{GC} \quad (18)$$

3.4. Closed-loop unified distillation for dense object detection

The feature-based HRAD can transfer more feature information to the detection head and benefit the response-based TPD. Correspondingly, the hierarchical re-weighting factor output from the detection head can be fed back to the HRAD. This process is similar to the closed-loop control system in the study of control science, which inspired the name of this method as the Closed-loop Unified Knowledge Distillation for dense object detection, its loss L_{CUD}:

$$L_{CUD} = L_{TPD} + L_{Feat} \quad (19)$$

4. Experiments

In this section, the implementation details of the proposed method is presented. The detection performance is also validated on the challenging MS COCO 2017 benchmark [34], and the comparison to the other state-of-the-art detectors is also reported.

4.1. Implementation details

The proposed work is implemented based on MMDetection framework [35] by referring its configuration. In the framework, ResNet [36] with FPN [37] is employed as the backbone network for overall evaluations, unless otherwise specified. For COCO dataset, we perform synchronized stochastic gradient descent(SGD) [38] by 16 images per minibatch (2 images per GPU) with 12 epochs. In ablations, all the experiments follow the 1× settings: the initial learning rate of 0.01, a weight decay of 0.0001 and momentum of 0.9. The learning rate is decreased by a factor of 10 after 8 epochs and 11 epochs iterations, respectively. For Pascal VOC dataset, the maximum training epoch is set to 4 and the learning rate is decreased by a factor of 0.1 after 3 epochs, when other settings are the same as the COCO dataset. The backbone network is initialized with the weights pre-trained on ImageNet [39], and the input images are resized to ensure their shorter edge to be 800 and the longer edge less than 1333. For other training and testing hyper-parameters, the protocol of GFLV2 [12] is followed exactly.

During the inference process, a single-scale testing approach is adopted with the same image size as in the single-scale training. For the operation of Non-Maximum Suppression (NMS) [40], an IoU threshold of 0.6 is applied to remove duplicate boxes.

4.2. Datasets

4.2.1. COCO [34] dataset

COCO Dataset is regarded as one of the most popular and authoritative benchmarks for the evaluation of different vision tasks. Our ablation experiments are trained on COCO2017 training set (115K images) and evaluated on COCO minival set (5K images). To compare with the other state-of-art detectors, the COCO AP is reported on the test-dev set (20K images).

Table 2
The comparison with Different Distillation Methods. The same teacher detector with ResNet-101-FPN backbone distills ResNet-50-FPN. 1× and 2× means the training schedule. 1×: single-scale training 12 epochs. 2×: multi-scale training 24 epochs.

Category	Method	AP	AP_{50}	AP_{75}
Baseline(GFLV2 [12])	S:ResNet-50 (1×)	41.0	58.5	45.0
	T:ResNet-101 (2×)	45.8	63.8	50.1
Feature-based	FitNet [41]	41.5	59.3	45.2
	DeFeat [6]	41.7	59.6	45.4
	Fine-grained [28]	41.9	60.0	45.5
	GID [42]	42.3	60.5	46.0
Attention-based	FKD [29]	42.7	60.3	46.5
	PGD [43]	44.1	61.7	48.2
	FGD [30]	44.1	61.8	48.1
Response-based	LD [1]	42.7	60.2	46.7
	Dist [44]	42.6	61.0	46.5
Our response-based	**TPD**	43.8	61.3	47.9
Our combined	**CUD**	**44.5**	**62.1**	**48.4**

4.2.2. PASCAL VOC [45] dataset

The PASCAL VOC Challenge was conducted between 2005 and 2012. One of the benchmarks for object detection. There are 20 classifications in this dataset. And PASCAL VOC dataset contains two parts, VOC2007 and VOC2012. VOC2007 contains 9963 labeled images, with a total of 24640 objects labeled. The VOC2012 dataset is an upgraded version of the VOC2007 dataset, with a total of 11530 images containing 27450 objects. For the evaluation we use the VOC2012 test set, which contains 2913 images with a total of 6929 objects.

4.3. Evaluation metrics

4.3.1. AP

6 evaluation metrics are employed in the experiment for a comprehensive evaluation of the detection performance including AP, AP_{50}, AP_{75}, AP_s, AP_m, AP_l. Specifically for COCO dataset, the evaluation of AP is regarded as the most important metric. Usually, AP (average precision) is averaged over multiple Intersection over Union (IoU) values with 10 IoU thresholds of .50:.05:.95. AP_{50} is calculated at a single IoU of 0.5 and AP_{75} is at 0.75. AP_s represents the objects size less than 32^2, AP_m represents the size between 32^2 and 96^2 and AP_l represents the size bigger than 96^2.

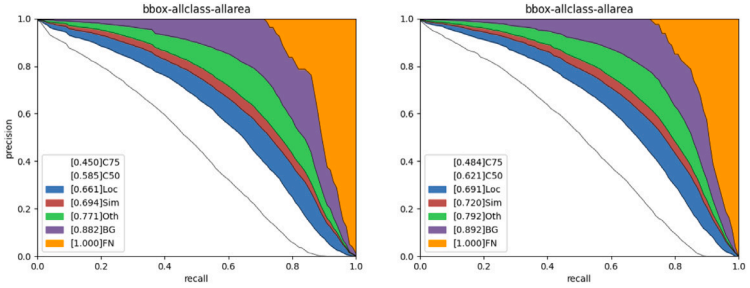
4.3.2. TIDE

TIDE [46] is a general toolbox for identifying object detection errors, which is a complement to AP as an evaluation metric. The main errors of TIDE contains 6 categories, "CLS" represents localized correctly but classified incorrectly, "Loc" represents classified correctly but localized incorrectly, "Both" classified incorrectly and localized incorrectly, "Dupe" represents the object would be correct if not for a higher scoring detection, "Bkg" represents detected background as foreground, "Miss" represents all undetected ground truth (false negatives) not already covered by classification or localization error. And the special errors of TIDE contain 2 categories, "FalsePos" represents false positive, "FalseNeg" represents False Negative.

4.4. Main results

The proposed CUD is compared with the other state-of-the-art distillation methods [1,6,28–30,41–44], under the same teacher–student (ResNet101-ResNet50) network pair in dense object detector GLFV2 [12]. And the results of the teacher–student detector have been shown in Table 1.

As shown in Table 2, the proposed TPD is a response-based method, which is designed from the perspective of the overall efficiency of the distillation process. Owing to fully exploiting the distribution of



GFLV2 CUD

Fig. 3. Visual comparisons on COCO evaluation of GFLV2 and our CUD, The teacher is ResNet-101 and the student is ResNet-50.

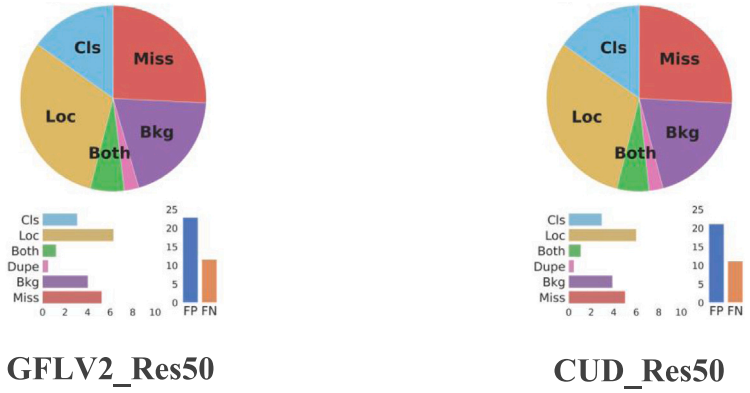


Fig. 4. Visual comparisons on TIDE evaluation of GFLV2 and our method CUD, The teacher backbone network is ResNet-101 and the student backbone network is ResNet-50.

Table 3
The main errors of GFLV2 and our method on the evaluation of TIDE.

Main errors						
Method	Cls ↓	Loc ↓	Both ↓	Dupe ↓	Bkg ↓	Miss↓
GFLV2 [12]	3.09	6.30	1.21	0.51	4.04	5.26
LD [1]	3.23	6.00	1.20	0.49	4.17	5.06
FKD [29]	3.09	6.03	1.16	0.49	3.99	5.16
FGD [30]	3.01	6.02	1.10	0.48	3.91	5.05
CUD	**2.95**	**6.01**	**1.09**	**0.47**	**3.89**	**5.02**

Table 4
The special errors of GFLV2 and our method on the evaluation of TIDE.

Special errors		
Method	FalsePos ↓	FalseNeg ↓
GFLV2 [12]	22.83	11.57
LD [1]	21.99	11.24
FKD [29]	22.08	11.11
FGD [30]	21.31	11.12
CUD	**21.09**	**11.09**

all branches in the teacher network, TPD is able to achieve a 2.8 AP improvement, and significantly surpass all the other response-based methods [1,44]. When further combining TPD and HRAD into the joint CUD, a 3.5 AP improvement can be obtained. Fig. 3 presents the visualization of experiment results on the evaluation on COCO val2017.

4.5. Evaluation on TIDE

We also make a comparison of GFLV2 [12], LD [1], FKD [29], FGD [30] and our CUD on the evaluation of TIDE. Tables 3 and 4 shows that our method get a better performance in all evaluation metrics of TIDE [46]. Furthermore, as shown in Fig. 4, the proportion of each error category in the pie chart is basically the same for both methods, indicating that our method is an overall improvement in all aspects for the baseline model, which has confirmed the excellent effect of our distillation method.

4.6. Evaluation on different detectors

To demonstrate that our proposed CUD can work in different types of object detectors, we apply the CUD to the two-stage detector Faster R-CNN [8], the one-stage detector CenterNet [21], and Yolo V3 [19]. As shown in Table 5, all the results show that our proposed CUD can effectively distill different types of detectors and is much better than LD [1].

4.7. Ablation study of TPD

4.7.1. Comparison of different distillation components in TPD

To evaluate the effectiveness of each component of the proposed distillation, we choose the ResNet-50-FPN as the student model and ResNet-101-FPN as the teacher model. The ablation study is reported in Table 6. It can be found that conducting "Classification Response Distillation (Cls)", "Localization Response Distillation (Loc)" and "Localization Quality Estimation Response Distillation (LQE)" can enhance the student performance by +1.3, +1.4, +0.8 AP, respectively. Particularly, the combination of Cls and Loc can obtain 43.3 AP, while the combination of three distillation can obtain 43.8 AP by complementary effect.

4.7.2. Comparison of different lightweight detectors

To further validate with different lightweight detectors, ResNet-101 with 45.8 AP is employed as the teacher model to distill a series of lightweight students, including ResNet-18, ResNet-34, and ResNet-50. As shown in Table 7, for the three given students, the AP can be stably improved by +2.9, +2.1, +3.1 without adding bells and whistles. Based on the results, the proposed method has shown a steadily improvement of various types of lightweight detectors.

Table 5
Our proposed CUD can be applied to different detectors, such as Faster R-CNN, CenterNet and Yolo V3. The results are reported on MS COCO val2017.

Detector	Backbone	Method	AP	AP_{50}	AP_{75}
Faster R-CNN [8]	ResNet-101(T)		39.8	60.1	43.3
	ResNet-50(S)		37.4	58.1	40.4
		LD [1]	38.6 (+1.2)	59.1	42.1
		CUD	**39.6 (+2.4)**	**60.0**	**43.1**
CenterNet [21]	ResNet-101(T)		34.6	53.0	36.9
	ResNet-18(S)		28.1	44.9	29.6
		LD [1]	29.6 (+1.5)	46.8	31.3
		CUD	**31.1 (+3.0)**	**48.0**	**32.9**
Yolo V3 [19]	DarkNet-53(T)		30.9	52.9	32.1
	MobileNetV2(S) [47]		22.2	41.9	21.4
		LD [1]	23.9 (+1.7)	43.3	23.0
		CUD	**25.1 (+2.9)**	**44.8**	**24.1**

Table 6
The teacher is ResNet-101 and the student is ResNet-50. The second row indicates the baseline GFLV2, The third to fifth rows are our "Classification Response Distillation (Cls)", "Localization Response Distillation (Loc)", and "Localization Quality Estimation Response Distillation (LQE)", respectively. The last two rows are the AP of parallel distillation(Cls& Loc) and TPD.

Cls	Loc	LQE	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
			41.0	58.5	45.0	24.3	44.6	53.4
√			42.3	60.8	46.2	24.7	46.2	55.5
	√		42.4	59.3	46.2	24.4	46.1	54.6
		√	41.8	59.3	45.7	24.1	45.2	54.5
√	√		43.4	60.8	47.5	25.3	47.3	56.6
√	√	√	**43.8**	**61.3**	**47.9**	**25.9**	**48.1**	**56.7**

Table 7
The same teacher detector with ResNet-101-FPN backbone distills different lightweight detectors with ResNet-18, ResNet-34, and ResNet-50 respectively. The results are reported on MS COCO val2017.

Teacher	Student	TPD	AP	AP_{50}	AP_{75}
ResNet-101	ResNet-18		36.3	52.8	39.5
		√	**39.2(+2.9)**	**56.0**	**42.8**
	ResNet-34		39.5	56.5	43.1
		√	**41.6(+2.1)**	**58.5**	**45.4**
	ResNet-50		41.0	58.5	45.0
		√	**43.8(+2.8)**	**61.3**	**47.9**

4.7.3. Different teacher distill same student

In the preceding section, we demonstrated the effectiveness of TPD in yielding favorable results within networks of the same category such as ResNet. In this section, we extend our investigation to assess TPD's performance in heterogeneous networks. As shown in Table 8, we employ TPD to distill knowledge from the ResNet-50 student network with various teacher networks, including ResNeXt-101, Res2Net-101, and Swin Transformer-Tiny. Remarkably, we observe enhancements of 2.8, 3.0, and 3.1, respectively. This compelling evidence showcases the versatility of our approach in delivering superior results across diverse backbone network architectures.

4.7.4. Extension to other dense object detectors

To verify the flexibility of incorporation with different dense object detectors, the proposed TPD has been incorporated into the other detectors, such as ATSS [23] and GFL [11], to have a consistent 3 AP improvement as shown in Table 9.

4.7.5. Separated VLR vs. combined VLR

For the evaluation of Valuable Localization Region (VLR) in Localization response distillation, the comparisons of positive region (PR), separated VLR (PR + VLR), and combined VLR (CVLR) are performed as well. Table 10 shows that when the training schedule is "1×", the combined VLR can achieve the best performance. When the training schedule is "2×", the best performance can be achieved with PR.

Table 8
The same student detector with ResNet-50-FPN backbone which is distilled by heterogeneous networks, such as ResNeXt-101, Res2Net-101, and Swin Transformer-Tiny. The results are reported on MS COCO val2017.

Backbone	TPD	AP	AP_{50}	AP_{75}
ResNeXt-101(T) [48]		46.7	64.9	51.0
ResNet-50(S) [36]		41.0	58.5	45.0
	√	**43.8(+2.8)**	**61.5**	**48.0**
Res2Net-101(T) [49]		48.7	67.1	53.2
ResNet-50(S) [36]		41.0	58.5	45.0
	√	**44.0(+3.0)**	**61.9**	**48.1**
Swin-Tiny(T) [50]		48.2	67.0	52.5
ResNet-50(S) [36]		41.0	58.5	45.0
	√	**44.1(+3.1)**	**62.0**	**48.1**

Table 9
Quantitative results of TPD on various effective object detectors. The teacher is ResNet-101 and the student is ResNet-50. The results are reported on MS COCO val2017.

Detector	TPD	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
ATSS [23]		39.2	57.3	42.4	22.7	43.1	51.5
	√	**42.3(+3.1)**	**60.1**	**45.7**	**25.3**	**46.5**	**55.0**
GFL [11]		40.2	58.4	43.3	23.3	44.0	52.2
	√	**43.0(+2.8)**	**61.1**	**46.7**	**25.5**	**47.4**	**55.7**

Table 10
Comparison of the different form of VLR, "TS" means the training schedule.

VLR	TS	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
PR	1×	43.6	61.5	48.0	25.8	47.7	56.7
	2×	45.6	63.0	50.0	28.0	49.8	58.9
PR + VLR	1×	43.6	61.3	47.9	26.0	47.8	56.6
	2×	45.2	63.0	49.7	27.8	49.2	57.5
CVLR	1×	43.8	61.3	47.9	25.9	48.1	56.7
	2×	45.4	63.2	49.9	27.9	49.4	57.8

Table 11
The comparison on PACACAL VOC dataset, and results are reported on VOC2012 test set. The same teacher detector with ResNet-101-FPN backbone distills different lightweight detectors with ResNet-18, ResNet-34, and ResNet-50 respectively. "TPD" is the proposed method.

Teacher	Student	TPD	AP	AP_{50}	AP_{75}
ResNet-101	ResNet-18		52.3	75.1	56.9
		√	**56.4(+4.1)**	**79.3**	**62.1**
	ResNet-34		55.8	78.1	61.4
		√	**58.5(+2.7)**	**80.8**	**64.3**
	ResNet-50		56.3	78.9	62.0
		√	**59.2(+2.9)**	**81.1**	**65.2**

4.7.6. Comparison of different lightweight detectors on PASCAL VOC dataset

We also make a comparison of different lightweight detectors on PASCAL VOC dataset to better demonstrate the effectiveness of TPD. The ResNet-101 with 61.0 AP is chosen as the teacher model to distill a series of lightweight student detectors, including ResNet-18, ResNet-34, and ResNet-50. As shown in Table 11, for the three given students, the AP can be stably improved by +4.1, +2.7, +2.9 without adding bells and whistles. Based on the results, the proposed method has shown a steadily improvement not only on COCO dataset, but also PASCAL VOC dataset, this also proves the robustness of TPD.

4.7.7. Weight factor of classification response distillation and localization quality estimation response distillation

Table 12
The ablation study of the weight factor λ_3 for classification response distillation, set $\lambda_3 = 2.5$ by default.

λ_3	AP	AP_{50}	AP_{75}
0	41.0	58.5	45.0
1.0	41.6	59.9	45.3
2.0	42.1	60.1	45.7
2.5	**42.3**	**60.8**	**46.2**
3.0	42.2	60.4	45.9
4.0	41.9	59.9	45.6

Table 13
The ablation study of the weight factor λ_5 for localization quality estimation response distillation, set $\lambda_5 = 10$ by default.

λ_5	AP	AP_{50}	AP_{75}
0	43.4	60.8	47.5
5	43.7	61.2	47.8
10	**43.8**	**61.3**	**47.9**
15	43.7	61.2	47.7
20	43.6	61.1	47.8
30	43.5	61.0	47.6

Table 14
The comparison with different attention mechanism methods in HRAD. IA means our proposed improved attention module, HR means our proposed hierarchical re-weighting factor.

Method	AP	AP_{50}	AP_{75}
Baseline(TPD)	43.8	61.3	47.9
TPD + Gcnet [33]	43.9 (+0.1)	61.5	47.9
TPD + FKD(Channel+Spatial+Non-local [51]) [29]	44.0(+0.2)	61.7	48.1
TPD + FGD(Channel+Spatial+Gcnet [33]) [30]	44.1(+0.3)	61.7	48.2
TPD + IA	44.2 (+0.4)	61.9	48.3
TPD + (IA + Gcnet)	44.3 (+0.5)	61.9	48.4
TPD + HRAD(IA + Gcnet + HR)	**44.5 (+0.7)**	**62.1**	**48.4**

Tables 12 and 13 show the ablation experiments of different weighting factors: λ_3 for classification response distillation, and λ_5 for localization quality estimation response distillation. Here, λ_4 is directly set 0.25 as mentioned in LD [1]. Meanwhile, the ablation experiments for classification response distillation is conducted on the baseline GFLV2 as shown in Table 12, and λ_3 is fixed to 2.5 in all the other experiments. As shown in Table 13, for the distillation of localization quality estimation response, its ablation experiments are conducted on the combination of all the proposed distillation by set $\lambda_5 = 10$ in all the other experiments.

4.8. Ablation study of HRAD

4.8.1. Comparison of different attention mechanism methods in HRAD

To demonstrate the effectiveness of HRAD more convincingly, a comparison of different attention mechanism methods in HRAD is shown in Table 14. When Gcnet [33] is combined with TPD, there is only 0.1 AP improvement. When FKD consisting of channel, spatial and Non-local [51] attention is combined with TPD, only 0.2 AP improvement has been achieved. When FGD consisting of channel, spatial and Gcnet [33] attention is combined with TPD, a 0.3 AP improvement has been achieved. When further combining TPD and our proposed improved attention (IA), a 0.4 AP improvement can be achieved. At the same time, when IA is combined with Gcnet [33], this improvement reaches 0.5 AP. Finally, when we combine TPD with our proposed HRAD, the overall result reaches 44.5, achieving a 0.7 AP improvement.

4.8.2. Different feature-based methods combined with response-based TPD

Table 15 shows the ablation study of different feature-based distillation by incorporating with TPD. When using the earlier proposed feature-based methods such as FitNet [41] and Defeat [6], the overall performance has almost no improvement. GID [42] and FKD [29] have only 0.2 AP improvement. Even FGD [30] which is already a very effective feature-based method, has only 0.3 AP improvement. But when combined with HRAD, a 0.7 AP is obtained, which demonstrates that the proposed HRAD and TPD are complementary.

Table 15
Different feature-based distillation methods when combined with same proposed response-based TPD.

Method	AP	AP_{50}	AP_{75}
Baseline(TPD)	43.8	61.3	47.9
TPD + FitNet [41]	43.8	61.2	48.0
TPD + Defeat [41]	43.9(+0.1)	61.3	47.9
TPD + GID [42]	44.0(+0.2)	61.7	48.3
TPD + FKD [29]	44.0(+0.2)	61.7	48.1
TPD + FGD [30]	44.1(+0.3)	61.7	48.2
CUD(TPD + HRAD)	**44.5 (+0.7)**	**62.1**	**48.4**

Table 16
Ablation study of weight factor α, β and γ on HRAD.

α β γ	1.1×10^{-3}	1.15×10^{-3}	1.2×10^{-3}	1.25×10^{-3}
mAP	44.3	44.4	**44.5**	44.3

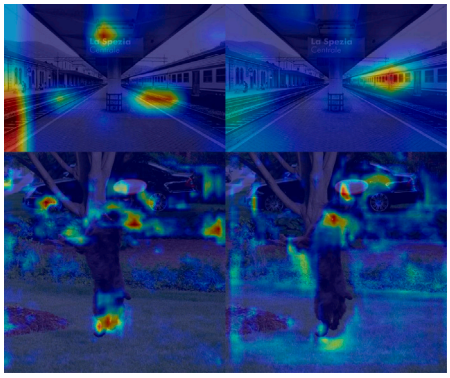


Fig. 5. Visualization of the heat maps of GFLV2 and HRAD from different pyramid layers in FPN, the first row is P_7 layer and the second row is P_4 layer.

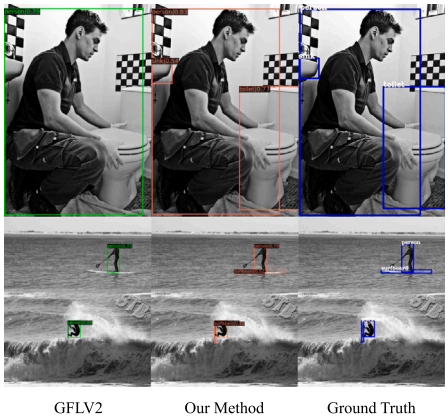


Fig. 6. Detection results of the GFLV2, Our Method and Ground Truth in infrared image scene. Our method is a great improvement over the GFLV2 that has not been distilled.

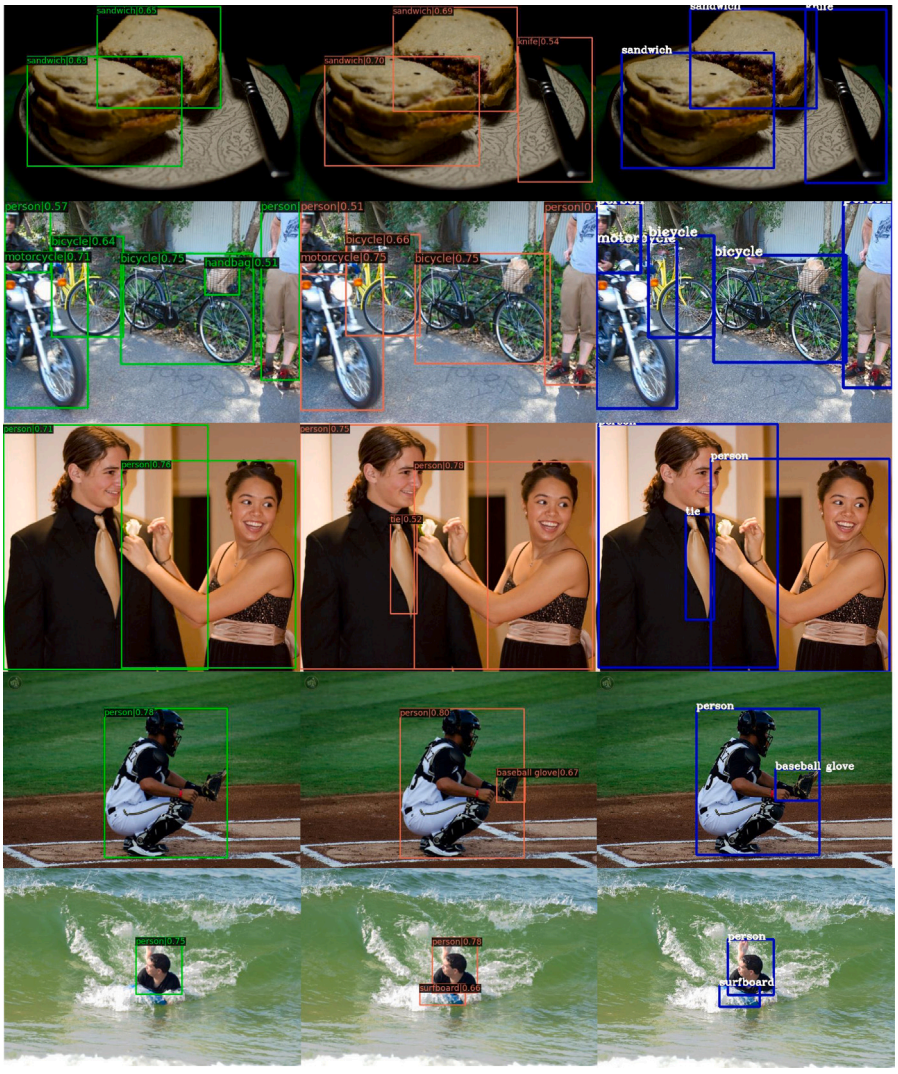


Fig. 7. Detection results of the GFLV2, Our Method and Ground Truth. Our method is a great improvement over the GFLV2 that has not been distilled.

4.8.3. Weight factor of HRAD

In Eq. (16), there are three weight factors α, β and γ for the losses of L_{maf}, L_{caf} and L_{saf}. As shown in Table 16, when α, β and γ are all fixed to 1.2×10^{-3}, Our HRAD can achieve the best performance.

4.8.4. Visualization of HRAD

To better understand the hierarchical re-weighting attention-based feature is distilled, the visualization heat maps of GFLV2 and HRAD from different pyramid layers in ResNet-50-FPN. The first row of Fig. 5 shows the heat maps of the P_7 layer of the FPN in GFLV2 and HRAD, focusing on large objects with anchor range of 512^2. It is obvious that the HRAD is more discriminating in the P_7 layer, better distinguishing the position of the train. In contrast, GFLV2 focuses its attention on the railroad tracks and station signs. The second row demonstrates the P_4 layer, and its anchor range is 64^2. Similarly, HRAD can more accurately distinguish the features contained in the ground truth such as frisbee, dog and car. These results prove the effectiveness of HRAD in each layer of FPN.

4.9. Detection results

Fig. 7 shows more detection results of GFLV2, Our Method and Ground Truth. At the same time, as shown in Fig. 6, our method not only obtains good results in the RGB image scene, but also obtains very good results in the infrared image scene.

5. Conclusion

In this work, a closed-loop unified knowledge distillation (CUD) for dense object detection is proposed, which makes the feature-based and response-based distillation unified and complementary. Accordingly, an effective response-based triple parallel distillation (TPD) is presented which takes full use of output response in dense object detection head. Then a hierarchical re-weighting attention distillation (HRAD) is proposed to enable student network to learn beyond the teacher network in feature imitation, as well as a reciprocal feedback between the classification-IoU joint representation of detection head and the attention-based feature. Experiments demonstrate that our CUD can be used as a unified distillation paradigm for dense object detention, and can be applied to other downstream tasks.

6. Limitation

Since our proposed HRAD requires the detectors to have similar number of layers and channels in the FPN layer to achieve the best results, and the Transformer-based detectors such as DETR series [52,53]

do not have the same number of channels and layers in FPN module as other kind of detectors, our CUD does not work well with the DETR series detectors, and we will explore this issue in our future work.

CRediT authorship contribution statement

Yaoye Song: Conceptualization, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Peng Zhang:** Funding acquisition, Resources, Software, Supervision, Writing – review & editing. **Wei Huang:** Project administration, Resources. **Yufei Zha:** Funding acquisition. **Tao You:** Data curation, Software. **Yanning Zhang:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

All authors declare that No conflict of interest exists.

Data availability

The data that has been used is confidential.

Acknowledgments

This work was jointly supported by grants 61971352 & 61862043 approved by National Natural Science Foundation of China. Natural Science Foundation of Ningbo (2021J048, 2021J049). The key grant 20204BC J22011 approved by Natural Science Foundation of Jiangxi Province in China.

References

[1] Z. Zheng, R. Ye, P. Wang, D. Ren, W. Zuo, Q. Hou, M.-M. Cheng, Localization distillation for dense object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.
[2] C.H. Nguyen, T.C. Nguyen, T.N. Tang, N.L. Phan, Improving object detection by label assignment distillation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 1005–1014.
[3] Z.-R. Wang, J. Du, Joint architecture and knowledge distillation in CNN for Chinese text recognition, Pattern Recognit. 111 (2021) 107722.
[4] P. Zhao, L. Xie, J. Wang, Y. Zhang, Q. Tian, Progressive privileged knowledge distillation for online action detection, Pattern Recognit. 129 (2022) 108741.
[5] Q. Li, S. Jin, J. Yan, Mimicking very efficient network for object detection, in: Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition, 2017, pp. 6356–6364.
[6] J. Guo, K. Han, Y. Wang, H. Wu, X. Chen, C. Xu, C. Xu, Distilling object detectors via decoupled features, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2154–2164.
[7] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: CVPR '14 Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
[8] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149.
[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2) (2020) 318–327.
[10] C. Wang, H. Wang, Cascaded feature fusion with multi-level self-attention mechanism for object detection, Pattern Recognit. 138 (2023) 109377.
[11] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, J. Yang, Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection, in: Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 21002–21012.
[12] X. Li, W. Wang, X. Hu, J. Li, J. Tang, J. Yang, Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11632–11641.
[13] Z. Cai, N. Vasconcelos, Cascade R-CNN: Delving into high quality object detection, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6154–6162.
[14] R. Girshick, Fast r-cnn, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2016.
[15] Y. Song, P. Zhang, W. Huang, Y. Zha, T. You, Y. Zhang, Object detection based on cortex hierarchical activation in border sensitive mechanism and classification-giou joint representation, Pattern Recognit. 137 (2023) 109278.
[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: European Conference on Computer Vision, Springer, 2016, pp. 21–37.
[17] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
[18] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7263–7271.
[19] J. Redmon, A. Farhadi, YOLOv3: An incremental improvement, 2018, arXiv e-prints.
[20] Z. Tian, C. Shen, H. Chen, T. He, FCOS: Fully convolutional one-stage object detection, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9627–9636.
[21] X. Zhou, D. Wang, P. Krähenbühl, Objects as points, 2019, arXiv preprint arXiv:1904.07850.
[22] Z. Yang, S. Liu, H. Hu, L. Wang, S. Lin, RepPoints: Point set representation for object detection, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9657–9666.
[23] S. Zhang, C. Chi, Y. Yao, Z. Lei, S.Z. Li, Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9759–9768.
[24] G. Hinton, O. Vinyals, J. Dean, et al., Distilling the knowledge in a neural network, 2015, arXiv preprint arXiv:1503.02531 2 (7).
[25] J. Gou, B. Yu, S.J. Maybank, D. Tao, Knowledge distillation: A survey, Int. J. Comput. Vis. 129 (6) (2021) 1789–1819.
[26] J. Ba, R. Caruana, Do deep nets really need to be deep? Adv. Neural Inf. Process. Syst. 27 (2014).
[27] S.I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, H. Ghasemzadeh, Improved knowledge distillation via teacher assistant, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 5191–5198.
[28] T. Wang, L. Yuan, X. Zhang, J. Feng, Distilling object detectors with fine-grained feature imitation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4933–4942.
[29] L. Zhang, K. Ma, Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors, in: International Conference on Learning Representations, 2020.
[30] Z. Yang, Z. Li, X. Jiang, Y. Gong, Z. Yuan, D. Zhao, C. Yuan, Focal and global knowledge distillation for detectors, 2021, arXiv preprint arXiv:2111.11837.
[31] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, IEEE Trans. Pattern Anal. Mach. Intell. PP (99) (2017).
[32] S. Woo, J. Park, J.Y. Lee, I.S. Kweon, CBAM: Convolutional block attention module, in: European Conference on Computer Vision, 2018.
[33] Y. Cao, J. Xu, S. Lin, F. Wei, H. Hu, Gcnet: Non-local networks meet squeeze-excitation networks and beyond, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.
[34] T.-Y. Lin, M. Maire, S.J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common objects in context, in: European Conference on Computer Vision, 2014, pp. 740–755.
[35] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, et al., MMDetection: Open mmlab detection toolbox and benchmark, 2019, arXiv preprint arXiv:1906.07155.
[36] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[37] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 936–944.
[38] N. Ketkar, Stochastic gradient descent, 2017, pp. 113–132.
[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.
[40] A. Neubeck, L.V. Gool, Efficient non-maximum suppression, in: 18th International Conference on Pattern Recognition (ICPR'06), Vol. 3, 2006, pp. 850–855.
[41] A. Romero, N. Ballas, S.E. Kahou, A. Chassang, C. Gatta, Y. Bengio, Fitnets: Hints for thin deep nets, 2014, arXiv preprint arXiv:1412.6550.
[42] X. Dai, Z. Jiang, Z. Wu, Y. Bao, Z. Wang, S. Liu, E. Zhou, General instance distillation for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7842–7851.
[43] C. Yang, M. Ochal, A. Storkey, E.J. Crowley, Prediction-guided distillation for dense object detection, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX, Springer, 2022, pp. 123–138.

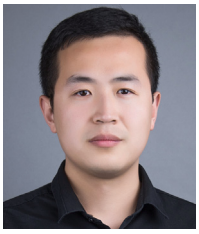
[44] T. Huang, S. You, F. Wang, C. Qian, C. Xu, Knowledge distillation from a stronger teacher, 2022, arXiv preprint arXiv:2205.10536.
[45] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2009) 303–308.
[46] D. Bolya, S. Foley, J. Hays, J. Hoffman, Tide: A general toolbox for identifying object detection errors, in: European Conference on Computer Vision, Springer, 2020, pp. 558–573.
[47] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.
[48] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1492–1500.
[49] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P. Torr, Res2net: A new multi-scale backbone architecture, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2) (2019) 652–662.
[50] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, 2021, arXiv preprint arXiv:2103.14030.
[51] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
[52] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, Springer, 2020, pp. 213–229.
[53] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable detr: Deformable transformers for end-to-end object detection, 2020, arXiv preprint arXiv:2010.04159.

Yaoye Song received the M.S. degree in School of Electrical and Information Engineering from the JiangSu University, China, in 2019. Currently, he is a Ph.D. student in School of Computer Science, Northwestern Polytechnical University, China. Her current research interests include object detection, machine learning and computer vision.

Peng Zhang received the B.E. degree from the Xian Jiaotong University, China in 2001. He received his Ph.D. from Nanyang Technological University, Singapore in 2011. He is now a professor in School of Computer Science, Northwestern Polytechnical University, China. Dr. Peng Zhang has published more than 80 research papers including CVPR, ACM Multimedia, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, IEEE Transactions on Medical Imaging and act as the PI in 3 grants of NSFC. His current research interests include computer vision, pattern recognition and machine learning. He is also the chief scientist in Mekitec OY, Finland.

Wei Huang obtained his B.Eng. and M.Eng. degrees from Harbin Institute of Technology, China in 2004 and 2006, respectively. He obtained his Ph.D. degree from Nanyang Technological University, Singapore in 2011. Before joining Nanchang University as an Associate Professor, he worked in the University of California San Diego, USA, as well as Agency for Science Technology and Research, Singapore as post-doctoral Research Fellows. Dr. Huang has published nearly 70 academic papers and acts as principal investigators in 12 national / provincial grants, including 3 NSFC grants and 2 provincial key grants. He received the best paper award of MICCAI-MLMI in 2010, the most interesting paper award of ICME-ASMMC in 2016 and was entitled the provincial young scientist of Jiangxi Province in 2015. Dr. Huangs research interests mainly include but not limited to machine learning, pattern recognition, computer vision and medical image processing.

Yufei Zha received the Ph.D. degree in information and communication engineering from Airforce Engineer University, Xian, China, in 2009. He is currently an Associate Professor with the Northwestern Polytechnical University, Xian, China. His current research interests include object detection, visual tracking, and machine learning.



Tao You received the Ph.D. degree in computer science from Northwestern Polytechnical University, Xi'an, China, in 2011. He is currently an Associate Professor and a Master's Supervisor with the School of Computer Science at Northwestern Polytechnical University, Xi'an, China. He was a Visiting Researcher with the Department of Computer, Purdue University, in 2010. His main research interests include real-time distributed computing and data mining, and complex networks. Dr. You's awards and honors include the Prize for National Defense Science and Technology, the Science and Technology Prize of Shaanxi Provincial Education Department, and the Prize for Xi'an Science and Technology Progress.

Yanning Zhang received the B.S. degree from the Department of Electronic Engineering, Dalian University of Technology, Dalian, China, in 1988, the M.S. degree from the School of Electronic Engineering, and the Ph.D. degree from the School of Marine Engineering, Northwestern Polytechnical University, Xian, China, in 1993 and 1996, respectively. She is currently a Professor at the School of Computer Science, Northwestern Polytechnical University. Her current research interests include computer vision and pattern recognition, image and video processing, and intelligent information processing. Dr. Zhang was the Organization Chair of the Asian Conference on Computer Vision 2009, and served as the program committee chairs of several international conferences.