General Instance Distillation for Object Detection

Xing Dai*[1] Zeren Jiang*†[1,2] Zhao Wu[1]
Yiping Bao[1] Zhicheng Wang[1] Si Liu[2] Erjin Zhou[1]
[1]MEGVII Technology [2]BeiHang University
daixinghome@gmail.com zeren.jiang99@gmail.com wuzhao@megvii.com
baoyiping@megvii.com wangzhicheng@megvii.com liusi@buaa.edu.cn zej@megvii.com

Abstract

In recent years, knowledge distillation has been proved to be an effective solution for model compression. This approach can make lightweight student models acquire the knowledge extracted from cumbersome teacher models. However, previous distillation methods of detection have weak generalization for different detection frameworks and rely heavily on ground truth (GT), ignoring the valuable relation information between instances. Thus, we propose a novel distillation method for detection tasks based on discriminative instances without considering the positive or negative distinguished by GT, which is called general instance distillation (GID). Our approach contains a general instance selection module (GISM) to make full use of feature-based, relation-based and response-based knowledge for distillation. Extensive results demonstrate that the student model achieves significant AP improvement and even outperforms the teacher in various detection frameworks. Specifically, RetinaNet with ResNet-50 achieves 39.1% in mAP with GID on COCO dataset, which surpasses the baseline 36.2% by 2.9%, and even better than the ResNet-101 based teacher model with 38.1% AP.

1. Introduction

In recent years, the accuracy of object detection has made a great progress due to the blossom of deep convolutional neural network (CNN). The deep learning network structure, including a variety of one-stage detection models [19, 23, 24, 25, 17] and two-stage detection models [26, 16, 8, 2], has replaced the traditional object detection and has become the mainstream method in this field. Furthermore, the anchor-free frameworks [13, 5, 32] have also achieved better performance with more simplified approaches. However, these high-precision deep learning based models are usually cumbersome, while a lightweight with high performance model is demanded in practical applications. Therefore, how to find a better trade-off between the accuracy and efficiency has become a crucial problem.

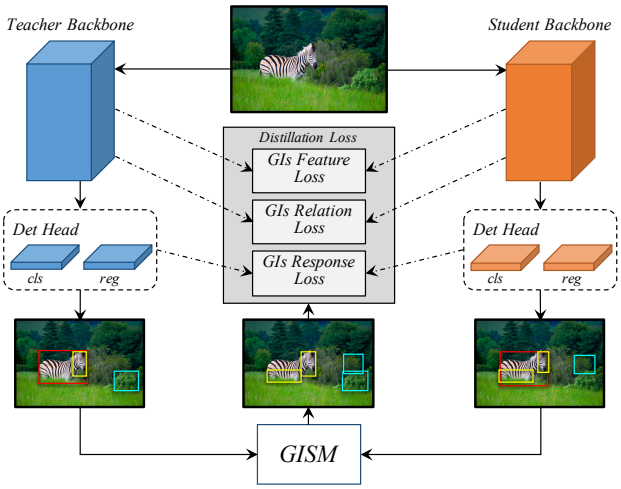


Figure 1. Overall pipeline of general instance distillation (GID). General instances (GIs) are adaptively selected by the output both from teacher and student model. Then the feature-based, relation-based and response-based knowledge are extracted for distillation based on the selected GIs.

Knowledge Distillation (KD), proposed by Hinton et al. [10], is a promising solution for the above problem. Knowledge distillation is to transfer the knowledge of large model to small model, thereby improving the performance of the small model and achieving the purpose of model compression. At present, the typical forms of knowledge can be divided into three categories [7], response-based knowledge [10, 22], feature-based knowledge [27, 35, 9] and relation-based knowledge [22, 20, 31, 33, 15]. However, most of the distillation methods are mainly designed for multi-class classification problems. Directly migrating the classification specific distillation method to the detection model is

*The first two authors contribute equally and the order is alphabetical.
†This work was done when Zeren was an intern at MEGVII Tech.

less effective, because of the extremely unbalanced ratio of positive and negative instances in the detection task. Some distillation frameworks designed for detection tasks cope with this problem and achieve impressive results, *e.g*. Li et al. [14] address the problem by distilling the positive and negative instances in a certain proportion sampled by RPN, and Wang et al. [34] further propose to only distill the near ground truth area. Nevertheless, the ratio between positive and negative instances for distillation needs to be meticulously designed, and distilling only GT-related area may ignore the potential informative area in the background. Moreover, current detection distillation methods cannot work well in multi detection frameworks simultaneously, *e.g*. two-stage, anchor-free methods. Therefore, we hope to design a general distillation method for various detection frameworks to use as much knowledge as possible effectively without concerning the positive or negative.

Towards this goal, we propose a distillation method based on discriminative instances, utilizing response-based knowledge, feature-based knowledge as well as relation-based knowledge, as shown in Fig 1. There are several advantages: (i) *We can model the relational knowledge between instances in one image for distillation.* Hu et al. [11] demonstrates the effectiveness of relational information on detection tasks. However, the relation-based knowledge distillation in object detection has not been explored yet. (ii) *We avoid manually setting the proportion of the positive and negative areas or selecting only the GT-related areas for distillation.* Though GT-related areas are almost informative, the extremely hard and simple instances may be useless, and even some informative patches from the background can be useful for students to learn the generalization of teachers. Besides, we find that the automatic selection of some discriminative instances between the student and teacher for distillation can make knowledge transferring more effective. Those discriminative instances are called general instances (GIs), since our method does not care about the proportion between positive and negative instances, nor does it rely on GT labels. (iii) *Our methods have robust generalization for various detection frameworks.* GIs are calculated upon the output from student and teacher model without relying on certain modules from a specific detector or some key characteristic, such as anchor, from a particular detection framework.

To sum up, this paper makes the following contributions:

- Define general instance (GI) as the distillation target, which can effectively improve the distillation effect of the detection model.
- Based on GI, we first introduce the relation-based knowledge for distillation on detection tasks and integrate it with response-based and feature-based knowledge, which makes student surpass the teacher.
- We verify the effectiveness of our method on the MSCOCO [18] and PASCAL VOC [6] datasets, including one-stage, two-stage and anchor-free methods, achieving state-of-the-art performance.

2. Related Work

2.1. Object Detection

The current mainstream object detection algorithms are roughly divided into two-stage and one-stage detectors. Two-stage methods [16, 8, 2] represented by Faster R-CNN [26] maintain the highest accuracy in the detection field. These methods utilize region proposal network (RPN) and refinement procedure of classification and location to obtain better performance. However, high demands for lower latency bring one-stage detectors [19, 23] under the spotlight, which achieve classification and location of targets through the feature map directly.

In recent years, another criterion divides detection algorithm into anchor-based and anchor-free methods. Anchor-based detectors such as [24, 17, 19] solve object detection tasks with the help of anchor boxes, which can be viewed as pre-defined sliding windows or proposals. Nevertheless, all anchor-based methods need to be meticulously designed and calculate a large number of anchor boxes which takes much computation. To avoid tunning hyper-parameters and calculation related to anchor boxes, anchor-free methods [23, 13, 5, 32] predict several key points of target, such as center and distance to boundaries, reach a better performance with less cost.

2.2. Knowledge Distillation

Knowledge distillation is a kind of model compression and acceleration approach which can effectively improve the performance of small models with guiding of teacher models. In knowledge distillation, knowledge takes many forms, *e.g*. the soft targets of the output layer [10], the intermediate feature map [27], the distribution of the intermediate feature [12], the activation status of each neuron [9], the mutual information of intermediate feature [1], the transformation of the intermediate feature [35] and the instance relationship [22, 20, 31, 33]. Those knowledge for distillation can be classified into the following categories [7]: response-based [10], feature-based [27, 12, 9, 1, 35], and relation-based [22, 20, 31, 33].

Recently, there are some works applying knowledge distillation to object detection tasks. Unlike the classification tasks, the distillation losses in detection tasks will encounter the extreme unbalance between positive and negative instances. Chen et al. [3] first deals with this problem by underweighting the background distillation loss in the classification head while remaining imitating the full feature map in the backbone. Li et al. [14] designs a distillation framework for two-stage detectors, applying the L2 distilla-

tion loss to the features sampled by RPN of student model, which consists of randomly sampled negative and positive proposals discriminated by ground truth (GT) labels in a certain proportion. Wang et al. [34] proposes a fine-grained feature imitation for anchor-based detectors, distilling the near objects regions which are calculated by the intersection between GT boxes and anchors generated from detectors. That is to say, the background areas will hardly be distilled even if it may contain several information-rich areas. Similar to Wang et al. [34], Sun et al. [30] only distilling the GT-related region both on feature map and detector head.

In summary, the previous distillation framework for detection tasks all manually set the ratio between distilled positive and negative instances distinguished by the GT labels to cope with the disproportion of foreground and background area in detection tasks. Thus, the main difference between our method and the previous works can be summarized as follows: (i) Our method does not rely on GT labels, nor does it care about the proportion between positive and negative instances selected for distillation. It is the information gap between student and teacher that guides the model to choose the discriminative patches for imitation. (ii) None of the previous methods take advantage of the relation-based knowledge for distillation. However, it is widely acknowledged that the relation between objects contains tremendous information even within one single image. Thus, based on our selected discriminative patches, we extract the relation-based knowledge among them for distillation, achieving further performance gain.

3. General Instance Distillation

Previous work [34] proposed that the feature regions near objects have considerable information which is useful for knowledge distillation. However, we find that not only the feature regions near objects but also the discriminative patches even from the background area have meaningful knowledge. Base on this finding, we design the general instance selection module (GISM), as shown in Fig 2. The module utilizes the predictions from both teacher and student model to select the key instances for distillation.

Furthermore, to make better use of the information provided by the teacher, we extract and take advantage of feature-based, relation-based and the response-based knowledge for distillation, as shown in Fig 3. The experimental results show that our distillation framework is general for current state-of-the-art detection models.

3.1. General Instance Selection Module

In detection model, predictions indicate the attention patches which are commonly meaningful areas. The difference of such patches between teacher and student model is also closely related to their performance gap. In order to quantify the difference for each instance and then select the discriminative instances for distillation, we propose two indicator: GI score and GI box. Both of them are dynamically calculated during each training step. For saving the computation resources during training, we simply calculate the L1 distance of classification score as GI score and choose box with higher score as GI box. Fig 2 illustrates the procedure of generating GI, and the score and box of which from each predicted instance r is defined as below.

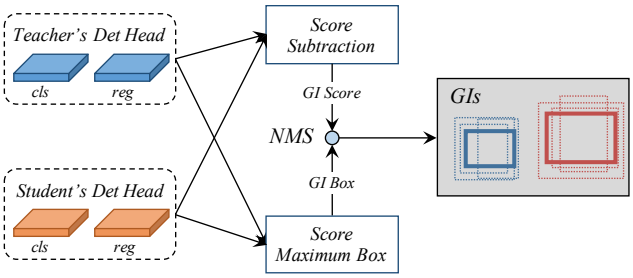


Figure 2. Illustration of the general instance selection module (GISM). To obtain the most informative locations, we calculate the L1 distance of classification scores from student and teacher as GI scores, and preserve regression boxes with higher scores as GI boxes. To avoid losses double counting, we use the non-maximum suppression (NMS) algorithm to remove duplicates.

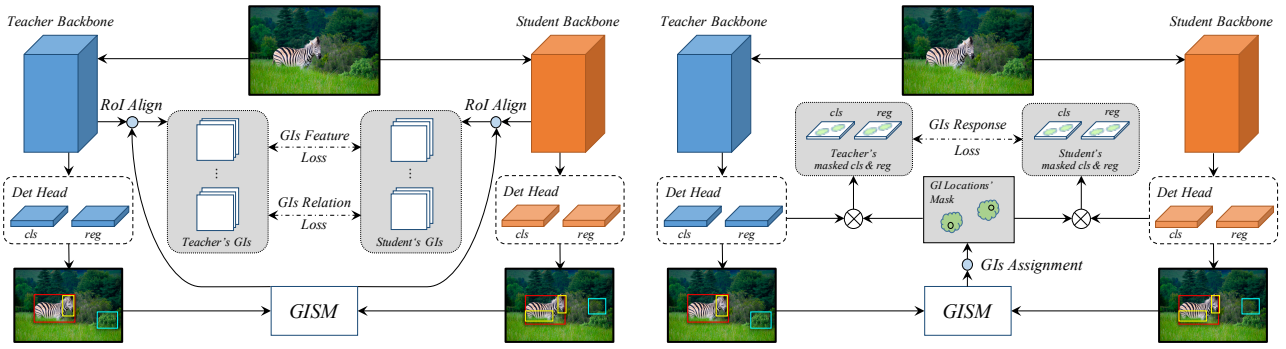
$$P^r_{GI} = \max_{0<c\leq C} |P^{rc}_t - P^{rc}_s|, \tag{1}$$

$$B^r_{GI} = \begin{cases} B^r_t, & \max\limits_{0<c\leq C} P^{rc}_t > \max\limits_{0<c\leq C} P^{rc}_s \\ B^r_s, & \max\limits_{0<c\leq C} P^{rc}_t \leq \max\limits_{0<c\leq C} P^{rc}_s \end{cases}, \tag{2}$$

$$GI = NMS(P_{GI}, B_{GI}), \tag{3}$$

where P_{GI} and B_{GI} denote GI score and GI box. For one-stage detectors, P_t and P_s are the classification score predicted by the teacher and student separately. As for two-stage detectors, P refers to the objectness score predicted by RPN. Meanwhile, B_t and B_s are the regression boxes predicted by the teacher and student, corresponding to the score P_t and P_s. R is the number of the predicted boxes and C is the number of classes. r, c are indexes in the dimension of R, C in subject to $0<c\leq C$ and $0<r\leq R$. Since we set the detection heads of teacher and student model pair to be exactly the same, so these two networks have equal R number of prediction boxes with one-to-one corresponding location.

Though we identified the indicator of GI scores and corresponding boxes, these instances with high GI scores are likely to be highly overlapped, thus leading to distillation loss double counting. To deal with these redundant and correlated regions, we use standard non-maximum suppression (NMS) to perform deduplication. Given a list of instances with GI scores and boxes, NMS works by iteratively selecting the instance with the highest GI score, and then removing all lower GI score instances that have high overlap with



(a) Feature-based and relation-based distillation

(b) Response-based distillation

Figure 3. Details of our method: (a) Selected GIs are used to crop the feature in student and teacher backbone by ROI Align. Then the feature-based and relation-based knowledge are extracted for distillation. (b) Selected GIs first generate a mask by GIs assignment. Then masked classification and regression head are distilled to utilize response-based knowledge.

the selected region. We use an IoU threshold of 0.3 to select dispersing instances. Moreover, only top K instances with the highest score are chosen as the final GI for distillation in each image.

3.2. Feature-based Distillation

Most of the SOTA detection models have introduced the Feature Pyramid Networks (FPN) [16], which can significantly improve the robustness of multi-scale detection. Since the FPN combines the feature of multiple backbone layers, we intuitively choose the FPN for distillation. To be specific, we crop the feature from the matching FPN layer according to the different size of each GI box.

Given that the target sizes vary greatly in detection tasks, directly performing pixel-wise distillation will make the model more incline to learn large targets. Therefore, as shown in Fig 3(a), we adopt the ROIAlign [8] algorithm, which resizes GI feature of different sizes to the same size and then perform distillation, treating each target equally. The feature-based distillation loss is as follows:

$$L_{Feature} = \frac{1}{K}\sum_{i=1}^{K}\|t_i - s'_i\|_2^2, \tag{4}$$

$$s' = f_{adapt}(s), \tag{5}$$

in which K is the number of GI selected by GISM with top K GI scores, t_i and s_i are the i^{th} GI feature extracted from the teacher and student model by ROIAlign algorithm, f_{adapt} is the linear adaptation function to adapt s_i to the same dimension as t_i.

3.3. Relation-based Distillation

Relational information [22, 20] between different objects has played a significant role in distillation for classification task. However, the relation-based knowledge distillation for detection tasks remains unexplored. Since the instances in the same scene are highly correlated, regardless of foreground or background, this correlation information can help the student network converge more effectively.

Owe to the informative GIs selected by GISM, we are able to take full advantage of the correlation between discriminative instances. Only performing one-to-one feature distillation is certainly not enough to import more knowledge. Therefore, to mine the valuable relation knowledge underlying a batch of GIs, we further introduce relation-based knowledge for distillation. Here we use Euclidean distance to measure the relevance of instances, and L1 distance to transfer knowledge. As shown in Fig 3(a), we additionally utilize the correlation information between GIs to distill knowledge from teacher to student. The loss expression is as follows:

$$L_{Relation} = \sum_{(i,j)\in\mathbb{K}^2} l\left(\frac{1}{\phi(t)}\|t_i - t_j\|_2, \frac{1}{\phi(s)}\|s'_i - s'_j\|_2\right),$$

$$\phi(x) = \frac{1}{|\mathbb{K}^2|}\sum_{(i,j)\in\mathbb{K}^2}\|x_i - x_j\|_2, \tag{6}$$

where $\mathbb{K}^2 = \{(i,j)\,|i \neq j, 1 \leqslant i,j \leqslant K\}$, and $\phi(\cdot)$ is a normalization factor for distance, and l denotes smooth L1 loss.

3.4. Response-based Distillation

[36] proposes that the performance gain from the knowledge distillation mainly due to the regularization of the respond-based knowledge from the teacher model. However, performing the distillation on the whole output of the detection head is detrimental to the performance of the student model. We speculate that this may be caused by the imbalance of the positive and negative samples of the detection tasks and the noise introduced by too many negative samples. Recently, some detection distillation methods [30, 3] only distill the positive sample on the detection head, ignoring the regularization effect of the discriminative nega-

Method	Faster R-CNN Res101-50		RetinaNet Res101-50		FCOS Res101-50	
	mAP	AP_{50}	mAP	AP_{50}	mAP	AP_{50}
teacher	56.3	82.8	57.3	81.9	58.4	81.6
student	54.2	82.2	55.4	80.9	56.1	80.2
Mimicking[14]	55.5	82.3	-	-	-	-
Fine-grained[34]	55.4	82.2	56.6	81.5	-	-
Fitnet[27]	55.1	82.2	55.8	81.4	57.0	80.3
Ours	**56.5**	**82.6**	**57.9**	**82.0**	**58.4**	**81.3**

Table 1. Comparison with previous work on PASCAL VOC with different detection frameworks. Some results are missing, as Mimicking and Fine-grained can only be applied to two-stage frameworks and anchor-based frameworks respectively.

tive samples. Therefore, we designed distillation masks for the classification branch and regression branch based on selected GIs, which is proved more effective than only using GT labels as the distillation mask.

However, since the definition of outputs from the detector head varies from model to model, we propose a general framework to perform the distillation on the detection head for different model, as shown in Fig 3(b). First of all, the distillation mask based on GIs is calculated as follows:

$$M = F_{Assign}(GIs), \tag{7}$$

where function F is label assignment algorithm, which is differ from model to model. It's input is the GI boxes and it's output is 1 when this output pixel is matched GI and 0 when it is not. *e.g.* For RetinaNet, we use IoU between anchors and GIs to determine whether it is masked or not. For FCOS, all the outputs outside GIs are masked.

Then response-based loss can be expressed as follows:

$$L_{Response} = \frac{1}{N_m}\sum_{i=1}^{R} M_i \left(\alpha L_{cls}\left(y_t^i, y_s^i\right) + \beta L_{reg}\left(r_t^i, r_s^i\right)\right),$$
$$N_m = \sum_{i=1}^{R} M_i, \tag{8}$$

in which y_t r_t are from teacher model while y_s r_s are from student model. y_t y_s are from output of the classification head. r_t r_s are from output of the regression head. L_{cls} and L_{reg} are the classification and regression loss function same as the task loss function of specific distilled model. It should be noted that, for two-stage detector, we distill the outputs of RPN instead for simplify.

3.5. Overall loss function

We trained the student model end-to-end, total loss for distilling student model is as follows:

$$L = L_{GT} + \lambda_1 L_{Feature} + \lambda_2 L_{Relation} + \lambda_3 L_{Response}, \tag{9}$$

where L_{GT} is task loss for detection model, λ_1, λ_2, λ_3 are hyper-parameters to balance each loss in the same scale.

4. Experiments

In order to verify the effectiveness and robustness of our method, we conduct experiments on different detection frameworks, heterogeneous backbones and few classes detection with COCO and Pascal VOC dataset. Specifically, following the setting in [26], for the Pascal VOC dataset, we choose the 5k trainval images split in VOC 2007 and 16k trainval images split in VOC 2012 for training and 5k test images split in VOC 2007 for test. While for COCO, we choose the default 120k train images split for training and 5k val images split for test. All the distillation performances are evaluated in average precision (AP).

We adopt the hyper-parameters $\{K = 10, \lambda_1 = 5 \times 10^{-4}, \lambda_2 = 40, \lambda_3 = 1, \alpha = 0.1, \beta = 1\}$ for all experiments by diagnosing the initial loss of each knowledge type and ensuring that all losses are within the same scale. Unless specified, we use 2x learning schedule to train 24 epochs (180000 iterations) on COCO dataset and 17.4 epochs (18000 iterations) on VOC dataset for distillation.

4.1. Different detection frameworks

We evaluate our method based on three state-of-the-art detection frameworks, anchor-based one-stage detector (RetinaNet), anchor-free one-stage detector (FCOS), and two-stage detector (Faster R-CNN). Among those three models, the distillation for feature-based and relation-based distillation is exactly the same. However, as the target definition for detection task in each model is different, the form of the response-based distillation loss is also different, *e.g.* following the original loss setting, we choose IoU loss and smooth L1 loss for FCOS and RetinaNet separately.

As for the backbone, we choose shallower student backbone with similar architecture of teacher model. To be specific, we choose ResNet-50 based model as student model, and ResNet-101 based model as teacher model. As shown in Table 1, we compare our method with SOTA detection distillation methods on Pascal VOC. The results shows that our method outperforms the previous SOTA methods to a large extent and even surpasses the teacher model.

As shown in Table 2, we also perform experiments with

Method	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	mAR	AR_S	AR_M	AR_L
Retina-Res101 (teacher)	38.1	58.3	40.9	21.2	42.3	51.1	54.4	34.1	59.1	70.5
Retina-Res50 (student)	36.2	55.8	38.8	20.7	39.5	48.7	52.1	33.7	55.3	68.6
+ Fitnet[27]	37.4	57.1	40.0	20.8	40.8	50.9	53.5	33.6	57.4	69.4
+ Fine-grained[34]	38.6	58.7	41.3	21.4	42.5	51.5	54.6	34.7	58.2	70.4
+ Ours	39.1	59.0	42.3	22.8	43.1	52.3	55.3	36.7	59.1	71.1
Our gain	**+2.9**	**+3.2**	**+3.5**	**+2.1**	**+3.6**	**+3.6**	**+3.2**	**+3.0**	**+3.8**	**+2.5**
FCOS-Res101 (teacher)	41.0	60.3	44.2	24.6	44.8	52.8	58.8	39.8	63.9	74.0
FCOS-Res50 (student)	38.5	57.0	41.3	21.4	41.8	50.7	56.1	34.6	60.3	72.0
+ Fitnet[27]	39.9	58.6	43.1	23.1	43.4	52.2	57.3	36.6	61.5	73.1
+ Ours	42.0	60.4	45.5	25.6	45.8	54.2	59.9	39.9	64.3	75.8
Our gain	**+3.5**	**+3.4**	**+4.2**	**+4.2**	**+4.0**	**+3.5**	**+3.8**	**+5.3**	**+4.0**	**+3.8**
R-CNN-Res101 (teacher)	39.6	60.6	43.1	22.7	43.3	51.9	53.3	32.8	57.5	67.7
R-CNN-Res50 (student)	38.3	58.8	41.7	21.4	41.6	50.5	51.5	30.9	55.2	65.9
+ Fitnet[27]	38.9	59.5	42.4	21.9	42.2	51.6	52.3	32.5	55.4	66.7
+ Mimicking[14]	39.6	60.1	43.3	22.5	42.8	52.2	52.9	33.1	56.1	67.6
+ Fine-grained[34]	39.3	59.8	42.9	22.5	42.3	52.2	52.4	32.2	55.7	67.9
+ Ours	40.2	60.7	43.8	22.7	44.0	53.2	53.9	33.5	57.6	68.8
Our gain	**+1.9**	**+1.9**	**+2.1**	**+1.3**	**+2.4**	**+2.7**	**+2.4**	**+2.6**	**+2.4**	**+2.9**

Table 2. Results of the proposed GID on COCO dataset with different detection frameworks.

COCO dataset. All student models get significant performance gains from the teacher by our method and reach a comparable result to the teacher models, *e.g.* the ResNet-50 based RetinaNet student model gets 2.9 absolute gain in mAP, which totally recovers the performance drop due to the shallow backbone. Especially, our method achieves a further APs gain compared to other feature-based methods, since we treat each instance equally, regardless of the proportion of the instance in the feature map. Those results demonstrate that our method is applicable to a variety of widely used detection frameworks.

4.2. Heterogeneous network backbones

To further verify the generalization of our methods, instead of using homogeneous ResNet backbones for distillation, we introduce two heterogeneous backbones. Specifically, we take MobileNet-v2 [28] based RetinaNet as student and ResNet-101 based one as teacher. As shown in Table 3, the lightweight MobileNet-V2 based detector gets a 2.5 absolute mAP gain even if the basic network module is different between student and teacher.

4.3. Distillation with fewer classes

[21] points out that the information distilled is linear in the number of classes, so distillation is considerably less efficient for models with few classes. However, our method will adaptively select highly informative areas to distill and take advantage of all kinds of knowledge from the teacher model. Thus, we get impressive results, as shown in Table 4, when only a single class (person) is considered in the COCO dataset. The student model still exceeds the teacher model by a large margin in terms of few classes.

Model	Teacher	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Retina-R101	-	38.1	58.3	40.9	21.2	42.3	51.1
Retina-Mob	-	31.0	48.9	32.7	16.4	33.8	42.6
Retina-Mob	Retina-R101	33.5	51.9	35.5	19.2	36.9	44.3
		+2.5	**+3.0**	**+2.8**	**+2.8**	**+3.1**	**+1.7**

Table 3. GID results on COCO dataset with heterogeneous network backbones.

Model	Teacher	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Retina-R101	-	52.1	81.2	54.6	32.8	59.3	71.8
Retina-R50	-	50.4	79.4	53.0	31.6	56.9	70.2
Retina-R50	Retina-R101	52.8	81.3	55.4	33.9	59.4	72.5
		+2.4	**+1.9**	**+2.4**	**+2.3**	**+2.5**	**+2.3**

Table 4. GID results on COCO dataset with only Person class.

4.4. Analysis

4.4.1 Visualization of General Instances

To better understand the general instances we selected to be distilled, we visualize the selected general instances and the corresponding heat maps of the GI score distribution from different training stages with RetinaNet-Res101-50 (ResNet-101 based teacher and ResNet-50 based student). As shown in Figure 4, the green boxes denote the ground-truth label, the other color boxes denote the general instance

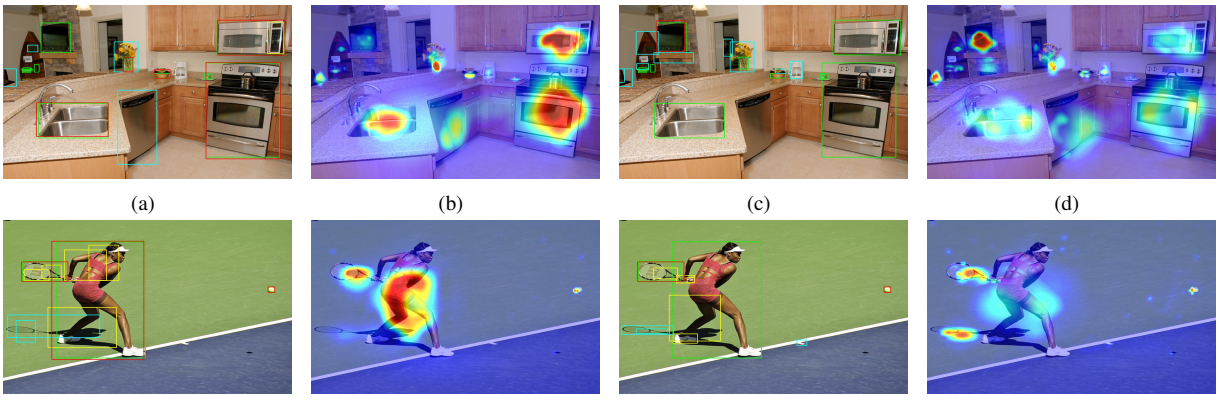


Figure 4. Examples from COCO of GIs selected by GISM and the corresponding heat maps of the GI score distribution with RetinaNet-Res101-50 model. The instances from (a)(e) are calculated by the teacher and student with 5000 iterations, while the instances from (c)(g) are calculated by the teacher and student model with 90000 iterations. The green, red, yellow and cyan boxes denote ground truth, positive, semi-positive and negative instances respectively, as defined in Sec 4.4.2. For clarity, we only visualize the GI with top10 GI score.

used to be distilled. The characteristic of general instances can be summarized into the following categories:

Key characteristic patches. As shown in the yellow boxes in Fig 4(e) 4(g), some key features of the athlete are selected such as the shoes and clothes, which are critical areas for distillation and are explained as visual concepts in [4].

Extra instances. Some confusing background areas with discriminative semantic information are also selected for distillation. For example, the oven-like machine and the shadow of the tennis racket shown in the cyan boxes in Fig 4(a) 4(g) are selected due to the inconsistent score distribution between the student model and teacher model, though some of those background instances do not include in the original 80 classes of the COCO dataset.

More informative positive instances. Compared with ground truth boxes (green), we only choose a subset among those for distillation as shown in the red boxes in Fig 4. In contrast to Online Hard Example Mining (OHEM) [29] which adaptively selects the hardest examples for training, in our method, the extremely hard instances are discarded for distillation, such as the small wine glasses in Fig 4(a), which are hard to be detected from both the student and teacher model. While in the late stage of training, the majority of ground truth instances are simple for both teacher and student models and some of them are ignored for distillation. As shown in Fig 4(c) 4(g), the athlete and the microwave oven are neglected, only some positive instances which are hard to learn from teacher remain to be distilled, like the tennis ball.

Time-varying distillation tendency. With the training goes on, the focus of the distillation targets has shifted from simple positive instances to small confusing patches, as shown in the heat maps of GI score in Fig 4(b) 4(f) from iteration 5000 and Fig 4(d) 4(h) from iteration 90000. This distillation tendency is similar to a human cognitive process. The above scenarios demonstrate that the GISM will adaptively choose the most informative and discriminative patches for distillation during training.

4.4.2 Performance gain from General Instance

To further analyze the contribution of each type of general instances and verify the effectiveness of GISM, we perform experiments on each type of general instances. We introduce an index named intersection over proposals (IoP) to help us separate those GIs:

$$IoP = \frac{area(GI \cap GT)}{area(GI)} \tag{10}$$

Then we define each type of GIs as follows:

$$GI = \begin{cases} \text{Pos} & \text{IoU} > 0.5 \\ \text{Semi-Pos} & \text{IoU} \leq 0.5,\ \text{IoP} > 0.7 \\ \text{Neg} & \text{IoU} \leq 0.5,\ \text{IoP} < 0.3 \end{cases}, \tag{11}$$

where the IoU means the intersection over union between GI and GT. Pos, Semi-Pos, Neg are short for positive, semi-positive and negative instances respectively. Besides, in ablation study, we ignore those instances with an IoP between 0.3 to 0.7 to analyse the contribution of each part of the general instances more clearly.

As shown in Table 5, distilling each type of instance can all bring performance gain to the student model, while combining all three types can achieve the best performance. One thing that's very noticeable is that performing distillation only on negative instances is still beneficial to the student model, which is a strong evidence that our approach effectively selects the useful information from the background

Model	RetinaNet Res101-50					
	GT	GI				
Positive	-	√	-	-	√	√
Semi-Positive	-	-	√	-	√	√
Negative	-	-	-	√	-	√
mAP	38.5	38.8	38.6	38.2	39.0	**39.1**

Table 5. Ablation Study for each type of General Instances, including positive, semi-positive and negative Instance. GT denotes that we use GT instance for distillation.

area while filters detrimental knowledge. Moreover, even if positive instances chosen by GISM is only a subset of GT instances, the result from positive instances surpasses that from GT instance, which indicate that some extreme hard or simple GT instances for both teacher and student will be detrimental for distillation. Besides, it is still effective to use the GT region for distillation. The essence is that the GT region is still the most informative and discriminative in the early stage of training. However, ignoring the hidden information in the background will make the student model fail to achieve better performance.

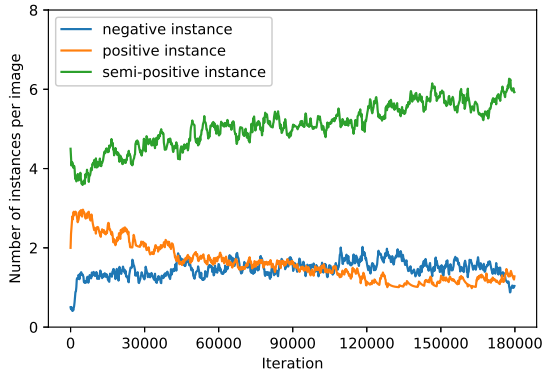


Figure 5. The curve of the instances number with training iteration on RetinaNet-Res101-50. For a better demonstration, we perform exponential smoothing on the instance number with $\alpha = 0.9$.

As shown in Fig 5, We drew the curves of these three types instances with the number of iterations. It can be seen that the proportion of positive distillation instances in the late training period is reduced, indicating that the improvement in the late training period is mainly brought by the discriminative background area and key characteristic patches.

4.4.3 Performance gain from various knowledge

In this subsection, we conduct several ablation experiments to understand how each type of knowledge component contributes to the final performance. As shown in Table 6, each distillation component improves the performance, especially for the feature-based and response-based knowledge which brings improvement about 1.7 mAP gains. The combination of the above knowledge achieves the best results, which brings about another 1.2 mAP gains compared with the best single component. That is to say, although the three types of knowledge contain some overlapping parts, we take advantage of the unique parts among them.

Model	Res50	RetinaNet Res101-50				
Feature-based	-	√	-	-	√	√
Relation-based	-	-	√	-	-	√
Response-based	-	-	-	√	√	√
mAP	36.2	37.9	37.3	37.9	38.7	**39.1**

Table 6. Ablation Study for each part of the distillation loss.

4.4.4 Varying top K for GISM

We investigate the influence of different GISM hyperparameter top K with RetinaNet-Res101-50 model. As shown in Table 7, when $K = 0$, no general instance is chosen, thus distillation loss is not applied. As the K increases, the student model gets a significant mAP gain, even with 5 distilled GIs, which is strong evidence that our approach selects the most worthy instances for distillation. The performance becomes stable and optimal when the K is at the range of 10 to 100, while it starts decreasing when K still goes further, which is mainly because the informative and discriminative GIs are overwhelmed by trivial instances.

Top K	0	5	10	40	100	300
mAP	36.2	38.9	**39.1**	39.0	**39.1**	38.6

Table 7. Hyperparameter analysis of top K GI score.

5. Conclusion

We propose GID framework that adaptively selects the most discriminative instances between teacher and student for distillation. Besides, our method effectively improves the performance of modern detection frameworks with feature-based, relation-based and response-based knowledge, and is applicable to various detection frameworks. The ablation study demonstrates that imitating some of the GT instances will do harm to the performance while even some instances from backgrounds can be helpful, which will give some insights for future distillation works.

Acknowledgment. This paper is supported by the National Key R&D Plan of the Ministry of Science and Technology ("Grid function expansion technology and equipment for community risk prevention", Project No. 2018YFC0809704).

References

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 1, 2

[3] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 742–751. Curran Associates, Inc., 2017. 2, 4

[4] Xu Cheng, Zhefan Rao, Yilan Chen, and Quanshi Zhang. Explaining knowledge distillation by quantifying the knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 7

[5] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6569–6578, 2019. 1, 2

[6] Mark Everingham, Luc Van Gool, C. K. I. Williams, J. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge, 2010. 2

[7] Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. Knowledge distillation: A survey, 2020. 1, 2

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 2, 4

[9] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. *CoRR*, abs/1811.03233, 2018. 1, 2

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 1, 2

[11] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[12] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer, 2019. 2

[13] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018. 1, 2

[14] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2, 5, 6

[15] Xiaojie Li, Jianlong Wu, Hongyu Fang, Yue Liao, Fei Wang, and Chen Qian. Local correlation consistency for knowledge distillation. In *European Conference on Computer Vision*, pages 18–33. Springer, 2020. 1

[16] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2, 4

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 2

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 2

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 1, 2

[20] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. Knowledge distillation via instance relationship graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 4

[21] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. Subclass Distillation. *arXiv e-prints*, page arXiv:2002.03936, Feb. 2020. 6

[22] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 4

[23] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1, 2

[24] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2

[25] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015. 1, 2, 5

[27] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2014. 1, 2, 5, 6

[28] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR*, abs/1801.04381, 2018. 6

[29] Abhinav Shrivastava, Abhinav Gupta, and Ross B. Girshick. Training region-based object detectors with online hard example mining. *CoRR*, abs/1604.03540, 2016. 7

[30] Ruoyu Sun, Fuhui Tang, Xiaopeng Zhang, Hongkai Xiong, and Qi Tian. Distilling object detectors with task adaptive regularization, 2020. 3, 4

[31] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation, 2019. 1, 2

[32] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1, 2

[33] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1, 2

[34] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 3, 5, 6

[35] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2

[36] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4