Adaptive Multi-task Learning for Few-shot Object Detection

Yan Ren[1*], Yanling Li[1,2*,†], and Adams Wai-Kin Kong[1]

[1] Nanyang Technological University, Singapore,
[2] The Second Research Institute of CAAC, China

Abstract. The majority of few-shot object detection methods use a shared feature map for both classification and localization, despite the conflicting requirements of these two tasks. Localization needs scale and positional sensitive features, whereas classification requires features that are robust to scale and positional variations. Although few methods have recognized this challenge and attempted to address it, they may not provide a comprehensive resolution to the issue. To overcome the contradictory preferences between classification and localization in few-shot object detection, an adaptive multi-task learning method, featuring a novel precision-driven gradient balancer, is proposed. This balancer effectively mitigates the conflicts by dynamically adjusting the backward gradient ratios for both tasks. Furthermore, a knowledge distillation and classification refinement scheme based on CLIP is introduced, aiming to enhance individual tasks by leveraging the capabilities of large vision-language models. Experimental results of the proposed method consistently show improvements over strong few-shot detection baselines on benchmark datasets. https://github.com/RY-Paper/MTL-FSOD

Keywords: Few-shot Object Detection · Multi-task Learning

1 Introduction

Rapid advancement of deep learning techniques has led to significant improvements in object detection [60]. However, it requires great effort to prepare densely annotated data for learning, such as refined object bounding boxes and their corresponding classification labels. In some data-scarce scenarios such as medical landmark detection [33] and animal species recognition [44], it is very difficult to collect adequate data. In response to these problems, few-shot object detection (FSOD) [1, 13, 15] has attracted growing research interest in recent years. FSOD aims to detect novel objects by generalizing knowledge learned from base (seen) classes to novel (unseen) classes, with sufficient training samples in the former and limited training samples in the latter. Several popular object detection frameworks, including Faster R-CNN [36], YOLO [35], and DETR [53],

[*] These authors contributed equally to this work

[†] This work was fully conducted during the author's PhD studies at NTU.

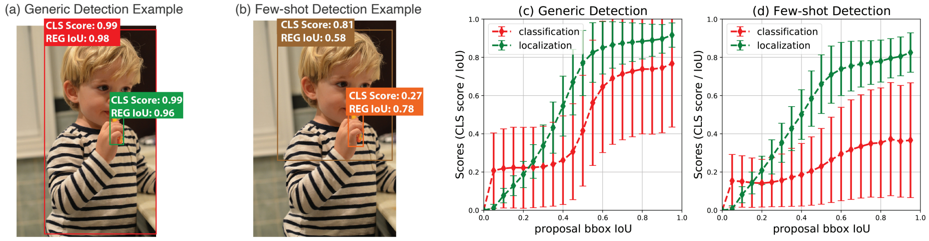

Fig. 1: An FSOD method [32] is trained under two distinct conditions: full-training (generic) and 10-shots (few-shot) on COCO-novel dataset. (a-b) Generic detection with abundant training data leads to balanced classification and regression IoU scores, while the few-shot setting shows higher score misalignments. (c-d) The gap between classification scores and regression IoU scores widens notably in few-shot scenarios compared to generic object detection. Metric details can be found in [46].

have been combined with meta-learning or transfer-learning techniques to build FSOD models, as demonstrated in previous studies such as [1, 10, 15, 41, 49]. While these methods have achieved many promising results, object detection under few-shot settings remains challenging due to the difficulty in preventing over-fitting and improving the generalization capability of the model.

Object detection is a multi-task learning problem involving two subtasks: classification for determining “what class the object belongs to" and localization for determining “where the object is". They have two distinct differences. The first is the conflicting translation and scale properties [16]. A properly trained classifier has features that are insensitive to the changes in object scale and position, as long as they belong to the same class, while localization relies on features that are responsive to the shifts in scale and position. The second is the conflicting backward task-gradients [3,31]. If the gradient norm of the classification loss with respect to the shared backbone parameters is much greater than that of the regression loss, then the training process would be dominated by the classification task, and vice versa. The discrepancy becomes more pronounced in few-shot settings as opposed to the generic detection problem (Fig. 1). Any bias toward either task would negatively affect the performance of FSOD. Nonetheless, most FSOD models utilize shared networks to carry out these two tasks, leading to an increased disparity between them. There are few studies utilizing task-specific models for FSOD in the current literature [27,32]. While strengthening individual tasks can indeed enhance overall performance in multi-task scenarios, multiple tasks with conflicting needs can hurt rather than facilitate the joint learning processes. Consequently, this could potentially result in lower performance compared to single-task learning [56].

To bridge the dissonance between the two specific tasks in this paper, we propose an adaptive multi-task learning method for FSOD. To handle task conflicts, a novel precision-driven gradient balancer is introduced to dynamically harmonize the learning process with a double-head R-CNN model. This balancer, pre-trained on a few-shot dataset, exhibits the capability of adaptively rescaling the backward gradients from the classification and localization heads. This adaptive rescaling not only proficiently alleviates the gradient conflicts but also substantially elevates the learning performance. Moreover, to enhance individual tasks, we devise a knowledge distillation and score refinement scheme (named KDSR) leveraging both image and text embeddings from a pre-trained CLIP [34]. During the training phase, CLIP is adapted as a teacher network to guide the learning of the classification head. During the inference phase, CLIP is further adapted to aggregate the detection scores. This scheme effectively inherits CLIP's powerful capabilities, particularly critical in data-scarce scenarios. In contrast to the previous method [19], which solely leverages CLIP's text encoder, KDSR aims to unleash the full potential of CLIP for FSOD by utilizing adapted image-text features. The experimental results demonstrate that the proposed method achieves state-of-the-art performance on benchmark datasets. The contributions of the proposed method are summarized as follows:

- A precision-driven gradient balancer can effectively mitigate task disparities and elevate overall learning efficiency.
- A KDSR scheme can enhance individual-task performance through knowledge distillation during training, and score refinement during inference, both using adapted image-text features of CLIP.
- Experimental results demonstrate that the proposed method achieves state-of-the-art FSOD results across benchmark datasets.

2 Related Work

2.1 Few-shot Object Detection

Existing FSOD methods can be categorized into two types: meta-learning based methods and transfer-learning based methods. Most FSOD methods are based on meta-learning [10,15,49,53], which focuses on learning to learn task-level knowledge that can be generalized to novel classes. Nevertheless, the meta-learning based methods require a complicated training process that depends on carefully structured query-support episodes as inputs. Additionally, separate feedforward processes must be conducted for each object class. Consequently, these methods usually suffer from a huge computational burden. Transfer-learning based methods learn novel concepts via simple fine-tuning on top of a pre-trained base model, which is more straightforward and efficient [41,45,52]. For example, a regularized transfer learning framework is developed in [1] to leverage the knowledge of base classes for enhancing the detection performance of novel classes. The Multi-scale Positive Sample Refinement (MPSR) [45] scheme enriches the transfer-learning based Faster R-CNN model with multi-scale features. A two-stage fine-tuning approach (TFA) [41] outperforms all previous meta-learning approaches. However, none of the methods fully addresses the disparities between the classification and localization tasks in few-shot scenarios.

2.2 Multi-task Learning

Multi-task learning [56] (MTL) leverages a joint model to concurrently address multiple predictive tasks. Joint training enables the extraction of shared concepts among related tasks, leading to the reduction of computation costs and the enhancement of data utilization efficiency. However, the negative transfer problem [24] arises when tasks have conflicting requirements. The conflict in learning dynamics can be detrimental, rather than beneficial, to the learning process. To mitigate this issue, three strategies can be considered: decoupling the models, addressing gradient modulation, and enhancing task-specific branches

Decoupled Models with task-specific branches have been considered by MTL architectures to minimize the negative transfer [28, 57]. In generic object detection, task-specific modules can offer superior classification accuracy and localization precision. For instance, Chen et al. [46] develop a double-head Faster R-CNN, splitting the shared RCNN head into classification and localization branches. Song et al. [38] address spatial misalignment between the tasks through task-aware spatial disentanglement. Nonetheless, the prior methods overlook the challenges posed by few-shot scenarios characterized by limited available data. A few FSOD methods have addressed the issue of task separation. For instance, DMNet [27] leverages decoupled representation transformation to adaptively generate representations, although its performance falls short of state-of-the-art (SOTA) results. DeFRCN [32] achieves comparable outcomes by decoupling the region proposal network (RPN) from the region-based convolutional neural network (RCNN). Nonetheless, these approaches do not comprehensively explore the potential benefits of task-specific branches in FSOD.

Gradient Modulation serves as another solution to address the issue of negative transfer [3]. The backpropagated gradients of different tasks exhibit conflicting directions or varying magnitudes. Conflicting gradients can result in imbalances or the dominance of certain tasks during the backpropagation toward the shared modules. To address this problem, several MTL methods have introduced gradient modulation techniques to merge them into a well-balanced joint gradient. For instance, gradient surgery is proposed in [51] to project task gradients onto the normal plane of the conflicting gradients. Nash Bargaining Solution is employed in [30] to treat the gradient combination as a cooperative bargaining game. Impartial multi-task learning method [23] utilizes a closed-form gradient balance method to mitigate task bias. However, existing MTL methods often emphasize gradient modulation via closed-form optimization techniques, disregarding the domain knowledge of task-specific data, and they have not been explored in FSOD. In FSOD, depending solely on static closed-form solutions may fall short in addressing the challenges posed by data scarcity.

Enhanced Task-specific Branches can contribute to enhancing the overall performance of MTL. Two task-specific strategies can be considered: 1) **Knowledge Distillation** (KD) guides the learning of task-specific student networks with the knowledge of individual single-task teacher networks [11]. For some MTL applications, the distilled student networks can match or even outperform the single-task teachers [5, 37]. For FSOD, a backdoor adjustment-based KD

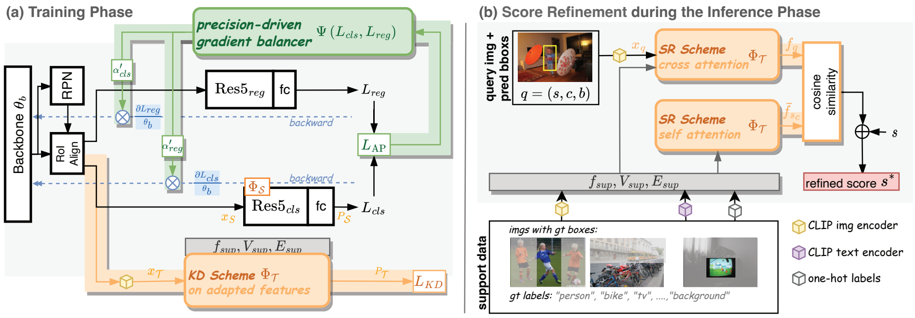

Fig. 2: Workflow of the proposed method. a) During training, a novel precision-driven gradient balancer is introduced within a double-head RCNN to address negative transfer issues in the backward task gradients. Meanwhile, an adapted-CLIP-based knowledge distillation module is developed to enhance the classification branch. b) During inference, a detection score refinement module is constructed to further enhance the FSOD performance. Notation details are given in Sec.3.

method is introduced in [19] with the text encoder of CLIP [34] as a teacher network. Nevertheless, relying solely on a pre-trained text encoder has not fully exploited the potential of CLIP. 2) **Refinement Strategy** aims to enhance the precision and accuracy of specific tasks. A range of studies has introduced various refinement strategies. For example, Cheng *et al.* [4] demonstrate a decoupled classification refinement (DCR) network that utilizes hard false positive samples. Li *et al.* [20] devise a few-shot correction network to learn from false positive samples of novel classes. Nonetheless, in few-shot scenarios, it is challenging to collect enough false positive samples to train the correction network adequately. In [32], an offline prototypical calibration block (PCB) utilizing a robust classifier pre-trained on ImageNet [6] is proposed for classification refinement in FSOD. Although large-scale pre-trained models like CLIP have demonstrated promising results in some few-shot classification methods [39, 54], none of the FSOD methods has considered employing adapted image-text embeddings for precision refinement. This paper aims to unleash the full potential of CLIP for FSOD by employing a combination of knowledge distillation and refinement strategies.

3 Methodology

3.1 Problem Formulation

Consistent with previous research on FSOD [1, 15, 41], transfer-learning based methods involve two training stages: base-training on a large dataset D_{base} with abundant annotated instances, followed by fine-tuning on a novel support dataset D_{novel} with very few annotated object instances. There are no overlap between the base classes C_{base} in D_{base} and the novel classes C_{novel} in D_{novel} , $C_{base} \cap C_{novel} = \varnothing$. In K-shot FSOD, exact K instance annotations of each novel class

Table 1: Performance comparison of gradient modulization methods on the enhanced baseline. No KDSR or PCB is involved. More details in Sec.4.3

Methods	mAP
Baseline (without PCB) [32]	16.7
Enhanced Baseline (without PCB)	**17.4**
Enhanced Baseline + PCGrad [51]	9.8
Enhanced Baseline + CAGrad [22]	17.9
Enhanced Baseline + GradVac [43]	17.3
Enhanced Baseline + IMTLG [23]	18.0
Enhanced Baseline + NashMTL [30]	17.2
Ours (without KDSR)	**18.6**

are given in training. In inference, a query set D_{query} with unseen objects of novel classes is used to evaluate the detection performance.

3.2 Enhanced Baseline with Double-head R-CNN

The majority of existing FSOD methods [10,41,49] use Faster R-CNN [36] as the basis of their architectures. Compared to other detection architectures such as DETR [53], Faster R-CNN is known for its efficiency and robustness. DeFRCN [32], which is built upon Faster R-CNN and achieves SOTA detection results, serves as the baseline of our method. The typical training loss L of Faster R-CNN is defined as

$$\min L = L_{rpn} + \delta_1 L_{rcnn} = L_{rpn} + \delta_1 \left(L_{cls} + \delta_2 L_{reg} \right), \tag{1}$$

where L_{rpn} encapsulates the loss within the RPN, combining a bounding box regression loss and an objectness classification loss to iteratively refine the region proposals. L_{rcnn} denotes the losses in RCNN consisting of the cross-entropy-based classification loss L_{cls} and the smooth L1-based localization loss L_{reg} [9]. The trade-off coefficients δ_1 and δ_2 are set to 1 in DeFRCN. Based on Eq.1, DeFRCN rescales the backward gradients of RPN and RCNN by constant rescaling coefficients β_{rpn} and β_{rcnn}, before propagating them back to the backbone parameters θ_b:

$$\theta_b \leftarrow \theta_b - \mu \left(\beta_{rpn} \cdot \frac{\partial L_{rpn}}{\partial \theta_b} + \beta_{rcnn} \cdot \frac{\partial L_{rcnn}}{\partial \theta_b} \right). \tag{2}$$

The rescaling approach yields a remarkable boost in detection performance. However, akin to previous methods, DeFRCN utilizes one shared RCNN network with two linear output layers to jointly handle object classification and bounding box regression without considering the conflicting properties of the two tasks.

To decouple classification from regression, we enhance the baseline with a double-head RCNN, dividing it into task-specific branches for classification and localization using Res5 residual blocks (see Res5$_{cls}$ and Res5$_{reg}$ in Fig. 2 (a)). The two Res5 branches have separate learning weights to acquire task-specific features, and the regression head is class-agnostic, ensuring complete separation. This straightforward enhancement leads to improved performance over the baseline, as shown in Tab. 1.

3.3 Adaptively Learning with Multiple Tasks

Building upon the enhanced baseline, the proposed method addresses the conflicting requirements of the two tasks by rescaling the backward gradients of the two task-specific branches before forwarding them back to the backbone.

$$\frac{\partial L_{rcnn}}{\partial \theta_b} = \alpha_{cls} \cdot \frac{\partial L_{cls}}{\partial \theta_b} + \alpha_{reg} \cdot \frac{\partial L_{reg}}{\partial \theta_b}. \tag{3}$$

where α_{cls} and α_{reg} represent the rescaling ratios of classification and regression gradients. We emphasize that α_{cls} and α_{reg} are applied to the backward gradients of the backbone, not to the task-specific branches. This is because the task-specific branches are entirely separate, eliminating any conflicts.

The optimal trade-off between α_{cls} and α_{reg} is vital to the performance. As previously mentioned, MTL techniques can effectively alleviate disparities between the two specific tasks. To demonstrate this, we apply a group of SOTA gradient modulation methods to the proposed double-head RCNN model (see Tab. 1). These methods employ closed-form optimizations to determine α_{cls} and α_{reg} with the aim of achieving a balanced backward gradient. Only two of these methods [22, 23] can obtain an improvement compared with the enhanced baseline. Note that none of these gradient modulation methods are originally developed for FSOD. Directly applying these closed-form methods to FSOD would neglect its properties, e.g., the objective of detection and the few-shot setting. More particularly, they lack the adaptability to incorporate task-specific few-shot data into the learning. We propose a precision-driven gradient balancer to adaptively leverage the available few-shot data and address the challenges posed by the conflicted gradients.

The performance of FSOD methods is commonly measured by the average precision (AP), which depends on both classification accuracy and localization precision. Instead of closed-form optimization, the proposed method defines an additional loss term of AP as L_{AP}. During the training process, L_{AP} can be directly utilized with the other losses to guide the generation of α_{cls} and α_{reg},

$$L_{AP} = AP_{max} - AP_{pred} = AP_{max} - \Psi(L_{cls}, L_{reg}), \tag{4}$$

in which a constant $AP_{max} = 100$ represents the maximum precision objective value, and AP_{pred} represents the predicted AP. Ψ is a simulated function of the classification and regression losses. AP indicates the detection performance, which depends on a series of classification and regression measurements. These measurements include precision and recall curves based on True/False Positives/Negatives, intersection over union (IoU) thresholds, and detection confidence scores. Thus, it is infeasible to construct the exact AP function in a few-shot setting for training. To solve this problem, we alternatively formulate AP_{pred} as a function Ψ of the task losses (Eq.4). Then the backward gradient of L_{AP} to the backbone parameters can be defined as:

$$\frac{\partial L_{AP}}{\partial \theta_b} = -\left(\frac{\partial \Psi}{\partial L_{cls}} \cdot \frac{\partial L_{cls}}{\partial \theta_b} + \frac{\partial \Psi}{\partial L_{reg}} \cdot \frac{\partial L_{reg}}{\partial \theta_b} \right). \tag{5}$$

From this formulation, we note that the gradients $\frac{\partial \Psi}{\partial L_{cls}}$ and $\frac{\partial \Psi}{\partial L_{reg}}$ are independent of the backbone parameters θ_b, such that we can construct a small network $\tilde{\Psi}$ to simulate Ψ. Data episodes (L_{cls}, L_{reg}, AP) can be sampled by fine-tuning the proposed model on the novel support dataset D_{novel} for several iterations, which can be utilized to train $\tilde{\Psi}$. The network details can be found in Sec.4.3. Then the partial gradients of L_{AP} in Eq.5 can be estimated by the gradients $\frac{\partial \tilde{\Psi}}{\partial L_{cls}}$ and $\frac{\partial \tilde{\Psi}}{\partial L_{reg}}$. With the consideration of AP maximization, the backward gradients of the double-head RCNN toward the backbone can be reformulated as:

$$\frac{\partial L_{rcnn}}{\partial \theta_b} = \gamma_{cls} \frac{\partial L_{cls}}{\partial \theta_b} + \gamma_{reg} \frac{\partial L_{reg}}{\partial \theta_b} + \gamma_{AP} \frac{\partial L_{AP}}{\partial \theta_b} = \alpha'_{cls} \frac{\partial L_{cls}}{\partial \theta_b} + \alpha'_{reg} \frac{\partial L_{reg}}{\partial \theta_b} \tag{6}$$

in which γ_{cls}, γ_{reg} and γ_{AP} are trade-off coefficients among different losses. The updated gradient scaling ratios α'_{cls} and α'_{reg} are defined as:

$$\alpha'_{cls} = \gamma_{cls} - \gamma_{AP} \frac{\partial \tilde{\Psi}}{\partial L_{cls}}, \alpha'_{reg} = \gamma_{reg} - \gamma_{AP} \frac{\partial \tilde{\Psi}}{\partial L_{reg}}. \tag{7}$$

The proposed precision-driven gradient balancer, denoted as α'_{cls} and α'_{reg}, has the capacity to dynamically balance the conflicting task gradients and adjust them according to different conditions of the losses, which can largely improve the FSOD performance (Tab. 1).

3.4 Task-Specific Learning with Pre-Trained CLIP

Enhanced single-task learning can boost MTL performance. The proposed method aims to fully exploit CLIP's task-specific learning capabilities. While CLIP's robust capabilities have been applied to downstream tasks, fully harnessing its potential for FSOD remains underexplored. Two key modules are introduced based on CLIP: knowledge distillation and detection score refinement.

Regarding knowledge distillation during the training process, CLIP serves as a teacher network $\Phi_{\mathcal{T}}$, while our proposed classification head takes on the role of a student network Φ_S (see Fig. 2 (a)). Given region proposals generated by RPN, the student network extracts ROI features from the backbone as x_S, while the teacher network utilizes the CLIP image encoder to obtain $x_{\mathcal{T}}$. The predicted classification logits $P_{\mathcal{T}}$ and P_S are respectively defined as

$$P_S = \Phi_S \left(x_S\right), \tag{8}$$

$$P_{\mathcal{T}} = \Phi_{\mathcal{T}} \left(x_{\mathcal{T}}, f_{sup}, V_{sup}, E_{sup}\right) = \eta \cdot \varrho \left(x_{\mathcal{T}} \cdot f_{sup}^T\right) V_{sup} + x_{\mathcal{T}} \cdot E_{sup}^T. \tag{9}$$

In Eq.9, $\Phi_{\mathcal{T}}$ is constructed as a key-value cache model based on the support set D_{novel}. f_{sup} represents the $(N+1)$-way-K-shot support RoI features, which also denotes the cache keys. V_{sup} is the one-hot vector generated from the support class labels in text. E_{sup} is the embeddings generated by CLIP's text encoder on the support class's textual labels. The function $\varrho(\cdot)$ is an attention model to assess the similarity between f_{sup} and $x_{\mathcal{T}}$, and η denotes the trade-off parameter between the two terms. The design of the teacher model draws inspiration from

[54] for few-shot classification. However, departure from the exclusive focus of [54] on foreground N-way-K-shot features, our teacher network also incorporates background shots as an additional class (the $(N+1)^{th}$ class). This adjustment ensures better alignment between the teacher network's logits and those of the student network, facilitating knowledge distillation. The background K shots are randomly cropped from the background regions of the support images, and these shots are then included in the support set.

Knowledge distillation aims at transferring the "dark knowledge" from teacher to student. To boost the performance of the transfer, a decoupled KD loss L_{KD} is formulated with target class knowledge distillation and non-target class knowledge distillation parts [58].

$$L_{\mathrm{KD}} = \alpha_k \left(\alpha_t \mathrm{KL} \left(P_{\mathcal{T}}^{tar} || P_S^{tar} \right) + \alpha_n \mathrm{KL} \left(P_{\mathcal{T}}^{non} || P_S^{non} \right) \right), \tag{10}$$

where KL denotes to KL-Divergence. P^{tar} represents the predicted logits for the target (ground-truth) class, while P^{non} refers to the non-target classes. Through knowledge distillation, the proposed model gains a substantial boost in its classification ability by learning from CLIP's expertise.

Regarding detection score refinement during inference, CLIP can be further utilized to obtain a more reliable detection score for the prediction (see Fig. 2 (b)). For a given query instance $q = (s, c, b)$ with predicted detection score s, class label c, and box coordinate b, which are obtained from our FSOD model, the RoI feature x_q is generated using CLIP's image encoder based on the predicted box. This x_q is then adapted into the adapted query feature $f_q = \Phi_T \left(x_q, f_{sup}, V_{sup}, E_{sup} \right)$, which benefits from cross-attention between the query and supports, and the adapted support features f_{sup} are enhanced with self-attention. The cross and self-attention layers play a crucial role in adapting both query and support features with the image-text correlation, thereby improving the representation ability of the adapted features towards novel classes. f_{sup} is further averaged along the K shots to generate the support prototypes $\bar{f}_{sup}$. With the adapted features, the detection score can be updated with the cosine similarity between the adapted query and support features :

$$s^* = \omega \cdot s + (1 - \omega) \frac{f_q \cdot \bar{f}_{sup_c}}{||f_q|| ||\bar{f}_{sup_c}||}, \tag{11}$$

where $\bar{f}_{sup_c}$ denotes the support prototype of the c^{th} class, and ω is a trade-off parameter. The refined detection scores leverage the discriminative capabilities of CLIP to mitigate false positives and enhance classification performance in few-shot scenarios.

4 Experiments

4.1 Datasets and Evaluation Settings

Our evaluation primarily centers on the detection performance of novel classes, utilizing the FSOD evaluation protocol that is commonly employed in the SOTA

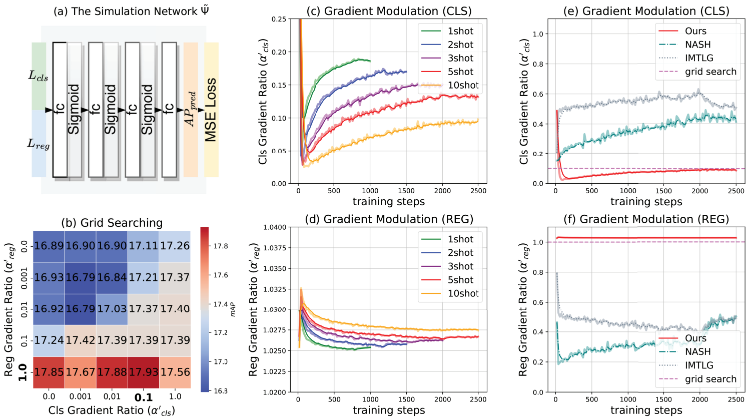

Fig. 3: (a) The architecture of the simulation network. (b) The results of grid searching on COCO-10shot with a batch size of 8. (c)-(d) The gradient-ratio changing curves during the fine-tuning stage. (e)-(f) The gradient-ratio changing curves compared with other gradient balancers on COCO 10-shot.

methods [1,32,47,49]. To ensure fair comparisons, all experiments are conducted on the two widely-used FSOD benchmarks, MS COCO [21] and PASCAL VOC [7], with the same settings and data split as previous works [32, 41]. The MS COCO dataset, consisting of 80 different object categories, is a widely used benchmark for FSOD. 60 base classes are employed for base-training, and 20 novel classes are used for fine-tuning. The PASCAL VOC dataset consists of 20 object categories. The base-training and fine-tuning stages are performed on three different base/novel splits. Each split contains 15 base classes and 5 novel classes. Following the previous work, we use the mean of the average precision (mAP) of the novel classes at varying intersections over union (IoU) thresholds (i.e., mAP, mAP_{50}, and mAP_{75}) as an evaluation metric for COCO (COCO style) and mAP_{50} at a fixed IoU threshold of 0.5 as an evaluation metric for VOC (VOC style). All our reported results are an average of multiple runs.

4.2 Implementation Details

The ResNet-101 architecture, pre-trained on the ImageNet dataset, is used as the backbone network in the proposed method. The SGD optimizer is used to train the network with a mini-batch size of 16, momentum of 0.9, and weight decay of $5e^{-5}$. Following FSCE [40], we double the maximum number (from 256 to 512) of proposals kept after NMS in order to increase positive samples for training. To ensure a fair comparison with existing methods, the proposed method initializes the fine-tuning model with the pre-trained model weights obtained from the base-training of DeFRCN. Specifically, during fine-tuning, the proposed method initializes the model weights of the Res5$_{cls}$ and Res5$_{reg}$ blocks with the same values obtained from the pre-trained Res5 model weights of DeFRCN over the

Table 2: Results on the VOC dataset over 3 splits with 1-, 2-, 3-, 5-, and 10-shot settings. The evaluation is conducted based on the performance of mAP_{50} in VOC style on novel classes. ⋆ indicates the results are reproduced by us using codebase shared by the authors. † indicates Faster R-CNN based models. The best results are bolded, and the second-best results are underlined.

VOC Split			Split 1					Split 2					Split 3				
Method	Venue	Category	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
FRCN-ft-full [36] †	NIPS15	meta	13.8	19.6	32.8	41.5	45.6	7.9	15.3	26.2	31.6	39.1	9.8	11.3	19.1	35.0	45.1
FSRW [15]	ICCV19	meta	14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9
MetaDet [42] †	ICCV19	meta	18.9	20.6	30.2	36.8	49.6	21.8	23.1	27.8	31.7	43.0	20.6	23.9	29.4	43.9	44.1
Meta R-CNN [49] †	ICCV19	meta	19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	45.4	14.3	18.2	27.5	41.2	48.1
CME [18]	AAAI19	meta	41.5	47.5	50.4	58.2	60.9	27.2	30.2	41.4	42.5	46.8	34.3	39.6	45.1	48.3	51.5
TIP [17] †	CVPR21	meta	27.7	36.5	43.3	50.2	59.6	22.7	30.1	33.8	40.9	46.9	21.7	30.6	38.1	44.5	50.9
MetaFasterRCNN [10] †	AAAI22	meta	43.0	54.5	60.6	66.1	65.4	27.7	35.5	46.1	47.7	51.4	40.6	46.4	53.4	59.9	58.6
KFSOD [55]	CVPR22	meta	44.6	-	54.4	60.9	65.8	37.8	-	43.1	48.1	50.4	34.8	-	44.1	52.7	53.9
Meta-DETR [53]	PAMI22	meta	40.6	51.4	58.0	59.2	63.6	37.0	36.6	43.7	49.1	54.6	41.6	45.9	52.7	58.9	60.6
ICPE [26] †	AAAI23	meta	54.3	59.5	62.4	65.7	66.2	33.5	40.1	48.7	51.7	52.5	50.9	53.1	55.3	60.6	60.1
LSTD [1]	AAAI18	transfer	8.2	1.0	12.4	29.1	38.5	11.4	3.8	5.0	15.7	31.0	12.6	8.5	15.0	27.3	36.3
NP-RepMet [50]	NIPS20	transfer	37.8	40.3	41.7	47.3	49.4	41.6	43.0	43.4	47.4	49.1	33.3.	38.0	39.8	41.5	44.8
MPSR [45] †	ECCV20	transfer	41.7	43.1	51.4	55.2	61.8	24.4	29.5	39.2	39.9	47.8	35.6	40.6	42.3	48.0	49.7
FSCE [40] †	CVPR21	transfer	44.2	43.8	51.4	61.9	63.4	27.3	29.5	43.5	44.2	50.2	37.2	41.9	47.5	54.6	58.5
TFA w/cos [41] †	CVPR21	transfer	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8
SRR-FSD [59] †	CVPR21	transfer	47.8	50.5	51.3	55.2	56.8	32.5	35.3	39.1	40.8	43.8	40.1	41.5	44.3	46.9	46.4
DeFRCN [32] ⋆ †	ICCV21	transfer	55.1	57.4	61.1	64.6	61.5	32.1	40.5	47.9	52.9	47.5	48.9	51.9	52.5	55.7	59.0
CoCo-RCNN [29]†	ECCV22	transfer	33.5	44.2	50.2	57.5	63.3	25.3	31.0	39.6	43.8	50.1	24.8	36.9	42.8	50.8	57.7
ModelCali [8]	ECCV22	transfer	40.1	44.2	51.2	62.0	63.0	33.3	33.1	42.3	46.3	52.3	36.1	43.1	43.5	52.0	56.0
DMNet [27]	TCyber23	transfer	34.7	50.7	54.0	58.8	62.5	31.3	28.2	41.8	46.2	52.7	38.6	40.0	43.4	48.9	48.9
D&R [19] †	AAAI23	transfer	60.4	64.0	65.2	64.7	66.3	37.9	46.8	48.1	52.7	53.1	55.7	57.9	57.6	60.6	61.9
NVAE [48] †	CVPR23	transfer	62.1	64.9	67.8	69.2	67.5	39.9	46.8	54.4	54.2	53.6	58.2	60.3	61.0	64.0	65.5
RISF [14] †	CVIU24	transfer	67.2	70.5	71.5	74.2	**73.9**	47.6	52.3	57.3	58.3	60.4	59.4	59.0	59.1	62.4	63.9
Ours	-	transfer	**68.9**	**71.5**	**72.1**	**74.5**	72.2	**65.5**	**69.8**	**73.5**	**74.4**	**73.1**	**68.8**	**69.8**	**70.0**	**71.6**	**71.9**

base dataset. As for fine-tuning, the learning rate is set to 0.01 for most of the experiments. Specifically, similar to Meta-DETR [53], the learning rates for the 5-shot and 10-shot PASCAL VOC cases are reduced to 0.001 to achieve better convergence. Following the same settings as DeFRCN, the Res5_{cls} block is frozen during fine-tuning, and the values of β_{rpn} and β_{rcnn} are set to 0 and 0.01, respectively. The positive fraction of random sampling in RPN is set to 0.5 for both the classification and localization tasks. CLIP utilizes the pre-trained ViT-L/14@336px model. The trade-off parameter η is set to 6 for COCO 1/2/3-shots and all VoC shots, and 1 for other COCO shots.

4.3 The Performance of Precision-Driven Gradient Balancer

A small network $\tilde{\Psi}$ (Fig. 3(a)) is built with an architecture of 4 linear layers and 3 Sigmoid activation layers in between. The input dimension of the network is 2 representing the two task losses, the output dimension is 1 denoting the AP value, and the dimension of the hidden layers is 8. Regarding the data episode collection, the proposed model is fine-tuned on the novel dataset D_{novel} for 200 iterations. Data episodes containing loss and average precision values are sampled every 10 iterations. In each iteration, RPN generates 200 proposals, creating a data episode for each. Sampling 20 iterations yields around 4k samples. The network undergoes training using normalized data episodes for 1000 iterations, employing a learning rate of $1e^{-3}$, the AdamW optimizer [25], and a batch size of

Table 3: Results on COCO for novel classes in 1-, 2-, 3-,5-, 10- and 30-shot settings. The evaluation is based on mAP , mAP_{50} and mAP_{75} in COCO style over multiple runs. ⋆ indicates the results are reproduced by us using the codebase shared by the authors. † indicates Faster R-CNN based models. The best results are bolded, and the second-best results are underlined.

COCO Shot				1	2	3	5	10			30		
Method	Venue	Category	Backbone	mAP	mAP	mAP	mAP	mAP	mAP_{50}	mAP_{75}	mAP	mAP_{50}	mAP_{75}
FRCN-ft-full [36] †	NIPS15	meta	ZFnet	1.7	3.1	3.7	4.6	5.5	10.0	5.5	7.4	13.1	7.4
FSRW [15]	ICCV19	meta	DarkNet-19	-	-	-	-	5.6	12.3	4.6	9.1	19.0	7.6
MetaDet [42] †	ICCV19	meta	ResNet-101	-	-	-	-	7.1	14.6	6.1	11.3	21.7	8.1
Meta R-CNN [49] †	ICCV19	meta	ResNet-101	-	-	-	-	8.7	19.1	6.6	8.7	19.1	6.6
CME [18]	AAAI19	meta	-	-	-	-	-	15.1	24.6	16.4	16.9	28.0	17.8
DCNet [12] †$	CVPR21	meta	ResNet-101	-	-	-	-	12.8	23.4	11.2	18.6	32.6	17.5
TIP [17] †	CVPR21	meta	ResNet-101	-	-	-	-	16.3	33.2	14.1	18.3	35.9	16.9
DAnA [2] †	TMM21	meta	ResNet-50	-	-	-	-	18.6	-	17.2	21.6	-	20.3
MetaFasterRCNN [10] †	AAAI22	meta	ResNet-101	5.1	7.6	-	-	12.7	25.7	10.8	16.6	31.8	15.8
FsDetView [47] †	PAMI22	meta	ResNet-18	4.5	6.6	7.2	10.7	12.5	27.3	9.8	14.7	30.6	12.2
Meta-DETR [53]	PAMI22	meta	ResNet-101	7.5	13.5	15.4	15.4	19.0	30.5	19.7	22.2	35.0	22.8
MPSR [45] † ▷	ECCV20	transfer	ResNet-101	5.1	6.7	7.4	-	9.8	17.9	9.7	14.1	25.4	14.2
TFA w/cos [41] †	CVPR21	transfer	ResNet-101	1.9	3.9	5.1	7.0	9.1	17.1	8.8	12.1	22.0	12.0
SRR-FSD [59] †	CVPR21	transfer	ResNet-101	-	-	-	-	11.3	23.0	9.8	14.7	29.2	13.5
DeFRCN [32] ⋆ †	ICCV21	transfer	ResNet-101	9.7	13.1	14.5	15.6	18.4	33.8	17.3	22.6	39.7	22.9
CoCo-RCNN [29]†	ECCV22	transfer	ResNet-101	5.2	-	-	-	16.4	26.5	16.5	19.2	32.9	21.0
DMNet [27]	TCyber23	transfer	ResNet-101	-	-	-	-	10.2	17.8	10.5	17.0	29.5	17.4
D&R [19] †	AAAI23	transfer	ResNet-101	8.3	12.7	14.3	16.4	18.7	-	-	21.8	-	-
ICPE [26] †	AAAI23	transfer	ResNet-101	-	-	-	-	19.3	27.9	18.0	23.1	32.9	19.2
NVAE [48] †	CVPR23	transfer	ResNet-101	9.5	13.7	14.3	15.9	18.7	-	-	22.5	-	-
RISF [14]	CVIU24	transfer	ResNet-101	11.7	15.9	**18.2**	**20.3**	21.9	39.9	-	24.4	43.2	-
Ours	-	transfer	ResNet-101	**12.8**	**16.9**	17.5	19.5	**22.7**	**40.0**	**22.3**	**25.2**	**43.3**	**25.3**

Table 4: The cross-domain FSOD performance on base-training with COCO base classes and fine-tuning on VOC novel classes.

Method	FRCN-ft-full [36]	FSRW [15]	MetaDet [42]	MetaRCNN [49]	MPSR [45]	DeFRCN [32]	Ours
mAP	31.2	32.3	33.9	37.4	42.3	41.0	**42.5**

8. The loss function employed is Mean Squared Error (MSE), which quantifies the difference between the sampled and predicted AP values. For the VOC dataset, mAP_{50} is utilized as the AP targets.

Tab. 1 shows a comparison between the SOTA MTL gradient modulation methods and the proposed precision-driven gradient balancer. All the models are finetuned on the COCO dataset in 10-shot with a batch size of 16. Through adaptively rescaling the backward gradients, the proposed balancer with the double-head RCNN boosts the average precision of FSOD by 11.3% compared with the baseline. In addition, during the fine-tuning, the proposed precision-driven gradient balancer can adaptively generate different ratios for different few-shot conditions (Fig. 3(c-d)). As an additional validation measure for the gradient balancer's performance, we conducted manual tuning of the gradient ratios via grid-searching (Fig. 3(b)). Surprisingly, the optimal ratios determined via grid-searching ($\alpha'_{cls} = 0.1, \alpha'_{reg} = 1.0$) closely resemble the choices made by the precision-driven gradient balancer (Fig. 3(e)-(f)). In contrast to grid-searching, which employs time-consuming brute-force searching, the gradient balancer provides a more time-efficient and parameter-efficient solution.

Table 5: The ablation studies on COCO-10shot. ◇ and ⋆ indicate the baseline and the enhanced baseline respectively.

Adaptive Multi-task Learning		Enhanced Task-specific Learning		mAP
Double-head RCNN	Precision-driven Gradient Balancer	Knowledge Distillation	Score Refinement	
-	-	-	-	16.7◇
-	-	-	PCB	18.5◇
-	-	-	✓	19.6
✓	-	-	-	17.4 ⋆
✓	✓	-	-	18.4
✓	✓	✓	-	18.6
✓	✓	✓	PCB	19.8
✓	✓	✓	✓	22.7

Table 6: Comparison of training and inference efficiency between the baseline method and our method using 4 RTX-6000 GPUs with a batch size of 4 per GPU.

Methods	Train (s/iter)	Inference (s/img)	Params (M)	mAP_{coco}
DeFRCN[32]	0.7	0.3	52.1	18.5
Ours	2.2	0.4	67.1	22.7

4.4 Comparison with State-of-the-Arts

The few-shot detection results on the PASCAL VOC dataset for all splits can be found in Tab. 2. Compared to the SOTA methods, our method achieves better or comparable results on novel classes. Compared with the second best, an improvement of 12.5% is observed in terms of the mean mAP_{50} across all configurations. Our method achieves more balanced precision scores on different VoC splits due to the adaptively learned gradient balancer and the KDSR scheme.

Table 3 shows the few-shot detection results for novel classes on the COCO datasets. Our method achieves either the highest or second-highest performance when compared to existing FSOD methods across all shots. The proposed method achieves an improvement of 3.7% over the second-best method in terms of mAP on COCO 10-shot. The proposed precision-driven gradient balancer, along with the KDSR scheme, has the capability to enhance performance in FSOD tasks. Compared to other methods, the proposed method retains the simplicity of the transfer-learning architecture while achieving outstanding performance under the more stringent mAP_{75} metric.

To evaluate the domain adaption ability of our method, the cross-dataset experiments are conducted following [32,45]. The results are presented in Tab. 4. The superior performance achieved by the proposed method indicates its strong generalization ability in few-shot cross-domain scenarios.

4.5 Ablation Studies & Parameter Analysis

The ablation study on our method is presented in Tab. 5. Unless otherwise specified, all experiments are performed on the COCO and VOC-split1 datasets in 10-shot with a batch size of 16. By gradually adding the proposed modules, the proposed method achieves increasing performance. The knowledge distillation module leads to a 1.1% improvement in the term of mAP via the learning from

Table 7: The parameter analysis of γ_{cls} and γ_{reg} for gradient balancer, and α_k for knowledge distillation.

α_k	0	0.05	**0.1**	0.2
COCO (mAP)	18.4	18.5	**18.6**	18.4

γ_{cls} & γ_{reg}	0	0.05	0.1	0.5	1
COCO (mAP)	-	-	-	0.5	**18.4**
VOC (mAP_{50})	62.8	**63.8**	61.4	63.1	62.2

Table 8: The parameter analysis of ω for score refinement.

ω	no-refine	0.2	0.4	0.6	0.8
COCO (mAP)	18.50	21.4	22.0	**22.7**	21.8

ω	no-refine	0.02	0.04	0.06	0.08
VOC (mAP_{50})	58.1	71.4	71.6	**71.8**	71.2

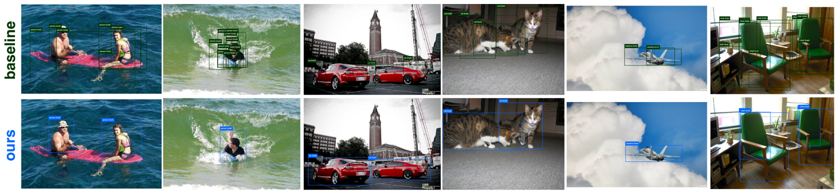

Fig. 4: Detection results visualization comparing the proposed method and baseline. Best viewed with zoom for optimal clarity.

the teacher network. With the score refinement module, the detection performance is increased by 4.1 in mAP. Due to the plug-and-play nature of the score refinement module, we incorporated it into the baseline method, resulting in a 5.9% performance boost for the baseline compared to PCB. The comparisons regarding training and inference efficiency have been depicted in Tab. 6.

The parameter analysis of γ_{cls} and γ_{reg} is given in Tab. 7. For the COCO dataset, employing too small values for γ_{cls} and γ_{reg} can lead to gradient vanishing, primarily influenced by the scale of the simulated gradients $\frac{\partial\tilde{\Psi}}{\partial L_{cls}}$ and $\frac{\partial\tilde{\Psi}}{\partial L_{reg}}$. Therefore, we set $\gamma_{cls} = \gamma_{reg} = 1.0$ for COCO. In the case of VOC, γ_{cls} and γ_{reg} are set to 0.05. The analysis of α_k is given in Tab. 8. For KD, the scaling ratio of the KD loss α_k is set to 0.1. α_t and α_n are set as 0.1 and 0.8. Moreover, the parameter analysis on the trade-off ω for score refinement can be found in Tab. 7. We further illustrate the detection results of both the proposed method and the baseline in Figure 4. Our method demonstrates superior IoU scores in terms of localization and higher detection scores in terms of classification.

5 Conclusion

This study proposes an adaptive multi-task learning method for few-shot object detection. The method constructs a novel precision-driven gradient balancer to mitigate task discrepancies in multi-task learning, allowing for adaptive adjustment of backward gradient ratios to achieve a balanced joint updating direction. Additionally, we introduce a CLIP-based knowledge distillation and score refinement scheme to enhance task-specific learning and further boost the performance of multi-task learning. Experimental results demonstrate that our method surpasses leading few-shot object detection methods on benchmark datasets.

Acknowledgements

This research is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

1. Chen, H., Wang, Y., Wang, G., Qiao, Y.: Lstd: A low-shot transfer detector for object detection. In: AAAI. vol. 32 (2018)
2. Chen, T.I., Liu, Y.C., Su, H.T., Chang, Y.C., Lin, Y.H., Yeh, J.F., Chen, W.C., Hsu, W.: Dual-awareness attention for few-shot object detection. IEEE TMM (2021)
3. Chen, Z., Badrinarayanan, V., Lee, C.Y., Rabinovich, A.: Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: International conference on machine learning. pp. 794–803. PMLR (2018)
4. Cheng, B., Wei, Y., Shi, H., Feris, R., Xiong, J., Huang, T.: Revisiting rcnn: On awakening the classification power of faster r-cnn. In: ECCV. pp. 453–468 (2018)
5. Clark, K., Luong, M.T., Khandelwal, U., Manning, C.D., Le, Q.V.: Bam! born-again multi-task networks for natural language understanding. arXiv preprint arXiv:1907.04829 (2019)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. Ieee (2009)
7. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV **88**, 303–308 (2009)
8. Fan, Q., Tang, C.K., Tai, Y.W.: Few-shot object detection with model calibration. In: ECCV. pp. 720–739. Springer (2022)
9. Girshick, R.: Fast r-cnn. In: ICCV. pp. 1440–1448 (2015)
10. Han, G., Huang, S., Ma, J., He, Y., Chang, S.F.: Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. In: AAAI. vol. 36, pp. 780–789 (2022)
11. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
12. Hu, H., Bai, S., Li, A., Cui, J., Wang, L.: Dense relation distillation with context-aware aggregation for few-shot object detection. In: CVPR. pp. 10185–10194 (2021)
13. Huang, G., Laradji, I., Vazquez, D., Lacoste-Julien, S., Rodriguez, P.: A survey of self-supervised and few-shot object detection. IEEE TPAMI (2022)
14. Jung, M.J., Han, S.D., Kim, J.: Re-scoring using image-language similarity for few-shot object detection. Computer Vision and Image Understanding p. 103956 (2024)
15. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting. In: ICCV. pp. 8420–8429 (2019)
16. Kim, J.U., Kim, S.T., Kim, E.S., Moon, S.K., Ro, Y.M.: Towards high-performance object detection: Task-specific design considering classification and localization separation. In: ICASSP. pp. 4317–4321. IEEE (2020)
17. Li, A., Li, Z.: Transformation invariant few-shot object detection. In: CVPR. pp. 3094–3102 (2021)

18. Li, B., Yang, B., Liu, C., Liu, F., Ji, R., Ye, Q.: Beyond max-margin: Class margin equilibrium for few-shot object detection. In: CVPR. pp. 7363–7372 (2021)
19. Li, J., Zhang, Y., Qiang, W., Si, L., Jiao, C., Hu, X., Zheng, C., Sun, F.: Disentangle and remerge: interventional knowledge distillation for few-shot object detection from a conditional causal perspective. In: AAAI. vol. 37, pp. 1323–1333 (2023)
20. Li, Y., Zhu, H., Cheng, Y., Wang, W., Teo, C.S., Xiang, C., Vadakkepat, P., Lee, T.H.: Few-shot object detection via classification refinement and distractor retreatment. In: CVPR. pp. 15395–15403 (2021)
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014)
22. Liu, B., Liu, X., Jin, X., Stone, P., Liu, Q.: Conflict-averse gradient descent for multi-task learning. NeurIPS **34**, 18878–18890 (2021)
23. Liu, L., Li, Y., Kuang, Z., Xue, J.H., Chen, Y., Yang, W., Liao, Q., Zhang, W.: Towards impartial multi-task learning. In: ICLR (2020)
24. Liu, S., Liang, Y., Gitter, A.: Loss-balanced task weighting to reduce negative transfer in multi-task learning. In: AAAI. vol. 33, pp. 9977–9978 (2019)
25. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
26. Lu, X., Diao, W., Mao, Y., Li, J., Wang, P., Sun, X., Fu, K.: Breaking immutable: Information-coupled prototype elaboration for few-shot object detection. In: AAAI. vol. 37, pp. 1844–1852 (2023)
27. Lu, Y., Chen, X., Wu, Z., Yu, J.: Decoupled metric network for single-stage few-shot object detection. IEEE Transactions on Cybernetics **53**(1), 514–525 (2022)
28. Ma, J., Zhao, Z., Yi, X., Chen, J., Hong, L., Chi, E.H.: Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 1930–1939 (2018)
29. Ma, J., Han, G., Huang, S., Yang, Y., Chang, S.F.: Few-shot end-to-end object detection via constantly concentrated encoding across heads. In: ECCV. pp. 57–73. Springer (2022)
30. Navon, A., Shamsian, A., Achituve, I., Maron, H., Kawaguchi, K., Chechik, G., Fetaya, E.: Multi-task learning as a bargaining game. arXiv preprint arXiv:2202.01017 (2022)
31. Oksuz, K., Cam, B.C., Kalkan, S., Akbas, E.: Imbalance problems in object detection: A review. IEEE TPAMI **43**(10), 3388–3415 (2020)
32. Qiao, L., Zhao, Y., Li, Z., Qiu, X., Wu, J., Zhang, C.: Defrcn: Decoupled faster r-cnn for few-shot object detection. In: ICCV. pp. 8681–8690 (2021)
33. Quan, Q., Yao, Q., Li, J., Zhou, S.K.: Which images to label for few-shot medical landmark detection? In: CVPR. pp. 20606–20616 (2022)
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
35. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: CVPR. pp. 7263–7271 (2017)
36. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. NeurIPS **28** (2015)
37. Rusu, A.A., Colmenarejo, S.G., Gulcehre, C., Desjardins, G., Kirkpatrick, J., Pascanu, R., Mnih, V., Kavukcuoglu, K., Hadsell, R.: Policy distillation. arXiv preprint arXiv:1511.06295 (2015)

38. Song, G., Liu, Y., Wang, X.: Revisiting the sibling head in object detector. In: CVPR. pp. 11563–11572 (2020)
39. Song, H., Dong, L., Zhang, W.N., Liu, T., Wei, F.: Clip models are few-shot learners: Empirical studies on vqa and visual entailment. arXiv preprint arXiv:2203.07190 (2022)
40. Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C.: Fsce: Few-shot object detection via contrastive proposal encoding. In: CVPR. pp. 7352–7362 (2021)
41. Wang, X., Huang, T., Gonzalez, J., Darrell, T., Yu, F.: Frustratingly simple few-shot object detection. In: International Conference on Machine Learning. pp. 9919–9928. PMLR (2020)
42. Wang, Y.X., Ramanan, D., Hebert, M.: Meta-learning to detect rare objects. In: ICCV. pp. 9925–9934 (2019)
43. Wang, Z., Tsvetkov, Y., Firat, O., Cao, Y.: Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. arXiv preprint arXiv:2010.05874 (2020)
44. Wertheimer, D., Hariharan, B.: Few-shot learning with localization in realistic settings. In: CVPR. pp. 6558–6567 (2019)
45. Wu, J., Liu, S., Huang, D., Wang, Y.: Multi-scale positive sample refinement for few-shot object detection. In: ECCV. pp. 456–472. Springer (2020)
46. Wu, Y., Chen, Y., Yuan, L., Liu, Z., Wang, L., Li, H., Fu, Y.: Rethinking classification and localization for object detection. In: CVPR. pp. 10186–10195 (2020)
47. Xiao, Y., Lepetit, V., Marlet, R.: Few-shot object detection and viewpoint estimation for objects in the wild. IEEE TPAMI **45**(3), 3090–3106 (2022)
48. Xu, J., Le, H., Samaras, D.: Generating features with increased crop-related diversity for few-shot object detection. In: CVPR. pp. 19713–19722 (2023)
49. Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., Lin, L.: Meta r-cnn: Towards general solver for instance-level low-shot learning. In: ICCV. pp. 9577–9586 (2019)
50. Yang, Y., Wei, F., Shi, M., Li, G.: Restoring negative information in few-shot object detection. NeurIPS **33**, 3521–3532 (2020)
51. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. NeurIPS **33**, 5824–5836 (2020)
52. Zhang, G., Cui, K., Wu, R., Lu, S., Tian, Y.: Pnpdet: Efficient few-shot detection without forgetting via plug-and-play sub-networks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3823–3832 (2021)
53. Zhang, G., Luo, Z., Cui, K., Lu, S., Xing, E.P.: Meta-detr: Image-level few-shot detection with inter-class correlation exploitation. IEEE TPAMI (2022)
54. Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free adaption of clip for few-shot classification. In: ECCV. pp. 493–510. Springer (2022)
55. Zhang, S., Wang, L., Murray, N., Koniusz, P.: Kernelized few-shot object detection with efficient integral aggregation. In: CVPR. pp. 19207–19216 (2022)
56. Zhang, Y., Yang, Q.: A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering **34**(12), 5586–5609 (2021)
57. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: ECCV. pp. 94–108. Springer (2014)
58. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: CVPR. pp. 11953–11962 (2022)
59. Zhu, C., Chen, F., Ahmed, U., Shen, Z., Savvides, M.: Semantic relation reasoning for shot-stable few-shot object detection. In: CVPR. pp. 8782–8791 (2021)

60. Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: A survey. Proceedings of the IEEE (2023)