scientific reports

OPEN

Instance mask alignment for object detection knowledge distillation

Zhen Guo[1,2], Pengzhou Zhang[1] & Peng Liang[2]

Knowledge distillation has proven to be an effective technique for enhancing object detection performance. However, the presence of different detector types often results in a significant performance gap between teacher and student models. In this paper, we propose an Instance Mask Alignment (IMA) knowledge distillation framework for object detection. Our framework leverages knowledge transformation operations to reduce the teacher-student gap, leading to notable performance improvements. We introduce instance mask distillation, which incorporates mask information to enhance the student model's ability to identify and focus on relevant regions or objects. Additionally, we introduce a cascade alignment module with instance standardization, utilizing an adaptive scale deflation module along the instance dimension. Through the integration of these cascade knowledge alignment modules, our proposed framework achieves substantial performance gains across various detector types. Extensive experiments conducted on the MS-COCO, PASCAL VOC and Cityscapes benchmarks demonstrate the effectiveness of our novel method, particularly its adaptability to heterogeneous detectors.

Keywords Knowledge distillation, Object detection, Instance mask, Feature alignment

Object detection as a fundamental task in computer vision[1], has been widely used in numerous applications such as smart factory, autonomous vehicles, surveillance systems, and medical scenarios. While deep learning has significantly advanced the performance of object detectors, these models often require substantial computational resources, limiting their deployment in resource-constrained environments[2]. Knowledge distillation[2,3], a technique that transfers knowledge from a larger teacher model to a smaller student model, has emerged as a promising solution to address this challenge. By leveraging the knowledge of the teacher model, the student model can learn from the teacher's representations and predictions, achieving comparable performance while being more computationally efficient.

Knowledge distillation has been extensively studied in the field of image classification[4]. However, applying knowledge distillation to the task of object detection presents unique challenges. Object detection involves not only the classification of objects but also their precise localization within the image. In addition, there exist various types of detectors, each with its own characteristics and response patterns. However, the presence of diverse detector architectures, such as two-stage, one-stage, and anchor-free detectors, poses significant challenges in the knowledge transfer process. These detectors often exhibit structural differences in their output representations, leading to a substantial gap between the teacher and student models, hindering the effective distillation of knowledge.

Existing knowledge distillation methods for object detection primarily focus on aligning the feature representations or output predictions between the teacher and student models[5]. However, these approaches often overlook the importance of instance-level information, such as object masks, which can provide valuable guidance for the student model to better identify and attend to relevant regions or objects within the input data. Moreover, the inherent differences between detector architectures, such as anchor-based versus anchor-free designs, can further exacerbate the teacher-student gap, necessitating additional alignment strategies. Effectively bridging this gap is crucial for achieving successful knowledge distillation and ensuring that the student model can learn from the teacher's expertise while maintaining high accuracy.

To address this issue, we propose the Instance Mask Alignment (IMA) framework for object detection knowledge distillation. In our proposed IMA framework, we introduce instance mask distillation, which incorporates mask information to improve the student model's ability to identify and attend to relevant regions or objects within the input data. Specifically, we distill the knowledge from the teacher's instance masks to the student model, encouraging the student to learn to predict accurate object masks during training. This instance-level guidance helps the student model better understand the spatial extent and boundaries of objects, leading

[1]State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China. [2]China Unicom Smart City Research Institute, Beijing 100048, China. ✉email: cathy.guozhen@cuc.edu.cn; zhangpengzhou@cuc.edu.cn

to improved detection performance. Furthermore, we introduce a cascade alignment module that consists of instance standardization and an adaptive scale deflation module in the instance dimension. The instance standardization step normalizes the instance-level features, thereby reducing the internal covariate shift and improving the training stability. The adaptive scale deflation module then adaptively scales the instance-level features based on their importance, allowing the student model to focus on the most relevant instances and mitigate the impact of irrelevant or noisy instances. By cascading these alignment modules, our framework effectively bridges the gap between different detector architectures, enabling successful knowledge transfer and performance improvements. This comprehensive approach enhances the student model's ability to predict image recognition and object detection tasks accurately.

We conduct extensive experiments on popular object detection benchmarks, including MS-COCO, PASCAL VOC, and extend experiments on the instance segmentation dataset Cityscapes. The results demonstrate that our IMA approach enhances the performance of student models while maintaining computational efficiency. The proposed IMA method, shown as Fig. 1 represents a promising solution for deploying efficient object detection models in resource-constrained scenarios.

In summary, our three key contributions are:

- Our paper presents the Instance Mask Alignment (IMA) framework for object detection knowledge distillation, which reduces the performance gap between teacher and student detectors.
- Our approach leverages instance mask information to enhance object detection performance. The cascade alignment module aligns feature representations between the teacher and student models, thereby reducing the performance gap across different detector types.
- Through extensive experiments conducted on multiple benchmarks, we demonstrate the effectiveness and adaptability of our proposed IMA framework in object detection knowledge distillation, with a thorough analysis of its strengths and limitations.

In accordance with the provided instructions, we proceed to present the theoretical foundations and methodology in the subsequent section.

Methods

In this section, we provide a detailed description of our proposed Instance Mask Alignment (IMA) framework for object detection knowledge distillation. We begin by revisiting the conventional knowledge distillation approach for object detection. We then introduce our proposed IMA framework, detailing its key components: Instance Mask Distillation and the Cascade Alignment Module.

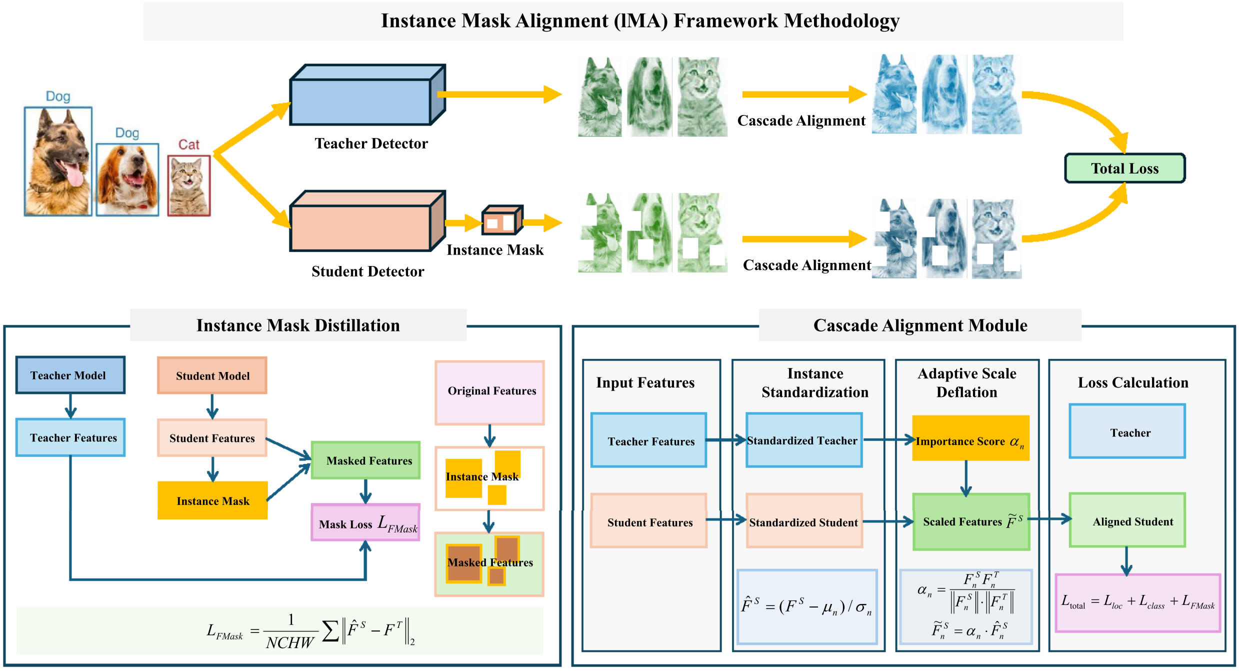

Fig. 1. Overview of the Instance Mask Alignment (IMA) framework for object detection knowledge distillation. Our distillation method follows a two-step process. Firstly, we extract instance feature maps from the teacher model. These feature maps are then used to generate an instance mask that aligns with the student features. To further enhance the alignment between the teacher and student features, we employ the Cascade Alignment technique, which includes Instance Standardization and Adaptive Scale Deflation. Finally, we calculate the total distillation loss.

Revisiting object detection knowledge distillation

We first revisit the general formulation of conventional detection knowledge distillation methods for a better understanding of our approach. Current feature distillation approaches encourage the student model *S* to mimic the intermediate features of the teacher model *T* by explicitly optimizing the feature distillation loss. Let $F^S \in \mathbb{R}^{N\times C\times H\times W}$ and $F^T \in \mathbb{R}^{N\times C\times H\times W}$ denote the middle-level features of the student and teacher models, respectively, where *N* is the number of instances, *C* is the number of channels, and *H* and *W* are the spatial dimensions. The purpose of conventional feature distillation is to minimize the feature distillation loss, which is described as follows:

$$L_{KD} = \frac{1}{NCHW} \sum_{n=1}^{N} \sum_{k=1}^{C} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathcal{D}_f \left(F^T_{n,k,i,j} - f_{align}(F^S_{n,k,i,j}) \right) \tag{1}$$

where $\mathcal{D}_f(\cdot)$ is the distance function measuring the difference between the intermediate features of the teacher and student models. The adaptation layer f_{align} is used to align the student's features F^S with the teacher's features F^T. In conventional feature distillation, the goal is to minimize the discrepancy between the student's and teacher's feature maps, thereby encouraging the student model to learn a similar intermediate representation as the teacher model.

Instance mask distillation

In our proposed method, we introduce an Instance Mask Distillation module to effectively transfer instance-level spatial information from the teacher to the student model. This module leverages instance masks to guide the student model in identifying and attending to relevant regions or objects within the input data, thereby enhancing detection performance.

Specifically, our instance mask distillation module utilizes a binary mask $M \in \mathbb{R}^{H\times W\times 1}$, which is generated randomly with a mask ratio $\zeta \in [0, 1)$, as defined below:

$$M_{i,j} = \begin{cases} 0, & R_{i,j} < \zeta \\ 1, & R_{i,j} \geq \zeta \end{cases} \tag{2}$$

Here, $R_{i,j}$ represents a random value sampled from a uniform distribution, denoted by $\mathcal{U}(0, 1)$, for each spatial location (*i*, *j*). The operation of element-wise multiplication, applied to the function $\mathcal{F}(F^S)$, effectively masks out certain frequency components.

$$\hat{F}^S = \mathcal{F}^{-1}(M \odot \mathcal{F}(F^S)) \tag{3}$$

The knowledge distillation loss term, L_{FMask}, is computed as the mean squared error between the masked student features, $\hat{F}^S$, and the teacher features, F^T, across all instances, spatial locations, and channels:

$$L_{FMask} = \frac{1}{NCHW} \sum_{n=1}^{N} \sum_{k=1}^{C} \sum_{i=1}^{H} \sum_{j=1}^{W} \left\| \hat{F}^S_{n,k,i,j} - F^T_{n,k,i,j} \right\|_2 \tag{4}$$

By minimizing this loss during training, we encourage the student model to learn to generate feature maps that are consistent with the teacher's feature maps in the masked regions. The guidance provided by the teacher's instance masks enables the student model to gain a more comprehensive understanding of the spatial extent and boundaries of objects, which in turn leads to enhanced detection performance.

The instance mask distillation module offers several advantages. Firstly, it provides a direct way to transfer instance-level spatial information from the teacher to the student, which is particularly beneficial for object detection tasks where accurate localization and segmentation are crucial. Secondly, by randomly masking different regions of the input during training, the student model is exposed to a diverse set of masked inputs, thereby promoting robustness and generalization. Finally, the instance mask distillation module can be easily integrated into existing knowledge distillation frameworks, in conjunction with other techniques such as feature mimicking or output distribution alignment.

Cascade alignment module

In addition to the Instance Mask Distillation module, we introduce a Cascade Alignment Module to further bridge the gap between different detector architectures and enable successful knowledge transfer. This module consists of two principal components: Instance Standardization and Adaptive Scale Deflation modules.

Instance standardization

The Instance Standardization step is designed to normalize the instance-level features, reducing the internal covariate shift and improving the training stability of the student model. This is particularly important when working with different detector architectures, as the feature distributions can vary significantly, thereby hindering the knowledge transfer process.

Specifically, we compute the mean $\mu_n \in \mathbb{R}^C$ and standard deviation $\sigma_n \in \mathbb{R}^C$ of the instance-level features across the spatial dimensions for each instance *n* as follows:

$$\mu_n = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} F^S_{n,c,i,j} \tag{5}$$

$$\sigma_n = \sqrt{\frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(F^S_{n,c,i,j} - \mu_n\right)^2} \tag{6}$$

We then normalize the instance-level features by subtracting the mean and dividing by the standard deviation, resulting in the standardized features, denoted by $\hat{F}^S$:

$$\hat{F}^S = \frac{F^S - \mu_n}{\sigma_n} \tag{7}$$

The Instance Standardization step serves to reduce the internal covariate shift, improving the training stability and enabling more effective knowledge transfer between different detector architectures.

Adaptive scale deflation

Following the Instance Standardization step, we introduce an Adaptive Scale Deflation module that adaptively scales the instance-level features based on their relative importance. This module enables the student model to concentrate on the most relevant instances and to mitigate the impact of irrelevant or noisy instances, thereby further enhancing the knowledge transfer process.

We compute an importance score $\alpha_n \in [0, 1]$ for each instance *n* based on the similarity between the student's and teacher's feature maps. Specifically, we compute the cosine similarity between the flattened student and teacher feature maps for each instance:

$$\alpha_n = \frac{F^S_n \cdot F^T_n}{\|F^S_n\| \, \|F^T_n\|} \tag{8}$$

Here, F^S_n and F^T_n represent the flattened student and teacher feature maps for instance *n*, respectively. We then apply a scaling factor γ_n to the standardized instance-level features $\hat{F}^S_n$, in accordance with the importance score, represented by α_n:

$$\gamma_n = \alpha_n \tag{9}$$

$$\tilde{F}^S_n = \gamma_n \hat{F}^S_n \tag{10}$$

By cascading the Instance Standardization and Adaptive Scale Deflation modules, our framework effectively aligns the instance-level features between the student and teacher models, thus bridging the gap between different detector architectures. This cascade of alignment operations enables successful knowledge transfer and performance improvements, as the student model can learn from the most relevant instances while mitigating the impact of irrelevant or noisy instances.

The Cascade Alignment Module offers several strengths. Firstly, the Instance Standardization step reduces the internal covariate shift, improving training stability and enabling more effective knowledge transfer between different architectures. Secondly, the Adaptive Scale Deflation module allows the student model to focus on the most relevant instances, further enhancing the knowledge transfer process and mitigating the impact of irrelevant or noisy instances. Finally, by cascading these two components, our framework can effectively bridge the gap between different detector architectures, enabling successful knowledge transfer and performance improvements.

Total optimization and inference

During training, the object detection loss and the distillation losses introduced by the proposed modules are jointly optimized. The total loss function is defined as:

$$L_{total} = L_{loc} + L_{cls} + L_{FMask} \tag{11}$$

where L_{loc} and L_{cls} are the localization loss and classification loss respectively for object detection, and L_{FMask} represents the loss from our Instance Mask Distillation module.

The localization loss L_{loc} measures the difference between the predicted bounding box for an object and the ground truth bounding box, typically using a smooth $L1$ loss:

$$L_{loc} = \sum_{i=1}^{N_{pos}} \text{smooth}_{L1} \left(b_i^{pred} - b_i^{gt} \right) \tag{12}$$

Here, N_{pos} represents the number of positive samples, b_i^{pred} denotes the predicted bounding box, and b_i^{gt} signifies the ground truth bounding box.

The classification loss L_{cls} measures the difference between the predicted class probabilities for an object and the ground truth class probabilities, typically using a cross-entropy loss:

$$L_{cls} = -\sum_{i=1}^{N_{pos}} \sum_{j=1}^{C} y_{ij} \log \left(p_{ij} \right) \tag{13}$$

Here, C is the number of classes, y_{ij} represents the ground truth label (0 or 1), indicating whether instance *i* belongs to class *j*, and p_{ij} denotes the predicted probability that instance *i* belongs to class *j*.

During inference, the student model generates object classification and location information using the features extracted from the input image, without relying on the teacher model or the distillation losses.

Theoretical foundation of instance mask alignment

The theoretical underpinning of our Instance Mask Alignment (IMA) framework is based on the observation that conventional knowledge distillation methods often struggle with the structural differences between different detector architectures. This is particularly evident when the teacher and student models employ different detection paradigms (e.g., two-stage vs. one-stage, or anchor-based vs. anchor-free).

From an information theory perspective, we argue that the teacher-student knowledge transfer process can be optimized by focusing on the most informative regions within the feature maps, namely the instance regions. By emphasizing these regions during the distillation process, we can ensure that the knowledge transferred from the teacher to the student is most relevant for the detection task.

Moreover, we observe that the feature distributions of different detector architectures can vary significantly, even when they are trained on the same dataset and for the same task. This distribution shift can hinder the knowledge transfer process, as the student model may struggle to mimic the teacher's feature representations. To address this issue, we introduce the concept of feature distribution alignment through instance standardization and adaptive scaling.

Formally, let $p_T(F)$ and $p_S(F)$ denote the probability distributions of the teacher's and student's feature maps, respectively. The goal of feature distribution alignment is to minimize the divergence between these distributions:

$$\min_S D(p_T(F)||p_S(F)) \tag{14}$$

where $D(\cdot||\cdot)$ is a divergence measure (e.g., KL divergence). However, directly minimizing this divergence is challenging due to the structural differences between the teacher and student models. Instead, we propose to align the distributions after applying a transformation $g(\cdot)$ to the feature maps:

$$\min_S D(p_T(g(F))||p_S(g(F))) \tag{15}$$

In our IMA framework, $g(\cdot)$ corresponds to the cascade of instance standardization and adaptive scale deflation, which we describe in detail in the above sections.

Pseudo-code for the IMA algorithm

To provide a clear understanding of our proposed IMA framework, we present a pseudo-code description of the main algorithm in Algorithm 1.

```
Require: Teacher model $T$, Student model $S$, Training dataset $\mathscr{D}$, Mask ratio $\zeta$
Ensure: Trained student model $S$
for each batch $(X,Y)$ in $\mathscr{D}$ do
    // Forward pass through teacher and student models
    $F^T \leftarrow \text{FeatureExtractor}(T,X)$
    $F^S \leftarrow \text{FeatureExtractor}(S,X)$
    // Instance Mask Distillation
    Generate random mask $M$ with mask ratio $\zeta$
    $\hat{F}^S \leftarrow \mathscr{F}^{-1}(M \odot \mathscr{F}(F^S))$
    Compute $L_{FMask}$ using Equation (6)
    // Cascade Alignment Module
    for each instance $n$ do
        // Instance Standardization
        Compute $\mu_n$ and $\sigma_n$ using Equations (7) and (8)
        $\hat{F}_n^S \leftarrow \frac{F_n^S - \mu_n}{\sigma_n}$
        // Adaptive Scale Deflation
        Compute $\alpha_n$ using Equation (10)
        $\gamma_n \leftarrow \alpha_n$
        $\tilde{F}_n^S \leftarrow \gamma_n \hat{F}_n^S$
    end for
    // Compute detection losses
    Compute $L_{loc}$ and $L_{cls}$ using Equations (13) and (14)
    // Compute total loss and update student model
    $L_{total} \leftarrow L_{loc} + L_{cls} + L_{FMask}$
    Update student model $S$ by minimizing $L_{total}$
end for
return Trained student model $S$
```

Algorithm 1. Instance Mask Alignment (IMA) for Object Detection Knowledge Distillation

	Method	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Teacher	FCOS-Res101	40.8	60.0	44.0	24.2	44.3	52.4
Student	FCOS-Res50	38.5	57.7	41.0	21.9	42.8	48.6
	GID[8]	42.0	60.4	45.5	25.6	45.8	54.2
	FRS[9]	40.9	60.3	43.6	25.7	45.2	51.2
	FGD[10]	42.1	–	–	**27.0**	**46.0**	54.6
	IMA (Ours)	**42.4**	**61.0**	**45.8**	26.6	45.9	**54.8**
Teacher	Faster RCNN-Res101	39.8	60.1	43.3	22.5	43.6	52.8
Student	Faster RCNN-Res50	38.4	59.0	42.0	21.5	42.1	50.3
	KD-Zero[11]	38.4	59.4	41.7	22.7	41.8	45.9
	FitNet[12]	38.8	59.6	41.8	22.3	42.2	50.7
	FGFI[13]	39.4	60.3	43.0	22.9	42.5	52.0
	FGD[10]	40.4	–	–	22.8	44.5	53.5
	IMA (Ours)	**40.6**	**60.9**	**43.9**	**23.0**	**44.5**	**54.0**
Teacher	RetinaNet101-Res101	38.9	58.0	41.5	21.0	42.8	52.4
Student	RetinaNet50-Res50	37.4	56.7	39.6	20.0	40.7	49.7
	KD-Zero[11]	36.8	56.6	39.4	21.9	40.6	48.2
	FitNet[12]	36.3	56.0	39.0	20.1	40.3	47.1
	FGFI[13]	37.3	57.1	40.0	21.0	41.5	49.7
	FGD[10]	39.6	–	–	**22.9**	**44.3**	**53.4**
	IMA (Ours)	**39.7**	**58.6**	**41.4**	22.7	42.9	51.3

Table 1. Main results on object detection. We use AP on different settings to evaluate results. Res101, Res50 represents using ResNet101 and ResNet50 as backbones. Significant values are in bold.

	Method	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Teacher	RetinaNet-ResNeXt101	41.6	61.4	44.3	23.9	45.5	54.5
Student	RetinaNet-Res50	37.4	56.7	39.6	20.0	40.7	49.7
	FGFI[13]	39.1	59.8	42.8	22.2	42.9	51.1
	COFD[14]	38.9	60.1	42.6	21.8	42.7	50.7
	FKD[15]	39.6	58.8	42.1	22.7	43.3	52.5
	FGD[10]	40.4	–	–	**23.4**	44.7	54.1
	IMA (Ours)	**41.0**	**60.2**	**43.6**	23.0	**45.2**	**55.0**
Teacher	Cascade Mask RCNN-ResNeXt101	45.6	64.1	49.7	26.2	49.6	60.0
Student	Faster RCNN-Res50	38.4	59.0	42.0	21.5	42.1	50.3
	FKD[15]	41.5	62.2	45.1	23.5	45.0	55.3
	IMA (Ours)	**41.6**	**62.3**	**45.5**	**23.5**	**45.3**	**55.3**
Teacher	RepPoints-ResNeXt101	44.2	65.5	47.8	26.2	48.4	58.5
Student	RepPoints-Res50	38.6	59.6	41.6	22.5	42.2	50.4
	FKD[15]	40.6	61.7	43.8	23.4	44.6	53.0
	FGD[10]	41.3	–	–	**24.5**	45.2	54.0
	IMA (Ours)	**42.3**	**63.1**	**45.8**	24.1	**46.4**	**55.9**

Table 2. More results on different backbone object detectors. Significant values are in bold.

The pseudo-code provides a step-by-step description of our IMA framework, including the Instance Mask Distillation module and the Cascade Alignment Module. The algorithm begins by extracting features from both the teacher and student models. It then applies the Instance Mask Distillation module, which generates a random mask and applies it to the student's features. Next, the Cascade Alignment Module is applied to each instance, which involves Instance Standardization and Adaptive Scale Deflation. Finally, the detection losses and the total loss are computed, and the student model is updated by minimizing the total loss.

Results

To evaluate the effectiveness of our proposed Instance Mask Alignment framework, we have conducted a series of comprehensive experiments across a range of object detectors, including two-stage, one-stage, and anchor-free architectures. We compare our method with state-of-the-art approaches and demonstrate superior performance on multiple evaluation metrics. Furthermore, we present experiments involving teacher models with heterogeneous backbones to demonstrate the versatility of our approach. Finally, we provide detailed ablation studies to validate the efficacy of our proposed techniques.

Datasets and Implementation Details Our experiments are performed on two widely-adopted object detection benchmarks: MS COCO and PASCAL VOC. The MS COCO dataset comprises 80 object categories with over 330,000 images, containing diverse object scales and challenging backgrounds. The PASCAL VOC dataset consists of 20 object categories with approximately 11,000 images. We evaluate the performance of the object detectors using standard metrics, such as mean Average Precision (mAP). For model optimization, we employ techniques like stochastic gradient descent (SGD) or Adam. The hyperparameters are set to $\alpha = 10$, and we use $L2$-loss for the function $\mathcal{D}_f(\cdot)$ across all experiments. We adopt a 2 x learning rate schedule and train for 24 epochs on the COCO dataset during the distillation process.

Main Results Table 1 presents the experimental results, comparing the baseline detectors with our distillation approach. It is evident that student detectors achieve superior performance when distilled from stronger teacher detectors based on more powerful backbones. IMA consistently achieves superior performance compared to both the baseline student models and other distillation methods. On the FCOS[6] detector, IMA attains the highest mAP of 42.4, outperforming GID, FRS, and FGD, and showing substantial improvements across most AP metrics, including AP_{50} and AP_{75}. For the Faster R-CNN detector, IMA achieves an mAP of 40.6, exceeding the best baseline (FGD) by 0.2 points, and delivering top performance on all detailed AP metrics, including AP_S, AP_M, and AP_L. Similarly, on RetinaNet[7], IMA yields the highest mAP of 39.7, outperforming FGD and other baselines, while maintaining competitive results across object scales. These results validate the generality and robustness of IMA across different detection architectures and demonstrate its effectiveness in improving student model performance through knowledge distillation.

Different Backbone Distillation Our approach is adaptable to distillation between heterogeneous backbones, enabling knowledge transfer from teachers with different architectures. we conduct experiments on various teacher-student detector pairs with different backbone architectures and leverage teacher detectors based on stronger backbones.Table 2 compares our results with other effective distillation techniques. IMA consistently outperforms existing state-of-the-art knowledge distillation methods, including FGFI, COFD, FKD, and FGD, across multiple detection frameworks. On the RetinaNet detector, where the teacher uses ResNeXt-101 and the student uses ResNet-50, IMA achieves the highest mAP of 41.0, significantly surpassing the baseline student model (37.4) and outperforming strong methods such as FGD (40.4) and FKD (39.6). IMA also delivers the best results across most sub-metrics, including AP_{50}, AP_{75}, AP_M, and AP_L, demonstrating its ability to effectively transfer both coarse and fine-grained knowledge from a stronger backbone. In the case of Cascade Mask R-CNN (teacher) to Faster R-CNN (student), IMA achieves an mAP of 41.6, slightly exceeding FKD (41.5) and

Method	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
GFL-Res101 (T)	44.9	63.1	49.0	28.0	49.1	57.2
GFL-Res50 (S)	40.2	58.4	43.3	23.3	44.0	52.2
FitNets[12]	40.7	58.6	44.0	23.7	44.4	53.2
Inside GT Box[17]	40.7	58.6	44.2	23.1	44.5	53.5
Defeat[18]	40.8	58.6	44.2	24.3	44.6	53.7
LD[17]	41.0	58.6	44.2	23.4	45.0	53.1
Main Region[17]	41.1	58.7	44.4	24.1	44.6	53.6
Fine-Grained[13]	41.1	58.8	44.8	23.3	45.4	53.1
GID[8]	41.5	59.6	45.2	24.3	45.7	53.6
SKD[19]	42.3	60.2	45.9	24.4	46.7	55.6
ScaleKD[20]	42.5	–	–	25.9	46.2	54.6
BCKD[21]	43.2	61.6	46.9	25.7	47.3	55.9
FGD[10]	43.4	61.7	47.0	26.2	47.4	56.4
CrossKD[22]	43.7	62.1	47.4	**26.9**	48.0	56.2
IMA (Ours)	**44.0**	**62.2**	**47.7**	26.8	**48.4**	**57.0**

Table 3. Comparison results in GFL framework on MS COCO. Significant values are in bold.

Method	Backbone	mAP@0.5
Faster R-CNN (Teacher)	ResNet-101	78.5
Faster R-CNN (Student)	ResNet-50	76.2
FitNet[12]	ResNet-50	77.1
FGFI[13]	ResNet-50	77.4
FGD[10]	ResNet-50	77.8
IMA (Ours)	ResNet-50	**78.5**
RetinaNet (Teacher)	ResNet-101	77.0
RetinaNet (Student)	ResNet-50	74.8
FitNet[12]	ResNet-50	75.2
FGFI[13]	ResNet-50	75.5
FGD[10]	ResNet-50	76.1
IMA (Ours)	ResNet-50	**76.7**

Table 4. Experimental results on the PASCAL VOC dataset. We report mAP at IoU threshold of 0.5. Significant values are in bold.

outperforming the student baseline by 3.2 mAP points. IMA also provides the best performance on all detailed AP metrics, including AP_{75} and large object detection (AP_L), confirming its robustness in two-stage detectors with high-capacity teachers. Similarly, for the RepPoints detector, IMA obtains the highest mAP of 42.3, outperforming FKD (40.6) and FGD (41.3). Notably, IMA achieves significant improvements in AP_{75} (45.8) and large object detection AP_L (55.9), indicating enhanced localization precision and better adaptation to scale variation. These results collectively demonstrate that IMA not only generalizes well across different detection architectures but also maintains strong performance under the challenging heterogeneous backbone setting, highlighting the effectiveness of the IMA method in bridging the architectural gap between teacher and student models. These results demonstrate the superior ability of our distillation models to capture and represent salient features, which consequently leads to enhanced detection performance.

GFL Framework Results Table 3 provides a detailed comparison of various knowledge distillation methods within the GFL[16] framework on the MS COCO dataset. The baseline student model (GFL-Res50) achieved an mAP of 40.2, while the teacher model (GFL-Res101) attained an mAP of 44.9. Among the existing methods, SKD and ScaleKD demonstrated significant improvements, achieving mAPs of 42.3 and 42.5, respectively. However, the proposed method achieved the best overall mAP of 44.0, closely aligning with the teacher model's performance. The results also highlight the superiority of the proposed method in terms of AP across different IoU thresholds (AP_{50} and AP_{75}) and object sizes (AP_S, AP_M, and AP_L). For instance, the proposed method achieved an AP_{50} of 62.2 and an AP_{75} of 47.7, surpassing all other methods. The gains in AP_S (+3.5) and AP_L (+4.8) further underscore the method's effectiveness in handling both small and large objects. Interestingly, the performance of the proposed method is particularly notable in the context of small object detection, where it achieved an AP_S of 26.8, second only to CrossKD. This indicates that the method addresses the challenges associated with detecting small objects, a common limitation of many KD techniques. Similarly, consistent improvements in AP_M and AP_L suggest that the method effectively balances performance across object scales.

	Method	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Teacher	RetinaNet-ResNeXt101	41.6	61.4	44.3	23.9	45.5	54.5
Student	RetinaNet-Res50	37.4	56.7	39.6	20.0	40.7	49.7
	+ KD	40.2	59.5	43.0	22.6	44.3	53.4
	+ Instance Mask	40.7	60.2	43.4	23.8	44.6	53.9
	+ Standardization	40.9	60.7	43.4	23.5	44.9	53.9
	+ Adaptive Scale	41.0	60.2	43.6	23.0	45.2	55.0

Table 5. Ablation studies on our proposed IMA using ResNet50-based RetinaNet with ResNeXt101-based RetinaNeXt serving as the teacher.

Model	Params (M)	FLOPs (G)	Mem (MB)	FPS
Single-Stage Detectors (RetinaNet)				
T: X101	95.86	424	367	29.4
T: R101	56.96	283	220	30.7
S: R50	37.97	215	148	41.9
Two-Stage Detectors (Faster R-CNN)				
T: X101	135.0	2014	528	20.6
T: R101	60.75	255	244	31.1
S: R50	41.75	187	171	42.1
Anchor-Free Detectors (RepPoints)				
T: X101	94.74	380	230	16.6
T: R101	55.84	239	224	24.5
S: R50	36.85	171	151	31.4

Table 6. Efficiency comparison of different object detectors. In addition to the number of parameters (Params) and FLOPs, we report CUDA memory usage (Mem) and inference speed (FPS), measured on an NVIDIA A100 GPU (80GB). All models are evaluated with an input resolution of 1088×800. T: teacher model, S: student model.

PASCAL VOC Results To further validate the generalizability of our proposed IMA framework, we conduct experiments on the PASCAL VOC dataset. Table 4 presents the experimental results on the PASCAL VOC dataset, comparing our method with baseline and other distillation approaches. Similar to the results on the MS COCO dataset, our method achieves consistent improvements across different detector architectures. For the Faster R-CNN with ResNet-50 backbone, our method improves the mAP by 2.3% compared to the baseline. These results further demonstrate the effectiveness and generalizability of our proposed IMA framework across different datasets and detector architectures.

Ablation Studies Table 5 presents ablation studies to analyze the impact of our proposed IMA approach and its components on the performance of a ResNet50-based RetinaNet student model. The teacher model, RetinaNet-ResNeXt101, achieves an mAP of 41.6, while the student model, RetinaNet-Res50, has a lower mAP of 37.4, indicating a significant performance gap compared to the teacher. Applying conventional Knowledge Distillation (KD) improves the student's mAP from 37.4 to 40.2. The introduction of our proposed Instance Mask Distillation module further boosts the student's mAP to 40.7, highlighting the benefit of transferring instance-level spatial information from the teacher. In addition, combining the Instance Standardisation component of our Cascade Alignment module increases the mAP to 40.9 by reducing internal covariate shift and improving training stability. Finally, the incorporation of the Adaptive Scale Deflation component, which adaptively scales instance-level features based on their importance, yields the highest mAP of 41.0. This step allows the student to focus on the most relevant instances during training, further enhancing knowledge transfer and mitigating the impact of irrelevant or noisy instances. Overall, the ablation studies demonstrate the effectiveness of our IMA approach and its components in bridging the performance gap between student and teacher models, even when their architectures differ significantly.

Computational Efficiency Analysis While the primary goal of knowledge distillation is to improve the performance of compact models, it is equally important to analyze the computational efficiency of the distilled models. Table 6 provides a comprehensive comparison of the computational efficiency of different object detectors used in our experiments. The results clearly demonstrate the computational advantages of the student models compared to their teacher counterparts. For instance, the RetinaNet with ResNet-50 backbone (student model) has approximately 60% fewer parameters and 49% lower FLOPs compared to the ResNeXt-101 backbone (teacher model), while achieving a 42% higher inference speed. Similar trends are observed for the two-stage and anchor-free detectors, where the student models consistently show significant reductions in computational requirements while maintaining competitive performance after distillation. This efficiency analysis underscores the practical value of our proposed IMA framework. By effectively transferring knowledge from computationally intensive

Method	Params (M)	FLOPs (G)	mIoU (%)
T: DeepLabV3-R101	84.74	695	78.07
S: PSPNet-R18	12.61	109	72.55
SKD[19]			73.29
IFVD[23]			73.71
CWD[24]			74.36
CIRKD[25]			74.73
MasKD[26]			75.34
IMA (Ours)			**75.99**

Table 7. Comparison of knowledge distillation methods for semantic segmentation on Cityscapes with the PSPNet-R18 student model. T: teacher model, S: student baseline model. Significant values are in bold.

Method	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline	37.4	56.7	39.6	20.0	40.7	49.7
0.1	40.7	60.2	43.4	23.8	44.6	53.9
0.5	40.9	60.7	43.4	23.5	44.9	53.9
1	41.0	60.2	43.6	23.0	45.2	55.0
5	39.2	58.1	42.2	22.4	43.2	51.0
10	38.5	57.8	41.0	21.8	42.5	50.8

Table 8. Ablation study of loss weight of L_{FMask} with RetinaNet-R50 as student, RetinaNet-X101 as teacher.

teacher models to more efficient student models, our approach enables the deployment of high-performance object detectors in resource-constrained environments, such as edge devices or real-time applications.

Extension to semantic segmentation

To demonstrate the versatility of our proposed IMA framework beyond object detection, we extend it to the task of semantic segmentation. Semantic segmentation, which aims to assign a semantic label to each pixel in an image, shares similar challenges with object detection in terms of the need for fine-grained spatial understanding. We adapt our IMA framework to semantic segmentation by applying the Instance Mask Distillation and Cascade Alignment modules to the feature maps of semantic segmentation models. In this context, the ”instances” correspond to regions of pixels belonging to the same semantic class. We evaluate our adapted approach on the Cityscapes dataset, using a PSPNet with ResNet-18 backbone as the student model and a DeepLabV3 with ResNet-101 backbone as the teacher model.

Table 7 presents the results of our semantic segmentation experiments, comparing our adapted IMA framework with several state-of-the-art knowledge distillation methods specifically designed for semantic segmentation. The results show that our adapted IMA framework outperforms all other knowledge distillation methods for semantic segmentation, achieving an mIoU of 75.99%, which is a 3.44% improvement over the baseline student model. This represents a significant step toward closing the gap with the teacher model (78.07% mIoU), all while maintaining the computational efficiency of the student model (12.61M parameters vs. 84.74M for the teacher, and 109G FLOPs vs. 695G for the teacher). Notably, our approach surpasses recent specialized methods such as MasKD (ICLR’23) and CIRKD (CVPR’22) by 0.65% and 1.26%, respectively. This is particularly impressive given that these methods were specifically designed for semantic segmentation, whereas our IMA framework was originally developed for object detection and adapted to semantic segmentation. These results demonstrate that the core principles underlying our IMA framework-namely, the use of instance-level mask information and feature alignment through standardization and adaptive scaling-are applicable beyond object detection and can effectively improve performance in other dense prediction tasks such as semantic segmentation.

Hyperparameter Analysis We experimentally analyze the impact of the hyperparameter of loss weight of L_{FMask} on detection results. Table 8 presents the findings obtained by varying the hyperparameter from 0.1 to 10.0. We observe that the best mAP result is achieved when using the hyperparameter value of 1.0 for knowledge distillation, providing insights into the optimal hyperparameter setting for our approach.

Precision-Recall Analysis The precision-recall curves illustrated in Fig. 2 provide insightful analysis into the effectiveness of our distillation method in enhancing the localization and classification capabilities of the student baseline detector. As illustrated in Fig. 2, using the dog class as an example, the student models trained with our distillation technique consistently outperform their baseline counterparts without distillation. We observe significant improvements in both mAP and inference time, indicating enhanced performance and computational efficiency of the distilled student models. Furthermore, we compare our method with other knowledge distillation approaches, such as attention transfer and feature map distillation. Our IMA method achieves superior mAP and inference time performance, highlighting its effectiveness in distilling knowledge for object detection tasks. The results clearly demonstrate that our approach significantly reduces various types of errors, effectively minimizing false detections, background errors, and missed ground truth instances. By distilling knowledge

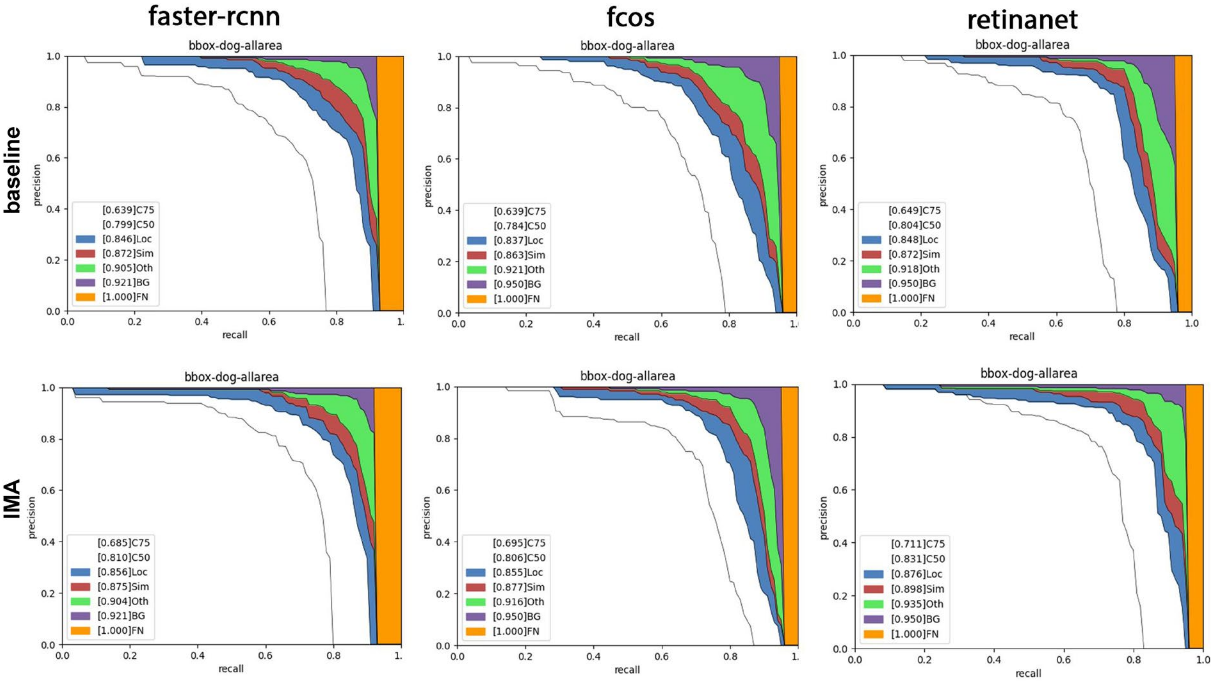

Fig. 2. Error analyses of baseline students (***First Row***) and students distilled by our approach (***Second Row***) on COCO benchmarks. C50 and C75: performance at specific IoU thresholds; Loc: localization errors; Sim and Oth: class confusion; BG: background discrimination; FN: detection completeness.

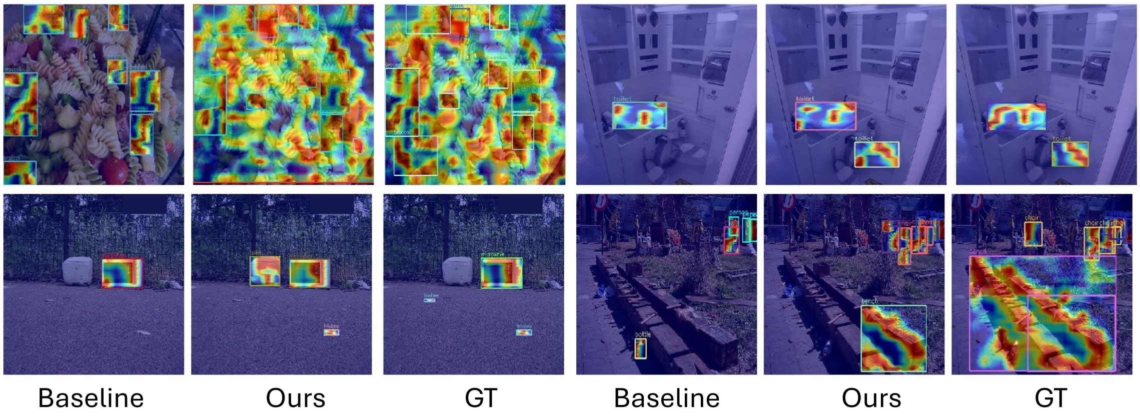

Fig. 3. Qualitative analysis of baseline Faster RCNN and Faster RCNN distilled by Baseline, Our IMA and GT on COCO benchmarks.

from the teacher model, the student detector exhibits improved specificity, accurately distinguishing between objects and background regions, thereby reducing false positive detections. Moreover, our method enhances the sensitivity of the student model, enabling it to detect and classify objects more effectively, addressing the issue of missed ground truth instances that baseline detectors often struggle with. Collectively, these findings underscore the ability of our distillation technique to significantly improve the localization and classification performance of the student detector, rendering it more reliable, robust, and accurate for object detection tasks across diverse scenarios.

Qualitative Analysis Figure 3 provides a qualitative visualization of the detection outputs from different detectors, allowing for a comprehensive comparison between our IMA distillation detector and the baseline student detector. The visual analysis highlights several critical advantages exhibited by our approach. Notably, our IMA distillation detector demonstrates superior performance in detecting small targets. This highlights the effectiveness of our distillation technique in transferring knowledge related to detecting and recognizing small-scale objects, a challenging task in object detection. Additionally, our method significantly reduces the occurrence of false positive detections, minimizing the identification of non-existent objects, which is crucial for reliable and interpretable detection systems. Furthermore, our approach effectively addresses false negatives, where baseline detectors fail to detect objects present in the scene, thereby enhancing the comprehensiveness and completeness of the detection process. It is of great importance to note that our IMA distillation technique mitigates performance degradation, ensuring robust and consistent detection accuracy across diverse scenarios. The qualitative analysis of the detection visualization results substantiates the superior accuracy in detecting small targets by our IMA distillation detector.

Discussion

Related work

Object Detection Object detection algorithms can be categorized as two-stage or one-stage. Two-stage methods, such as Faster R-CNN[27] and Cascade R-CNN[28] maintain high accuracy through the use of RPN and the refinement of classification and location. One-stage detectors including SSD[29] and YOLO[30] have lower latency by direct prediction from feature maps. Recent methods also distinguish between anchor-based and anchor-free detectors. Anchor-based detectors, such as SSD[29] and Faster R-CNN[31], rely on predefined anchor boxes. However, many anchors pose challenges. Anchor-free methods, including CenterNet[32] and CornerNet[33], predict key points like center, achieving better performance with fewer costs. Addressing the foreground-background imbalance is crucial. Two-stage detectors employ sampling and OHEM[34] to reduce the background. One-stage approaches, such as RetinaNet[7], introduce focal loss. Anchor-free detectors like FCOS[6] and FoveaBox[35] eliminate anchors, reducing operations and tuning. Recent work also proposes dynamic label assignment[36,37] to better define samples. Additionally, the DETR family of detectors[38] has gained popularity due to the powerful feature encoding capabilities of transformer blocks. These detectors are capable of encoding highly expressive features and have emerged as a significant trend in the object detection community.

Knowledge Distillation for Object Detection As one of the popular model compression and performance optimization strategies, knowledge distillation[4,39] has been widely explored in various domains, including object detection, image classification, and natural language processing. In the context of object detection, existing methods primarily focus on aligning the feature representations or output predictions between the teacher and student models. For example, some approaches employ feature mimicking techniques, where the student model's feature maps are guided to mimic the teacher's feature maps through additional loss terms or attention mechanisms. For instance, Zheng et al.[40] emphasize a valuable localization region to leverage classification and localization information. Yang et al.[41] propose a multi-scale imitation function of the core features for the distillation of the adaptive reinforcement control. Dai et al.[8] go a step further by distilling discriminative patches between students and teachers. Zhang et al.[42] propose a method for structured knowledge distillation using an attention mechanism for guided distillation. However, these previous works only employ a fixed teacher for experiments, without exploring the relationship between teacher and student performance in object detection tasks. Some recent studies have introduced new approaches for knowledge distillation in object detection. FGD[10] aligns the attention between teacher and student models, while PKD[43] maximizes the Pearson Correlation Coefficient between their feature representations. Huang et al.[44] propose a DISK method, which uses correlation-based loss to better capture interclass relationships. Liu et al.[45] propose a cross-architecture distillation method. Prediction mimicking, commonly used in classification distillation, has also been adapted for object detection. For example, Tu et al.[5] propose a dynamic distillation method in which teacher and student networks can learn from each other. Lv et al.[46] propose a gap-free feature imitation method to decouple the encode and decode distillation process. Zhang et al.[47] propose an explainable distillation method that uses the class activation map to exploit information about both the structure and the label. Song et al.[48] propose a closed-loop method combining hierarchical re-weighted attention distillation and detection head classification for dense object detection. In contrast to these methods, our approach focuses on aligning feature knowledge to reduce gaps in teacher-student detectors.

Limitations and future work

While our proposed IMA framework demonstrates superior performance in bridging the gap between teacher and student models, several limitations and areas for future improvement can be identified:

Computational Overhead The Instance Mask Distillation module and Cascade Alignment Module introduce additional computational overhead during training. Although this overhead is only present during the training phase and does not affect inference, it may still impact the training efficiency, especially for large-scale datasets. Future work could explore more efficient implementations of these modules, potentially leveraging techniques such as pruning or quantization to reduce the computational cost.

Hyperparameter Sensitivity As shown in our hyperparameter analysis, the performance of our IMA framework is sensitive to the choice of the loss weight of L_{FMask}. This sensitivity may require extensive hyperparameter tuning for optimal performance, which can be time-consuming and resource-intensive. Future research could investigate more robust methods for automatically determining the optimal hyperparameter values or develop adaptive schemes that adjust the hyperparameters during training.

Teacher Model Selection Our results show that the performance of the student model is influenced by the choice of the teacher model. However, we have not systematically explored the relationship between the teacher and student model architectures and the resulting performance. Future research could investigate this relationship more thoroughly, potentially providing guidelines for selecting the most suitable teacher model for a given student model.

Conclusion

In this paper, we propose a novel Instance Mask Alignment Knowledge Distillation (IMA) framework for object detection. Our framework effectively bridges the gap between different detector architectures through a cascade of knowledge transformation operations. The introduction of instance mask distillation enables the student model to learn from the teacher's instance-level mask information, improving its ability to identify and attend to relevant objects. Furthermore, the cascade alignment module, consisting of instance standardization and adaptive scale deflation, enhances the training stability and guides the student model towards the most crucial instances. Extensive experiments on various detectors and multiple benchmarks demonstrate the significant performance improvements achieved by our IMA framework.

Our work provides several valuable insights into the knowledge distillation process for object detection. First, we identify the importance of instance-level information in the distillation process, demonstrating that this information can significantly enhance the student model's performance. Second, we show that addressing the feature distribution gap between teacher and student models through instance standardization and adaptive scaling is crucial for effective knowledge transfer. Finally, we empirically validate the effectiveness of our approach across a range of detector architectures and datasets, highlighting its generalizability and robustness.

Despite these strengths, our approach does have limitations, particularly in terms of computational overhead during training and sensitivity to hyperparameters. These limitations suggest promising directions for future research, including exploring more efficient implementations, investigating automatic hyperparameter tuning methods, and extending the approach to other computer vision tasks. Furthermore, future work could explore more efficient ways to incorporate instance-level information, such as using lightweight attention mechanisms or sparse representations, and extend the proposed approach to other computer vision tasks, such as instance segmentation or panoptic segmentation.

Data availability

The datasets used and analyzed during the current study are available in the COCO (https://cocodataset.org/), PASCAL VOC (http://host.robots.ox.ac.uk/pascal/VOC/) and Cityscapes (https://www.cityscapes-dataset.com/).

Received: 27 October 2024; Accepted: 19 May 2025
Published online: 25 September 2025

References

1. Zou, Z., Chen, K., Shi, Z., Guo, Y. & Ye, J. Object detection in 20 years: A survey. *Proc. IEEE* **111**, 257–276 (2023).
2. Li, Z. et al. When object detection meets knowledge distillation: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 10555–10579 (2023).
3. Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015).
4. Gou, J. *et al.* Neighborhood relation-based knowledge distillation for image classification. *Neural Netw.* 107429 (2025).
5. Tu, Z., Liu, X. & Xiao, X. A general dynamic knowledge distillation method for visual analytics. *IEEE Trans. Image Proc.* **31**, 6517–6531 (2022).
6. Tian, Z., Shen, C., Chen, H. & He, T. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 9627–9636 (2019).
7. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision* 2980–2988 (2017).
8. Dai, X. *et al.* General instance distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 7842–7851 (2021).
9. Zhixing, D. et al. Distilling object detectors with feature richness. *Adv. Neural Inf. Process. Syst.* **34**, 5213–5224 (2021).
10. Yang, Z. *et al.* Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 4643–4652 (2022).
11. Li, L., Dong, P., Li, A., Wei, Z. & Yang, Y. Kd-zero: Evolving knowledge distiller for any teacher-student pairs. *Adv. Neural Inf. Process. Syst.* **36**, 69490–69504 (2023).
12. Romero, A. *et al.* Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014).
13. Wang, T., Yuan, L., Zhang, X. & Feng, J. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 4933–4942 (2019).
14. Heo, B. *et al.* A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 1921–1930 (2019).
15. Zhang, L. & Ma, K. Towards accurate and efficient detectors. In *International Conference on Learning Representations, Improve Object Detection with Feature-Based Knowledge Distillation* (2020).
16. Li, X. et al. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* **33**, 21002–21012 (2020).
17. Zheng, Z. *et al.* Localization distillation for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9407–9416 (2022).
18. Li, Q., Jin, S. & Yan, J. Mimicking very efficient network for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 6356–6364 (2017).
19. De Rijk, P., Schneider, L., Cordts, M. & Gavrila, D. Structural knowledge distillation for object detection. *Adv. Neural Inf. Process. Syst.* **35**, 3858–3870 (2022).
20. Zhu, Y. *et al.* Scalekd: Distilling scale-aware knowledge in small object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 19723–19733 (2023).
21. Yang, L. *et al.* Bridging cross-task protocol inconsistency for distillation in dense object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 17175–17184 (2023).
22. Wang, J. *et al.* Crosskd: Cross-head knowledge distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 16520–16530 (2024).
23. Wang, Y., Zhou, W., Jiang, T., Bai, X. & Xu, Y. Intra-class feature variation distillation for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16* 346–362 (Springer, 2020).
24. Shu, C., Liu, Y., Gao, J., Yan, Z. & Shen, C. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 5311–5320 (2021).
25. Yang, C. *et al.* Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 12319–12328 (2022).
26. Huang, T. *et al.* Masked distillation with receptive tokens. arXiv preprint arXiv:2205.14589 (2022).

27. Ren, S., He, K., Girshick, R. & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1137–1149 (2016).
28. Cai, Z. & Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 6154–6162 (2018).
29. Liu, W. *et al.* Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* 21–37 (Springer, 2016).
30. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 779–788 (2016).
31. Lin, T.-Y. *et al.* Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2117–2125 (2017).
32. Duan, K. *et al.* Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 6569–6578 (2019).
33. Law, H. & Deng, J. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)* 734–750 (2018).
34. Shrivastava, A., Gupta, A. & Girshick, R. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 761–769 (2016).
35. Kong, T. et al. Foveabox: Beyound anchor-based object detection. *IEEE Trans. Image Process.* **29**, 7389–7398 (2020).
36. Nguyen, C. H., Nguyen, T. C., Tang, T. N. & Phan, N. L. Improving object detection by label assignment distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* 1005–1014 (2022).
37. Deng, C., Wang, M., Liu, L., Liu, Y. & Jiang, Y. Extended feature pyramid network for small object detection. *IEEE Trans. Multimed.* **24**, 1968–1979 (2021).
38. Dai, L., Liu, H., Tang, H., Wu, Z. & Song, P. Ao2-detr: Arbitrary-oriented object detection transformer. *IEEE Trans. Circuits Syst. Video Technol.* **33**, 2342–2356 (2022).
39. Wang, L. & Yoon, K.-J. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 3048–3068 (2021).
40. Zheng, Z. et al. Localization distillation for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 10070–10083 (2023).
41. Yang, Y. et al. Adaptive knowledge distillation for lightweight remote sensing object detectors optimizing. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–15 (2022).
42. Zhang, L. & Ma, K. Structured knowledge distillation for accurate and efficient object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 15706–15724 (2023).
43. Cao, W. et al. Pkd: General distillation framework for object detectors via pearson correlation coefficient. *Adv. Neural Inf. Process. Syst.* **35**, 15394–15406 (2022).
44. Huang, T., You, S., Wang, F., Qian, C. & Xu, C. Dist+: Knowledge distillation from a stronger adaptive teacher. *IEEE Trans. Pattern Anal. Mach. Intell.* (2025).
45. Liu, Y. et al. Cross-architecture knowledge distillation. *Int. J. Comput. Vis.* **132**, 2798–2824 (2024).
46. Lv, Y., Cai, Y., He, Y. & Li, M. Drkd: Decoupling response-based distillation for object detection. *Pattern Recognit.* **161**, 111275 (2025).
47. Sun, T., Chen, H., Hu, G. & Zhao, C. Explainability-based knowledge distillation. *Pattern Recognit.* **159**, 111095 (2025).
48. Song, Y. et al. Closed-loop unified knowledge distillation for dense object detection. *Pattern Recognit.* **149**, 110235 (2024).

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2022YFC3302100).

Author contributions

Z.G. and P.Z. conceived the experiment(s), Z.G. drafted the manuscript, Z.G. and P.L. conducted the experiment(s), Z.G. and P.L. analysed the results. P.Z. advised on model designs/experiments. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.G. or P.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.