DCSF-KD: Dynamic Channel-wise Spatial Feature Knowledge Distillation for Object Detection

Tao Dai[1,4], **Yang Lin**[1], **Hang Guo**[3], **Jinbao Wang**[2,5], **Zexuan Zhu**[1,2,4,*]
[1]College of Computer Science and Software Engineering, Shenzhen University
[2]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University
[3]Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[4]Shenzhen City Key Laboratory of Embedded System Design, Shenzhen, China
[5]Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen, China
{daitao.edu,liny24.edu,cshguo}@gmail.com, {wangjb, zhuzx}@szu.edu.cn

Abstract

Knowledge distillation (KD) has recently gained great success in the field of object detection. By transferring the knowledge of the spatial or channel domain from the teacher model to the student model, it allows for a more compact representation with minimal performance loss. Despite this progress, existing KD methods typically treat knowledge from spatial or channel domains independently, ignoring the exploitation of the mutual relationship between these domains. In this work, we first explore the connection between spatial and channel domains and find there exists a strong correlation between them, i.e. the salient channels tend to contain significant object regions in the spatial domain. Motivated by this observation, we propose DCSF-KD, a novel Dynamic Channel-wise Spatial Feature Knowledge Distillation framework for object detection by fully exploiting both spatial and channel knowledge. Specifically, we introduce channel-wise spatial feature distillation and global channel attention distillation, using information from both domains to improve the accuracy of the student network. Experiments demonstrate that our DCSF-KD outperforms existing detection methods on both homogeneous and heterogeneous teacher-student network pairs. For example, when using the MaskRCNN-Swin detector as the teacher, and based on RetinaNet and FCOS with ResNet-50 on MS COCO, our DCSF-KD can achieve 41.9% and 44.1% mAP, respectively.

Code — https://github.com/LinY-ct/DCSF-KD

Introduction

Knowledge distillation, aiming to compress a large teacher model to obtain a compact student model without loss of performance, has been widely applied across various computer vision tasks, such as object detection. Currently, the principal approaches to knowledge distillation in object detection focus on the efficient transfer of features within the detector's Neck layer. These methods can generally be divided into two main categories: The first method involves knowledge transfer of spatial dimension features, which includes identifying and differentiating key region features within the network that are sensitive to the detection task, such as distinguishing between foreground and background features (Dai et al. 2021). This approach achieves commendable performance through carefully designed weighting functions that identify key features. In addition, there exist methods that focus on transferring the overall relationships of characteristics in the spatial domain, such as GI (Yang et al. 2022) and PKD (Cao et al. 2022), which achieve satisfactory results without the need for complex distinction functions. The second way involves feature transfer on the channel dimension, where each channel feature represents different visual patterns of the image, adaptively focusing on different region features of the image. For instance, CWD (Shu et al. 2021) converts feature outputs into a distribution form, utilizing the minimization of KL divergence distance between teacher and student distributions for knowledge distillation.

Our observations indicate that the first manner faces challenges in designing complex distinction functions to identify sensitive features in detection tasks and exhibits performance instability across different detection architectures. For example, in FRS (Zhixing et al. 2021), the background region is also considered to contain crucial information for the detection task, suggesting that effective features cannot be simply identified through foreground-background differentiation. However, while holistic distillation methods simplify the process, they tend to transfer a large amount of redundant information during the distillation process, negatively affecting the final performance. Thus, the current methods based on channel feature transfer in object detection knowledge distillation overlook the varying importance of each channel, therefore achieving suboptimal results.

In this paper, we introduce a general Dynamic Channel-wise Spatial Feature Knowledge Distillation (DCSF-KD) for object detection, which aims at efficiently transferring key detection knowledge by combining prominent channel features with channel-wise spatial features. As shown in Fig. 1, in object detection tasks, different channel features represent different visual patterns of the image. We reshape the normalized features of the FPN layer into a representation for each channel. Inspired by SENet (Hu, Shen, and Sun 2018), the knowledge of features across different channels varies in importance for the task. Through Global Average Pooling (GAP) operations, we generate weight values for each channel, emphasizing the significance of diverse channel features. In this way, we can adaptively focus on

*Corresponding author: Zexuan Zhu

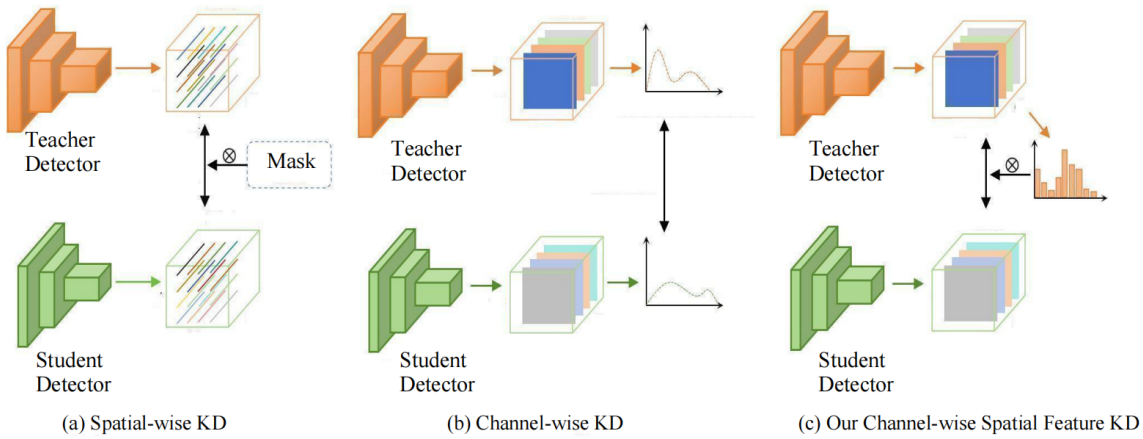

Figure 1: We compare the general forms of spatial distillation and channel distillation. Dynamic Channel-wise Spatial Feature Knowledge Distillation proposes to fully leverage knowledge from two feature domains: channel-wise spatial features and global activation knowledge of channels.

channel-wise spatial features according to channel weights. DCSF-KD offers two main advantages: first, our method does not require the design of complex weighting functions to differentiate sensitive features of the detection task, representing the crucial knowledge in knowledge distillation in a simple and efficient manner. Secondly, DCSF-KD is applicable to various architectures of object detectors, addressing the issue of instability in the distillation process across different architectures in practical applications. By extracting feature knowledge from the FPN layer, DCSF-KD is adaptable to various types of detectors.

Without bells and whistles, our method significantly enhances the performance of the student detector, achieving faster training convergence. In this paper, we conduct comprehensive experiments on the MS COCO dataset. We achieve 41.0% mAP and 40.7% mAP in teacher-student pairs for RetinaResX101-R50 and FasterRCNNRes101-Res50 in both one-stage and two-stage settings, respectively; and 41.9% mAP in heterogeneous detectors of MaskRCNN-Swin and Retian-Res50, substantially surpassing previous state-of-the-art methods.

The main contributions are summarized as follows.

- We propose dynamic channel-wise spatial feature knowledge distillation for object detection, which, without the need to design complex weighting functions to differentiate important areas in space, can adaptively transfer the key region features of the image and represent the importance of different channels by the activation values of channels.
- We show that different channels have varying levels of importance to the detection task, with the network being more sensitive to channels with higher activation values.
- Experiments indicate that our DCSF-KD, when deployed across various architectures of object detectors, surpasses the performance of existing mainstream methods.

Related Work

Object Detection

Object detection aims to identify the categories and location information of objects within images. In recent years, object detection algorithms based on deep learning have become the mainstream, categorized into two main types: two-stage (Chen et al. 2019a; Girshick 2015; Girshick et al. 2014; Guo et al. 2021b) and single-stage algorithms (Li et al. 2020a; Lin et al. 2017; Redmon and Farhadi 2018; Zhang et al. 2020). Two-stage algorithms initially select candidate regions in an image and subsequently refine these regions for object classification and precise localization, such as Faster R-CNN (Ren et al. 2016). On the other hand, single-stage algorithms perform regression and classification globally, directly predicting the positions and categories of objects, offering real-time capabilities at the expense of some detection accuracy. Examples of single-stage object detectors include FCOS (Tian et al. 2019) and RetinaNet (Lin et al. 2020). More recently, networks like DETR (Carion et al. 2020; Li et al. 2022a; Liu et al. 2022; Lin et al. 2017), leveraging the powerful representation capabilities of Transformers, have emerged as a new trend in object detection tasks. In these works, the Feature Pyramid Network (FPN) layer is employed to capture objects of varying scale sizes, making our method applicable across diverse architectures of object detectors.

Knowledge Distillation

Knowledge Distillation (KD) (Hinton, Vinyals, and Dean 2015), as a model compression strategy, was initially applied to image classification tasks (Tian, Krishnan, and Isola 2019; Zagoruyko and Komodakis 2016; Zhao et al. 2022; Cho and Hariharan 2019) . The formative approaches, inspired by the FitNets (Romero et al. 2014) concept, entailed fitting intermediate features within the backbone of two-stage object detectors for KD (Chen et al. 2017). Presently, the predominant methodologies are categorized into spatial distillation (Cao

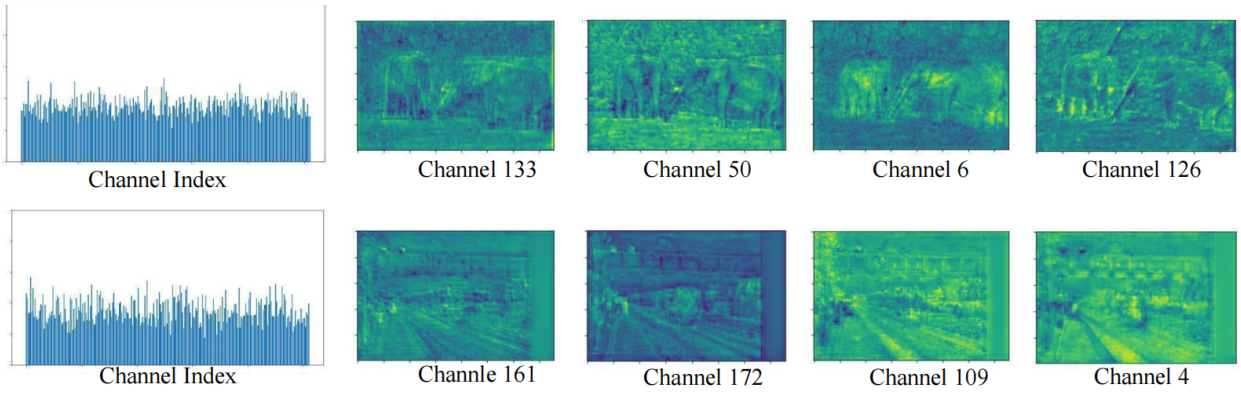

Figure 2: We visualized single-channel images from the MS COCO training set, along with the distribution of channel attention weights. The four visualized channel images represent the scenarios of the smallest weight, the second smallest weight, the second largest weight, and the largest weight, respectively.

et al. 2022; Wang et al. 2019; Zhang and Ma 2020) and channel distillation (Li et al. 2022b) based on the KD approach.

Spatial distillation: Spatial distillation methods are concentrated on pivotal regions pertinent to detection tasks. For example, Quanquan Li et al. (Li, Jin, and Yan 2017) proposed focusing on the outputs of the RPN layer instead of the entire feature map to simplify the task; Tao Wang et al. (Wang et al. 2019) introduced learning the spatial regions around GT boxes; Linfeng Zhang et al. (Zhang and Ma 2020) applied a Gaussian Mask to distinguish between foreground and background features, emphasizing spatial region features sensitive to detection. Successive techniques such as FGD (Yang et al. 2022) and FRS (Guo et al. 2021a) have built on this concept, offering foreground-background attention mechanisms and maximum classification score strategies, respectively, to achieve notable results. Moreover, methodologies like PKD (Cao et al. 2022) and GI (Dai et al. 2021) have enhanced distillation generality and efficacy by transferring holistic spatial feature relationships.

Channel distillation: In contrast, channel distillation methods, such as those advocated by Zhou et al. (Zhou et al. 2020) focus on the knowledge within each channel, employing weighted MSE to reduce the disparity in channel features. CSC (Park and Heo 2020) conveys related knowledge through the calculation of pairwise relationships between spatial and channel features. Channel exchanging (Wang et al. 2020) posited that channel features signify key attributes, showcasing substantial generality and efficacy suitable for cross-modal knowledge tasks. CWD (Shu et al. 2021) converting channel features into a distribution via softmax to minimize the disparity in channel features between teacher and student, emerges as a practical and efficient method. These channel distillation strategies, primarily targeting the transfer of related knowledge between teacher-student outputs and acknowledging the varied significance of different channels to the network, warrant further investigation for their efficacy.

Addressing the need to simultaneously consider per-channel spatial feature knowledge and channel importance, this work proposes a novel Dynamic Channel-wise Spatial Feature Knowledge Distillation for Object Detection.Our methodology mitigates the difficulties in identifying critical spatial features within spatial feature distillation and the redundancy inherent in channel distillation techniques. As a Dynamic Channel-wise Spatial Feature Knowledge Distillation (DCSF-KD) approach, diverging from prior channel centric methods such as CWD(Shu et al. 2021), DCSF-KD adaptively zeroes in on detection-sensitive region features according to channel weights. Compatible with diverse object detector architectures, DCSF-KD secures state-of-the-art (SOTA) achievements on the MS COCO dataset.

Method

Spatial Distillation

We first define the general form of the spatial distillation method. In object detection knowledge distillation methods, specific approaches divide the detector into three parts: (1) the backbone network, which is used for feature extraction; (2) FPN multi-scale layers, used for integrating features of different scales; (3) detection network output layers, used for generating regression and classification scores. For feature-based knowledge distillation, intermediate feature representations from the teacher and the student at the output of the FPN layers are selected, denoted by $T \in R^{C\times H\times W}$ and $S \in R^{C\times H\times W}$, respectively. Therefore, the feature-based distillation loss between T and S can be expressed as:

$$\mathcal{L}_{\text{feat}} = \sum_{l=1}^{L} \frac{1}{N_l} \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathcal{L}_{\varepsilon} \left(\nu(\phi(S_{l,c,h,w})), \nu(T_{l,c,h,w})\right) \tag{1}$$

where L, C, H, W represent the number of FPN layers, channels, height, and width respectively, and $N_l = C \times H \times W$ is the total number of elements in the lth output scale. Additionally, $\mathcal{L}_{\varepsilon}$ represents a distance metric function, applied to different models' outputs to minimize the differences between feature map outputs. $\phi(\cdot)$ is an adaptation layer to ensure that T and S dimensions are consistent, typically a 1×1 convolution layer. $\nu(\cdot)$ is defined as a weighting func-

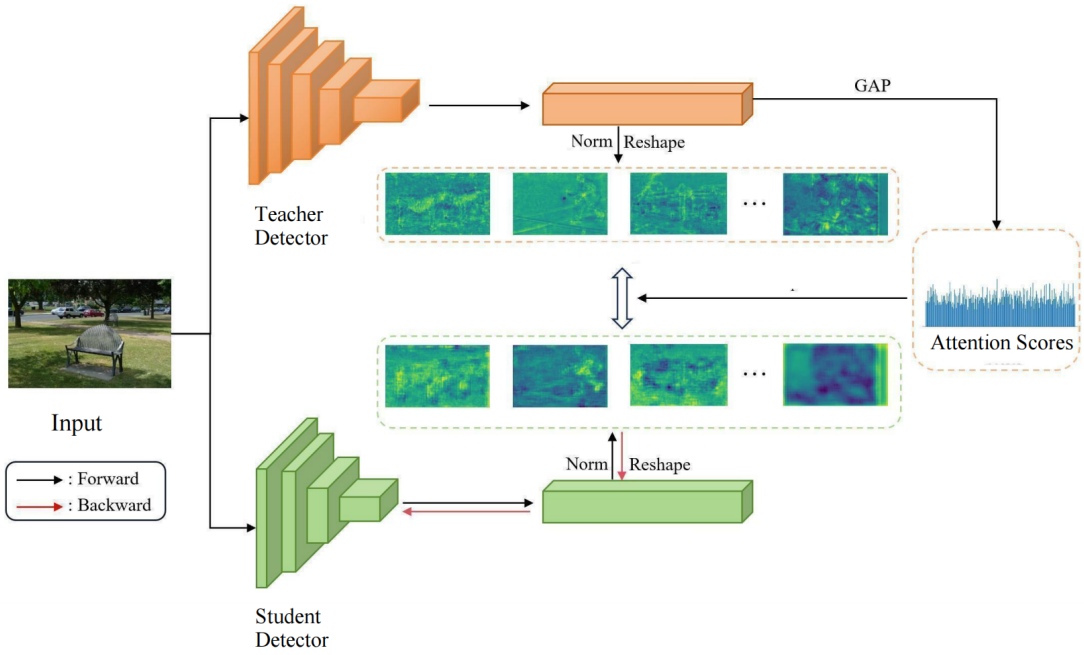

Figure 3: Overall Framework of DCSF-KD, featuring adaptive channel weights, which involve global pooling of the teacher network's FPN layer and the application of a softmax function to yield attention scores between 1 and 10. The reshape operation is designed to represent the normalized features on a channel-wise basis.

tion, focusing on crucial region features of the detection network. In the feature-based distillation loss approach, different methods process the outputs of the FPN layer differently, for example, multiplying by a weight matrix that separates the foreground or background, or according to a class score weight matrix, making feature distillation focus on spatial key areas of the object detection task.

Channel-wise Distillation

Unlike spatial distillation methods, channel distillation approaches do not require the design of complex spatial masks and can naturally represent various important features of an image, such as measuring the differences between teacher and student using the form of activation distributions.In the channel distillation approach, we convert feature outputs to a per-channel basis. Similar to the method described above, we define:

$$\mathcal{L}_{\text{feat}} = \sum_{l=1}^{L} \frac{1}{N_l} \sum_{c=1}^{C} \sum_{w=1}^{H \times W} \mathcal{L}_{\varepsilon} \left(\nu(\phi(S_{l,h,w,c})), \nu(T_{l,h,w,c}) \right) \tag{2}$$

where $\phi(\cdot)$ represents the alignment function, and $\nu(\cdot)$ signifies the transformation function for the outputs of T and S, such as converting into a distribution form. $\mathcal{L}_{\varepsilon}$ denotes the distance metric function. In the CWD[31] method, T and S are transformed into the channel distribution form using softmax, subsequently utilizing KL divergence to assess the distance between the output distributions.

$$\nu(F_c) = \frac{\exp(F_{c,i}/\tau)}{\sum_{i=1}^{W \cdot H} \exp(F_{c,i}/\tau)} \tag{3}$$

$$\mathcal{L}_{\text{feat}}(T, S) = \frac{\tau^2}{C} \sum_{c=1}^{C} \sum_{i=1}^{W \cdot H} \nu(T_{c,i}) \cdot \log \frac{\nu(S_{c,i})}{\nu(S_{c,i})} \tag{4}$$

Channel distillation allows for the adaptive use of knowledge from each channel without needing to design complex weighting functions in the spatial dimension. Nonetheless, different channels vary in importance to the network, with channels exhibiting higher activation values typically indicating region features more sensitive to the object detection network.

Dynamic Channel-wise Spatial Feature Knowledge Distillation

Traditional knowledge distillation methods primarily focus on finding per-pixel relationships or local pixel similarities in the spatial dimension for knowledge transfer. However, these methods often struggle to accurately determine which areas in the space are effective and salient regions of

Method	schedule	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Retina-ResX101 (T)	2x	40.8	60.5	43.7	22.9	45.6	54.6
Retina-Res50 (S)	2x	37.4	56.7	39.6	20.0	40.7	49.7
FKD(Zhang and Ma 2020)	2x	39.6 (+2.2)	58.8	42.1	22.7	43.3	52.5
FRS(Zhixing et al. 2021)	2x	40.1 (+2.7)	59.5	42.5	21.9	43.7	54.3
FGD(Yang et al. 2022)	2x	40.4 (+3.0)	59.9	43.3	23.4	44.7	54.1
PKD(Cao et al. 2022)	2x	40.8 (+3.4)	60.3	43.4	23.0	45.1	54.7
DCSF-KD(Ours)	2x	**41.0(+3.6)**	**60.4**	**43.7**	**23.9**	**45.3**	**54.8**
FasterRCNN-Res101 (T)	2x	39.8	60.1	43.3	22.5	43.6	52.8
FasterRCNN-Res50 (S)	2x	38.4	59.0	42.0	21.5	42.1	50.3
GID(Dai et al. 2021)	2x	40.2 (+1.8)	60.7	43.8	22.7	44.0	53.2
FRS(Zhixing et al. 2021)	2x	40.4 (+2.0)	60.8	44.0	23.2	44.4	53.1
FGD(Yang et al. 2022)	2x	40.4 (+2.0)	60.7	44.3	22.8	44.5	**53.5**
PKD(Cao et al. 2022)	2x	40.5 (+2.1)	60.9	44.4	22.6	44.8	53.1
DCSF-KD(Ours)	2x	**40.7(+2.3)**	**61.0**	**44.6**	**23.2**	**45.0**	**53.5**
FCOS-Res101 (T)	2x+ms	41.2	60.4	44.2	24.7	45.3	52.7
FCOS-Res50 (S)	2x+ms	39.1	58.4	41.6	24.0	42.7	48.7
FRS(Zhixing et al. 2021)	2x+ms	42.2 (+3.1)	60.6	45.6	**27.1**	46.5	53.0
FGD(Yang et al. 2022)	2x+ms	42.3 (+3.2)	60.8	45.8	26.1	46.7	53.3
PKD(Cao et al. 2022)	2x+ms	42.8 (+3.7)	61.4	46.2	25.9	**47.2**	54.6
DCSF-KD(Ours)	2x+ms	**42.9(+3.8)**	**61.6**	**46.4**	26.2	47.0	**54.8**
RepPoints ResNeXt101 (T)	2x	44.2	65.5	47.8	26.2	48.4	58.5
RepPoints Res50 (S)	2x	38.6	59.6	41.6	22.5	42.2	50.4
FGD(Yang et al. 2022)	2x	41.3 (+2.7)	-	-	**24.5**	45.2	54.0
PKD(Cao et al. 2022)	2x	42.3 (+3.7)	63.1	45.4	23.9	46.6	56.5
DCSF-KD(Ours)	2x	**42.4(+3.8)**	**63.2**	**45.6**	24.2	**46.8**	**56.5**
TOOD-ResX101 (T)	2x+ms	47.6	68.5	51.6	30.6	51.4	59.7
TOOD-Res50 (S)	1x	42.4	59.7	46.2	25.4	45.5	55.7
PKD(Cao et al. 2022)	2x	46.0 (+3.6)	63.7	50.0	**28.5**	50.0	**60.1**
DCSF-KD(Ours)	2x	**46.2(+3.8)**	**63.9**	**50.1**	28.1	**50.3**	59.9
GFL-R101 (T)	2x	44.9	63.1	49.0	28.0	49.1	57.2
GFL-R50 (S)	2x	40.2	58.4	43.3	23.3	44.0	52.2
PKD(Cao et al. 2022)	2x	43.3(+3.1)	61.3	46.9	25.2	47.9	56.2
CrossKD(Wang et al. 2023)	2x	43.7(+3.5)	62.1	47.4	**26.9**	48.0	56.2
DCSF-KD(Ours)	2x	**44.1(+3.9)**	**62.3**	**47.8**	25.9	**48.4**	**57.4**

Table 1: Results on the MS COCO dataset with homogeneous teacher detectors. '1x' corresponds to 12 epochs, '2x' corresponds to 24 epochs, and 'ms' refers to a composite training strategy setting. The best results are highlighted in **bold**.

the teacher network, and the teacher network tends to have an excess of redundant information. Although some studies have attempted to identify salient regions by differentiating the importance of foreground and background, the key areas in object detection are not limited to the foreground; the background also contains important information. In deep convolutional neural networks, channels of an image represent different visual patterns, corresponding to different image features, which can be represented as sensitive area features for the object detection task. We introduce a novel Dynamic Channel-wise Spatial Feature Knowledge Distillation approach: for spatial distillation, we extract spatial feature information on a per-channel basis; for channel distillation, inspired by the concepts from AT (Zagoruyko and Komodakis 2016) and SENet (Hu, Shen, and Sun 2018), we utilize Global Average Pooling (GAP) to compute the attention scores for each channel.

The overall framework of DCSF-KD is illustrated in Fig. 3 .In knowledge distillation, the feature maps of the student network often do not match the dimensions of the teacher network. After dimension alignment, the outputs of the FPN layers from the teacher and student networks are denoted as $T \in R^{L\times C\times H\times W}$ and $S \in R^{L\times C\times H\times W}$, where L is the number of FPN layers, C is the number of channels, and H and W are the height and width of the feature maps, respectively. First, the DCSF-KD method processes T and S to reduce the numerical differences in the FPN layer outputs through the batch normalization method $\phi(\cdot)$.

$$T' = \phi(T), \quad S' = \phi(S). \tag{5}$$

where $T' \in R^{L\times C\times H\times W}$ and $S' \in R^{L\times C\times H\times W}$.

Considering the superior performance of the teacher network over the student network, channels with higher activation values in the teacher network indicate more important features for the detection task. By applying the Global Average Pooling (GAP) operation, we calculate the attention weights for each channel of the teacher network.

$$W_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} T'_{l,c,i,j} \tag{6}$$

After that, we use the softmax funciton to constrain the weight values between 1 and 10, obtaining the spatial feature weight W_c of each channel of the teacher network. Through the softmax, we can represent the relative relationship of different channel weight values. As shown in Fig. 2, we visualize the two channels with the lowest and highest weights. It can be observed that the visualized images with higher channel weight scores have higher activation values in the key and small object areas of detection, appearing as brighter areas in the image.

$$\mathcal{L}_{\text{DCSF-KD}} = \sum_{l=1}^{L} \frac{1}{N_l} \sum_{c=1}^{C} \sum_{w=1}^{H \times W} (T_{l,c,w} - S_{l,c,w})^2 \times W_c. \tag{7}$$

In summary, DCSF-KD adaptively focuses on key channel region features, and batch normalization on the FPN layer prevents the influence of extreme values in heterogeneous distillation, enhancing the stability and generality of knowledge distillation. The overall loss function for training the student network is given by:

$$\mathcal{L} = \mathcal{L}_{\text{GT}} + \alpha \mathcal{L}_{\text{DCSF-KD}}. \tag{8}$$

where $\mathcal{L}_{\text{GT}}$ is the detection training loss, and α is a hyperparameter used to balance the detection training loss and distillation loss.

Experiments

Experimental settings

To verify the effectiveness and robustness of our DCSF-KD method for object detection, we conducted experiments on various detection frameworks using the MS COCO dataset (Lin et al. 2014), which includes 120,000 training images and 5,000 test images. Standard data processing and default image settings were applied. We evaluated performance primarily using Mean Average Precision (mAP) and additional AP scores at different thresholds and scales such as AP_{50}, AP_{75}, AP_S, AP_M, and AP_L.

We adopted the training strategy from PKD (Cao et al. 2022) for initializing student models and setting experimental parameters. All experiments were conducted on 4 NVIDIA 3090Ti GPUs, processing two images per GPU, using mmdetection (Chen et al. 2019b) and mmrazor (Contributors 2021) frameworks based on PyTorch (Paszke et al. 2017).

DCSF-KD uses α to balance between detection loss ($\mathcal{L}_{\text{GT}}$) and adaptive channel spatial feature transfer ($\mathcal{L}_{\text{DCSF-KD}}$). We set $\alpha = 2$ for all two-stage models and $\alpha = 1$ for single-stage models. For training across all detectors, we used an SGD optimizer with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 0.0001 for 24 epochs. For 12-epoch experiments, a compound strategy with LinearLR and MultiStepLR optimizers was used, starting with a linear learning rate change from iterations 0-500 with a starting factor of 0.001, and reducing the learning rate by a factor of 0.1 during epochs 8-11.

Main Results

In this section, we compare five mainstream knowledge distillation methods for object detection. The FKD (Zhang and Ma 2020) and FGD (Yang et al. 2022) methods propose separating foreground and background to focus on transferring features of the foreground regions. GID (Dai et al. 2021) conducts targeted distillation by identifying valuable regions between the teacher and student networks. FRS (Zhixing et al. 2021) utilizes classification score weights as a criterion for the importance of features. PKD (Cao et al. 2022) concentrates on the correlation in the spatial dimension. The best results are highlighted in **bold**.

Homogeneous Detector Experimental Results. DCSF-KD can be applied to object detectors across different frameworks. Experiments were carried out on five popular detectors, including a two-stage detector (Faster RCNN (Ren et al. 2016)), an anchor-free detector (FCOS (Tian et al. 2019)), and anchor-based single-stage detectors (RetinaNet (Lin et al. 2020), RepPoints (Yang et al. 2019) , and TOOD (Feng et al. 2021)) . In comparison with mainstream methods, Tab. **??** illustrates the performance of DCSF-KD against current state-of-the-art methods on the MS COCO dataset, achieving the best results. The student network exhibited significant improvement in the AP metric by transferring the channel feature knowledge from the teacher network. For example, a 3.8% increase in mAP was obtained in the FasterRCNN detector using ResNet-50. These outcomes underscore the capability of DCSF-KD to effectively transfer critical regional knowledge in object detection networks, proving to be effective and generalizable in both single-stage and two-stage detectors.

Heterogeneous Detector Experimental Results. Most knowledge distillation algorithms to date are designed for homogeneous detectors, limiting their application to teacher detectors with identical architectures. Conversely, DCSF-KD is versatile for heterogeneous detector distillation, facilitating the effective transfer of FPN layer features from powerful teacher networks across diverse frameworks. This section elaborates on experiments conducted with high-performance object detection teacher networks, as illustrated in Tab. **??**. In comparison with Tab. **??**, it is observed that student networks exhibit enhanced performance when guided by teacher networks with superior accuracy and capabilities. For example, with teacher network architectures such as Mask-RCNN-Swin (He et al. 2017) and GFL-Res101 (Li et al. 2020b), the FCOS-Res50 student model achieves mAP scores of 43.9% and 43.5%, respectively. These outcomes affirm that DCSF-KD can efficaciously transfer knowledge from object detectors of disparate structures relative to the student model, and student networks yield better results under the tutelage of robust heterogeneous teacher networks, corroborating the efficacy and adaptability of DCSF-KD.

Method	schedule	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask RCNN-Swin (T)	3x+ms	48.2	69.8	52.8	32.1	51.8	62.7
Retina-Res50 (S)	2x	37.4	56.7	39.6	20.0	40.7	49.7
PKD(Cao et al. 2022)	2x	41.5(+4.1)	60.6	44.1	22.9	45.2	**56.4**
DCSF-KD(Ours)	2x	**41.9(+4.5)**	**61.2**	**44.7**	**23.2**	**46.1**	56.3
Mask RCNN-Swin (T)	3x+ms	48.2	69.8	52.8	32.1	51.8	62.7
FCOS-Res50 (S)	2x+ms	39.1	58.4	41.6	24.0	42.7	48.7
PKD(Cao et al. 2022)	2x+ms	43.9(+4.8)	62.3	47.5	27.2	48.0	57.1
DCSF-KD(Ours)	2x+ms	**44.1(+5.0)**	**62.6**	**47.9**	**27.4**	**48.5**	**57.2**
GFL-Res101 (T)	2x+ms	44.9	63.1	49.0	28.0	49.1	57.2
FCOS-Res50 (S)	2x+ms	39.1	58.4	41.6	24.0	42.7	48.7
PKD(Cao et al. 2022)	2x+ms	43.5(+4.4)	61.9	47.1	26.5	47.7	55.7
DCSF-KD(Ours)	2x+ms	**43.7(+4.6)**	**62.2**	**47.4**	**26.6**	**48.0**	**55.7**

Table 2: Results on the MS COCO dataset with heterogeneous teacher detectors. The best results are highlighted in **Bold**.

Detector Pairs	Methods	mAP	AP_S	AP_L
Retina-ResX101 Retina-Res50	CKD	40.6	22.6	54.1
	CWD	40.8	22.7	**55.3**
	DCSF-KD	**41.0**	**23.9**	54.8
RepPoints ResNeXt101 RepPoints Res50	CKD	41.6	23.8	54.0
	CWD	42.0	24.1	55.0
	DCSF-KD	**42.4**	**24.2**	**56.5**

Table 3: Ablation Study Results.

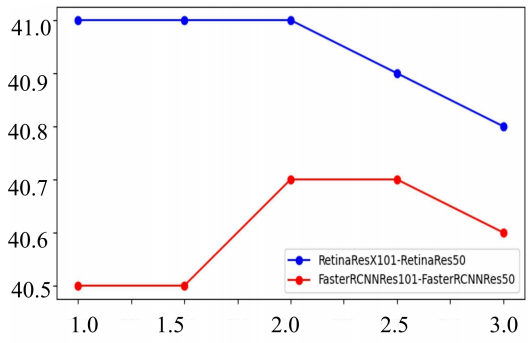

Figure 4: mAP results for RetinaResX101-RetinaRes50 and FasterRCNNRes101-FasterRCNNRes50 with varying α.

Ablation Studies

In this subsection, we explore the differences between a knowledge distillation method based solely on channel feature transfer (CKD) and DCSF-KD , which does not employ channel attention weights. Additionally, we compare these with the CWD (Shu et al. 2021), which is a mainstream channel-feature-based object detection knowledge distillation approach. The results in Tab. 3 indicate that incorporating adaptive channel attention weights, which focus on channels with higher activation values, our method can obtain mAP gains by 0.2 0.8 mAP, compared with CKD. Besides, our DCSF-KD method significantly surpasses the advanced CWD, establishing a new state-of-the-art for channel feature-based KD methods.

Sensitivity study of loss weight α. We employ the loss weight hyperparameter α to balance between the detection training loss and the distillation loss. This balance was explored within the context of the single-stage object detector teacher-student pair RetinaResX101-RetinaRes50 and the two-stage object detector teacher-student pair FasterRCNNRes101-FasterRCNNRes50 across varying settings of α. As depicted in Fig. 4, DCSF-KD demonstrates insensitivity to changes in the α parameter, with a mere 0.2% decrease in mAP in the least favorable scenario. This underscores the method's robust versatility and stability across different architectural frameworks of detectors, facilitating the identification of a relatively consistent α setting to maintain the stability of distillation outcomes.

Conclusion

In this paper, we introduce DCSF-KD, a simple yet effective knowledge distillation method aimed at enhancing the performance of lightweight object detectors. DCSF-KD leverages both spatial and channel domain feature knowledge, adaptively capturing channel features critical to the detection task from the teacher network without the need for designing complex weighting functions to differentiate important regions. DCSF-KD is applicable to both homogeneous and heterogeneous detector pairs, achieving faster convergence and introducing only a single hyperparameter, thus ensuring stable distillation efficacy. Our results indicate that DCSF-KD enhances distillation efficiency and achieves state-of-the-art performance. In the future, we plan to extend our method to end-to-end detectors (such as the DETR series) and other related domains, such as 3D object detection and instance segmentation.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China, under Grant (62302309, 62471310), Shenzhen Science and Technology Program (JCYJ20220818101014030) and the Guangdong Provincial Key Laboratory (Grant No. 2023B1212060076).

References

Cao, W.; Zhang, Y.; Gao, J.; Cheng, A.; Cheng, K.; and Cheng, J. 2022. Pkd: General distillation framework for object detectors via pearson correlation coefficient. *Advances in Neural Information Processing Systems*, 35: 15394–15406.

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

Chen, G.; Choi, W.; Yu, X.; Han, T.; and Chandraker, M. 2017. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30.

Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. 2019a. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4974–4983.

Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. 2019b. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*.

Cho, J. H.; and Hariharan, B. 2019. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4794–4802.

Contributors, M. 2021. Openmmlab model compression toolbox and benchmark.

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

Feng, C.; Zhong, Y.; Gao, Y.; Scott, M. R.; and Huang, W. 2021. Tood: Task-aligned one-stage object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3490–3499. IEEE Computer Society.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

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

Guo, J.; Han, K.; Wang, Y.; Wu, H.; Chen, X.; Xu, C.; and Xu, C. 2021a. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2154–2164.

Guo, J.; Han, K.; Wu, H.; Zhang, C.; Chen, X.; Xu, C.; Xu, C.; and Wang, Y. 2021b. Positive-unlabeled data purification in the wild for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2653–2662.

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

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

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

Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L. M.; and Zhang, L. 2022a. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13619–13627.

Li, G.; Li, X.; Wang, Y.; Zhang, S.; Wu, Y.; and Liang, D. 2022b. Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 1306–1313.

Li, Q.; Jin, S.; and Yan, J. 2017. Mimicking very efficient network for object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 6356–6364.

Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; and Yang, J. 2020a. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33: 21002–21012.

Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; and Yang, J. 2020b. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33: 21002–21012.

Lin, T.; Goyal, P.; Girshick, R. B.; He, K.; and Dollár, P. 2020. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2): 318–327.

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

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 740–755.

Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2022. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. In *The Tenth International Conference on Learning Representations*.

Park, S.; and Heo, Y. S. 2020. Knowledge distillation for semantic segmentation using channel and spatial correlations and adaptive cross entropy. *Sensors*, 20(16): 4616.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.

Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6): 1137–1149.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Shu, C.; Liu, Y.; Gao, J.; Yan, Z.; and Shen, C. 2021. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5311–5320.

Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. FCOS: Fully Convolutional One-Stage Object Detection. In *International Conference on Computer Vision*, 9626–9635.

Wang, J.; Chen, Y.; Zheng, Z.; Li, X.; Cheng, M.-M.; and Hou, Q. 2023. CrossKD: Cross-Head Knowledge Distillation for Dense Object Detection. *arXiv preprint arXiv:2306.11369*.

Wang, T.; Yuan, L.; Zhang, X.; and Feng, J. 2019. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4933–4942.

Wang, Y.; Huang, W.; Sun, F.; Xu, T.; Rong, Y.; and Huang, J. 2020. Deep multimodal fusion by channel exchanging. *Advances in neural information processing systems*, 33: 4835–4845.

Yang, Z.; Li, Z.; Jiang, X.; Gong, Y.; Yuan, Z.; Zhao, D.; and Yuan, C. 2022. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4643–4652.

Yang, Z.; Liu, S.; Hu, H.; Wang, L.; and Lin, S. 2019. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9657–9666.

Zagoruyko, S.; and Komodakis, N. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*.

Zhang, L.; and Ma, K. 2020. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*.

Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; and Li, S. Z. 2020. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9759–9768.

Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 11953–11962.

Zhixing, D.; Zhang, R.; Chang, M.; Liu, S.; Chen, T.; Chen, Y.; et al. 2021. Distilling object detectors with feature richness. *Advances in Neural Information Processing Systems*, 34: 5213–5224.

Zhou, Z.; Zhuge, C.; Guan, X.; and Liu, W. 2020. Channel distillation: Channel-wise attention for knowledge distillation. *arXiv preprint arXiv:2006.01683*.