中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2025)01-0161-12

论文引用格式:Song T, Zhang J T, Li W W, Zhao M F, Ran L, Ye D X, Yang Y C and Yue D H. 2025. Knowledge-enhanced feature editing reconstruction distillation. Journal of Image and Graphics, 30(01):0161-0172(宋涛, 张景涛, 李沩沩, 赵明富, 冉璐, 叶定兴, 杨贻晨, 岳岱衡. 2025. 知识增强的特征编辑重建蒸馏. 中国图象图形学报, 30(01):0161-0172)[DOI:10.11834/jig.230890]

知识增强的特征编辑重建蒸馏

宋涛[1,2],张景涛[1*],李沩沩[1],赵明富[1],冉璐[1],叶定兴[1],杨贻晨[1],岳岱衡[1]

1. 重庆理工大学光纤传感与光电检测重庆市重点实验室,重庆　400054; 2. 重庆工业大数据创新中心有限公司,重庆　400708

摘　要:**目的**　知识蒸馏作为一种有效可行的模型压缩方法现已在图像分类领域得到了广泛的研究。由于目标检测网络的复杂性加大了知识蒸馏的难度,现有应用于目标检测的知识蒸馏方法效果并不理想。为此,提出一种适用于目标检测任务的知识蒸馏方法,即基于知识增强的特征编辑重建蒸馏,实现对目标检测模型的有效压缩。**方法**　针对现有方法只在教师模型和学生模型的对应特征层之间进行蒸馏,无法充分利用教师模型中隐藏的"暗知识"的问题,通过空间注意力和通道注意力分别对教师特征进行自下而上和自上而下的多尺度特征融合以进行知识增强;针对当教师模型和学生模型能力差距过大时,学生模型无法理解教师模型的知识导致蒸馏性能受限的问题,通过将教师模型的部分特征作为先验知识融合到学生模型的特征中,以缩小教师模型和学生模型之间的表征能力差距,并删除学生模型特征的边缘、轮廓等细节信息,实现特征编辑,然后迫使学生模型利用剩余特征结合先验知识通过一个简单的卷积块来恢复删除的细节信息,进行特征重建,使学生模型在此过程中得到正向的反馈,从而学习到更好的特征。**结果**　实验在两个数据集上采用基于ResNet50(residual network)的RetinaNet(retina network)、Faster R-CNN(faster region-based convolutional neural network)、FCOS(fully convolutional one-stage object detection)3种不同类型的检测器与最新的4种方法进行了比较,VOC2007(visual object classes 2007)测试集的平均精度均值(mean average precision,mAP)对比基线分别提高了2.1%、2.7%和3.8%;NEU-DET(Northeastern University surface defect database)测试集的mAP对比基线分别提高了2.7%、2.6%和2.1%,均高于当前性能最优的算法。**结论**　本文所提出的方法能充分挖掘出教师模型的能力,有效提升学生模型的性能,同时适用于多种类型的目标检测器。

关键词:模型压缩;知识蒸馏;知识增强;特征重建;目标检测

Knowledge-enhanced feature editing reconstruction distillation

Song Tao[1,2], Zhang Jingtao[1*], Li Weiwei[1], Zhao Mingfu[1], Ran Lu[1], Ye Dingxing[1], Yang Yichen[1], Yue Daiheng[1]

1. Chongqing Key Laboratory of Fiber Optic Sensing and Photoelectric Detection, Chongqing University of Technology, Chongqing　400054, China; 2. Chongqing Industrial Big Data Innovation Center Co., Ltd., Chongqing　400708, China

Abstract: **Objective** In recent years, convolutional neural networks have had great potential in various fields of computer vision due to their excellent feature extraction ability. As the performance of the model increases, the size of the model

收稿日期:2024-01-10;**修回日期**:2024-06-18;**预印本日期**:2024-06-25

*通信作者:张景涛　jtzhang134@qq.com

基金项目:重庆市科技局基础与前沿研究计划资助项目(cstc2021jcyj-msxmX0348);重庆理工大学科研创新团队项目(2023TDZ014);国家博士后科学基金项目(2022M710543);重庆理工大学研究生创新项目(gzlcx20233080)

Supported by: Basic and Frontier Research Program of Chongqing Science and Technology Bureau(cstc2021jcyj-msxmX0348); National Science Foundation for Post-doctoral Scientists of China(2022M710543)

becomes increasingly bloated, and the large number of parameters makes its inference speed slow. Even with the help of GPU acceleration, the real-time demand of many application scenarios cannot be met. In addition, the occupied memory and storage space increase the cost of use. As a result, these large models are difficult to deploy and run on mobile devices or embedded platforms with limited arithmetic power and storage space, restricting their promotion. Therefore, how to compress large deep neural network models is a key issue. Knowledge distillation is a simple and effective model compression method. Unlike the idea of model pruning or parameter quantization, knowledge distillation is essentially a special model training method. It does not directly make changes to the model structure or parameters; rather, it compresses the volume without changing the model structure and parameters. It is also used in model training to learn the hard labels in the training set aside from the hard labels in the training set and guide the teacher's model by using the teacher's model classification. In addition to learning the hard labels in the training set, we use the classification output of the teacher's model as soft labels to guide the learning of the student model so that the hidden "dark knowledge" in the teacher's model, which is powerful but has a bloated network structure, is transferred to the student's model. The student's model has a relatively simple network structure and a smaller parameter volume. This knowledge transfer enables the student's model, which has fewer model parameters and a faster reasoning speed, to achieve comparable accuracy with the teacher's model, thus achieving model compression. However, the target detection task requires classifying the target and outputting the specific position of the target in the picture. This step is not possible just by learning the labels output from the teacher's model. Therefore, the traditional knowledge distillation method used for classification does not work well on the target detection task. Moreover, the network structure of the detector is more complex, so existing knowledge distillation methods based on the target detection task usually let the features of the student model directly learn the features of the teacher model for distillation instead of learning the labels of the teacher. Existing methods also have multiple limitations. Therefore, a new knowledge distillation method applicable to the target detection task is proposed, i. e. , feature editing reconstruction distillation based on knowledge enhancement, to achieve effective compression of the target detection model. **Method** Two modules are constructed to address two common problems of the current knowledge distillation methods for the target detection task: 1) knowledge enhancement module and 2) feature editing and reconstruction module. To address the problem in which the existing methods distill only between the corresponding feature layers of the teacher's model and the student's model, which cannot fully utilize the hidden "dark knowledge" in the teacher's model, the knowledge enhancement module is introduced to enhance the knowledge of the teacher's model through spatial attention and channel attention, respectively. Therefore, the knowledge enhancement module is introduced to perform bottom-up and top-down multi-scale feature fusion of teacher features through spatial attention and channel attention, respectively, for knowledge enhancement. As the performance of the teacher model continues to improve, the ability gap between students and teachers becomes increasingly larger, and the trend of performance enhancement of the student model gradually reaches saturation or even decreases. The feature edit reconstruction module is used to construct a new distillation paradigm to fuse some features of the teacher's model into the features of the student's model as a priori knowledge to narrow the representation capability gap between the teacher's model and the student's model. This module then randomly deletes detailed information such as edges and contours of the student model's features through a pixel-level mask to realize feature editing, and then forces the student model to use the remaining features in combination with the a priori knowledge through a simple convolutional block to recover the deleted detail information for feature reconstruction, and in the process of model learning to optimize the quality of the reconstructed feature maps, the student's original feature maps are back-propagated through feature reconstruction, thus learning features with stronger representational capabilities. **Result** Experiments were conducted based on three different types of detectors: retina network (RetinaNet), faster region-based convolutional neural network (Faster R-CNN), and fully convolutional one-stage object detection (FCOS), on the generalized target detection dataset visual object classes 2007 (VOC2007) and the steel surface defects dataset Northeastern University surface defect database (NEU-DET) by using the teacher's model, student's model of ResNet101-ResNet50. First, feature map visualization found that the feature map of the distilled detector responds significantly less to the background noise information and pays more attention to the foreground critical region. A comparison of the visualization of the detection results show the distilled detector significantly improves the situation of misdetection and omission. An evaluation of the detection results using the mAP metrics shows

that the mAP comparison baselines for the VOC2007 test set improved by 2. 1%, 2. 7%, and 3. 8%, respectively, and the mAP comparison baselines for the NEU-DET test set improved by 2. 7%, 2. 6%, and 2. 1%, respectively. **Conclusion** In this study, a new knowledge distillation method for target detection task is proposed, and experimental results show that this method can significantly improve the focus of the detector feature map on the key target region and reduce the interference of noise. Thus, the false detection and leakage rate is reduced, and the accuracy improvement is better than that of several state-of-the-art algorithms. The proposed method is suitable for general target detection datasets and specialized defect detection datasets, with good generalization performance, and can be applied to many types of detectors at the same time.

Key words: model compression; knowledge distillation; knowledge enhancement; feature reconstruction; target detection

0 引　言

卷积神经网络(convolutional neural network, CNN)在计算机视觉领域的图像分类(丁维龙 等, 2024)、目标检测(杜艳玲 等,2023)和语义分割(栗风永 等,2023)等方向得到了广泛应用和深入研究。然而,基于深度学习的方法在实际应用中仍然存在一些问题。首先,大而深的深度神经网络模型对运算的实时性要求很高,即使借助GPU(graphics processing unit)加速,仍然无法满足许多应用场景的实时性需求。其次,这些应用模型在使用深度卷积神经网络进行特征提取和处理时,会带来庞大的内存计算开销和内存占用,增加了存储和计算成本,导致这些深度模型难以在移动设备或嵌入式平台上部署运行,限制了其推广(高晗 等,2021)。因此,在不影响深度网络模型效果的前提下,压缩网络模型成为一个重要的研究问题。通过压缩网络模型,可以减小模型的尺寸和计算量,从而提高模型在资源受限环境下的使用效率和性能。目前已经有一些方法提出以用于网络模型的压缩和优化,如模型剪枝(郑香平和梁循,2022)、参数量化(冯烨 等,2020)和知识蒸馏(Hinton 等,2015)等。这些方法通过减少模型中的冗余参数、减小参数的表示精度或者利用已训练好的大模型的知识来实现网络的压缩。与模型剪枝或参数量化的思路不同的是,Hinton 等人(2015)提出的知识蒸馏并不是直接对模型体积进行改动和压缩,而是在不改变模型结构和参数的前提下,通过将教师模型的分类输出作为软标签来引导学生模型学习,从而将性能强大但网络结构臃肿的教师模型中隐藏的"暗知识"转移到网络结构相对简单、参数体积较小的学生模型中,以达到提升学生模型精度的目的。目前已经有许多工作(Xie 等,2020; Ge 等 2020; Pham 等,2021)通过知识蒸馏显著提高了学生分类网络的准确率。然而,目标检测任务既需要对目标进行分类,还需要输出目标在图像中的具体位置,仅仅通过学习教师模型输出的标签是不够的,这也是传统用于分类的知识蒸馏方法在目标检测任务上效果不理想的原因。

Chen 等人(2017)首先在目标检测任务上成功应用了知识蒸馏,通过自适应层将教师和学生的特征图进行维度匹配后直接蒸馏,并使用加权交叉熵损失来解决目标检测任务中类别严重失衡问题。FGFI(fine-grained feature imitation)(Wang 等,2019)方法和 TADF(task adaptive distillation framework)(Sun 等,2020)方法认为背景区域对蒸馏有害,分别利用细粒度掩码和高斯掩码将蒸馏区域控制在前景真实框附近。然而,Guo 等人(2021)认为背景区域同样包含着对蒸馏有用的"暗知识",单独蒸馏背景区域甚至可以达到与蒸馏前景相同的效果。Dai 等人(2021)提出了一种提取实例之间关系信息的蒸馏方法。Yang 等人(2022)在对前背景进行解耦的基础上,通过空间注意力和通道注意力对前景进行蒸馏,并提取了背景中的全局信息。然而,这些方法都只在教师和学生对应的特征层之间进行蒸馏,没有充分利用教师网络中隐藏的"暗知识"。另外,Cho 和 Hariharan (2019)发现,学生模型的性能与教师模型的性能并不总是正相关的,当学生和教师之间的表征能力差距过大时,学生将无法"理解"教师所传递的知识,从而导致性能饱和或下降。

针对现有目标检测蒸馏算法所存在的问题,本文提出基于知识增强的特征编辑重建蒸馏(feature editing reconstruction distillation based on knowledge enhancement, KE-FERD)。KE-FERD 由知识增强模块(knowledge enhancement module, KEM)和特征编辑重建模块(feature editing and reconstruction mod-

ule, FERM)组成。KEM分别通过空间注意力和通道注意力对教师特征进行自下而上和自上而下的多尺度特征融合,使学生网络可以跨层级学习到更加丰富的空间信息和语义信息。另外,KE-FERD没有让学生特征对教师特征进行直接的模仿。受掩码自编码器(masked autoencoders, MAE)(He等,2022)架构实现图像重生成的启发, FERM通过简单的卷积实现特征图的重建。首先将教师网络的部分特征作为先验知识与学生网络特征进行融合,以缩小教师和学生模型之间的表征能力差距,降低学生模型学习的难度。然后删除此特征图中的边缘、轮廓等细节信息,迫使学生网络结合先验知识通过一个由简单的卷积块组成的特征重建模块来恢复这些信息,在特征重建的过程中学生模型可以得到反馈从而学习到更好的特征。综上所述,本文的创新主要有:1)提出一种新的教师特征增强方法。传统的知识蒸馏只学习对应特征层的知识,本文通过知识增强模块的加入使学生网络可以跨层级地学习到更加丰富的空间信息和语义信息。2)提出一种新的蒸馏范式。传统的知识蒸馏直接对教师特征进行模仿,本文通过迫使学生模型结合教师的先验知识从空间细节被删除的特征图中恢复教师特征以获得正向的反馈来进行蒸馏。

1 本文方法

目前,针对目标检测的知识蒸馏方法大多是基于特征的蒸馏,分别提取教师网络和学生网络中间层的特征图,并迫使学生特征图对教师特征图进行模仿,蒸馏过程中学生特征为"输入",教师特征作为"目标"。此方法可以表示为

$$L_{\text{fea}}\left(F^{\text{t}},F^{\text{s}}\right)=\sum_{k=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(F_{k,i,j}^{\text{t}}-f\left(F_{k,i,j}^{\text{s}}\right)\right)^2 \tag{1}$$

式中,F^{t}和F^{s}分别表示从教师网络和学生网络的Neck层提取的特征图,C、H、W分别表示特征图的通道数以及高和宽。f表示对齐特征图通道数的自适应层,通常由卷积核大小为1×1的卷积层组成。现有的知识蒸馏方法只在对应特征层间进行蒸馏,没有充分利用教师模型中隐藏的"暗知识"。另外,随着教师模型性能的提升,学生模型的性能提升却逐渐饱和,甚至会出现性能下降。为了解决上述问题,本文提出KE-FERD。KE-FERD由KEM和FERM组成。训练时教师网络和学生网络同时进行前向传播,冻结教师网络,仅学生网络进行反向传播以更新参数。分别提取教师和学生模型backbone中的特征图,将教师模型backbone中各个stage的输出特征图通过KEM模块进行增强以生成新的特征,然后将学生特征输入FERM中进行特征重建。蒸馏的目标是使用学生特征尽可能地去恢复KEM生成的特征,整体结构如图1所示。

1.1 知识增强模块

目前基于ResNet的目标检测网络中,用于提取图像特征的主干网络(backbone)通常包括stage1—stage4的4个阶段,从底层到顶层所输出的特征图尺寸逐级减小,通道数逐级增加,这意味着底层特征具有更大的图像分辨率,包含更多的空间位置信息而语义信息较少;高层特征具有更复杂的维度表示,包含更抽象的语义信息但由于特征图分辨率较低使得小目标信息丢失。现有的目标检测知识蒸馏方法大多只在教师和学生的对应层之间进行特征蒸馏,无法充分利用教师特征中隐藏的"暗知识"。为了同时兼顾小目标的检测率和分类识别准确度,本文提出了KEM用于增强知识蒸馏过程中的教师特征,希望学生网络的底层特征学习到教师网络的高层特征语义信息,学生网络的高层特征学习到教师网络底层的空间位置信息,从而增强学生网络特征的表征能力,并提高教师网络中知识的利用率。如图1所示,本文分别构建了自底向上和自顶向下的两条跨层级特征融合路径。与CBAM(convolutional block attention module)(Woo等,2018)不同的是,KEM没有串联式地提取同一特征图的空间注意力和通道注意力,而是在自底向上和自顶向下的路径中分别通过空间注意力和通道注意力引导特征融合。通过类似DANet(dual attention network)(Fu等,2019)中的方法提取空间注意力和通道注意力,如图2所示。

整个特征融合的步骤如下:

1)自底向上。首先从教师的主干网络中提取特征C_2^{t},C_3^{t},C_4^{t},C_5^{t},中间特征M_2^{t}由C_2^{t}直接生成,同时将C_2^{t}进行下采样后与C_3^{t}通过空间注意力引导融合生成M_3^{t},同理依次生成M_4^{t},M_5^{t}。此过程可以表示为

$$M_2^{\text{t}}=C_2^{\text{t}} \tag{2}$$

$$M_3^{\text{t}}=s_{\text{att}}(C_3^{\text{t}})+s_{\text{att}}(down_{\text{sam}}(M_2^{\text{t}})) \tag{3}$$

$$M_4^{\text{t}}=s_{\text{att}}(C_4^{\text{t}})+s_{\text{att}}(down_{\text{sam}}(M_3^{\text{t}})) \tag{4}$$

$$M_5^{\text{t}}=s_{\text{att}}(C_5^{\text{t}})+s_{\text{att}}(down_{\text{sam}}(M_4^{\text{t}})) \tag{5}$$

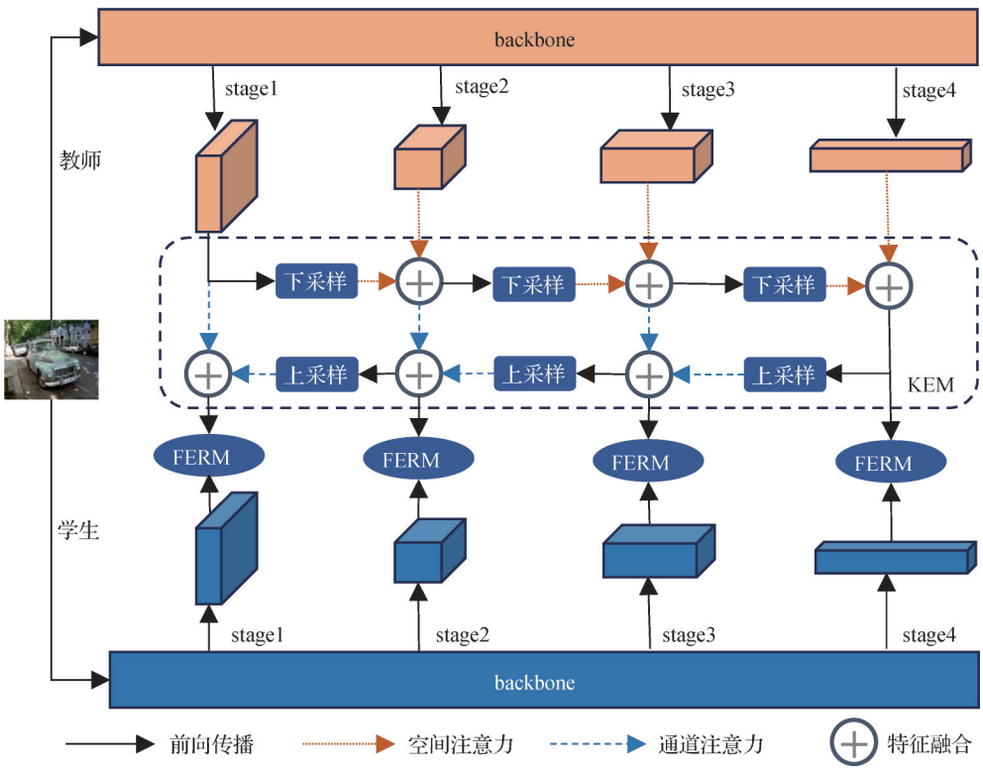


图 1　KE-FERD（包括 KEM 和 FERM 两个模块）

Fig. 1　KE-FERD (including both KEM and FERM modules)

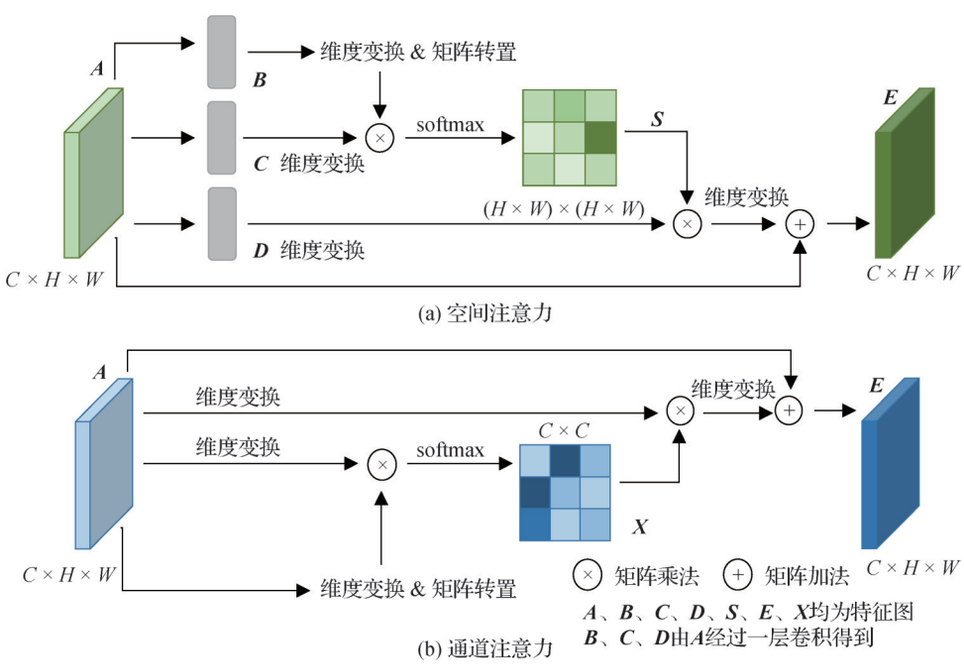


图 2　空间注意力和通道注意力的提取方式

Fig. 2　How spatial attention and channel attention are extracted ((a) spatial attention; (b) channel attention)

式中，$down_{sam}$表示下采样，通过卷积核大小为 3 × 3 的卷积层实现，s_{att}表示空间注意力。这条融合路径使高层特征获取到了更多来自底层的空间位置信息。

2）自顶向下。M_5^t直接生成F_5^k，然后将M_5^t通过 1 × 1 卷积与M_4^t进行通道对齐，再进行上采样后与M_4^t通过通道注意力引导融合生成F_4^k，同理依次生成F_3^k，F_2^k，此过程可以表示为

$$F_5^k = M_5^t \tag{6}$$

$$F_4^{\mathrm{k}} = c_{\mathrm{att}}(M_4^{\mathrm{t}}) + c_{\mathrm{att}}(up_{\mathrm{sam}}(f_{\mathrm{align}}(M_5^{\mathrm{t}}))) \tag{7}$$

$$F_3^{\mathrm{k}} = c_{\mathrm{att}}(M_3^{\mathrm{t}}) + c_{\mathrm{att}}(up_{\mathrm{sam}}(f_{\mathrm{align}}(M_4^{\mathrm{t}}))) \tag{8}$$

$$F_2^{\mathrm{t}} = c_{\mathrm{att}}(M_2^{\mathrm{t}}) + c_{\mathrm{att}}(up_{\mathrm{sam}}(f_{\mathrm{align}}(M_3^{\mathrm{t}}))) \tag{9}$$

式中，f_{align}表示用于特征图通道对齐的自适应层，up_{sam}表示上采样，通过最近邻插值法实现，c_{att}表示通道注意力。这条融合路径使底层特征获取到了更多来自高层的语义信息。

1.2 特征编辑重建模块

与现有方法不同的是，本文没有让学生特征对教师特征进行直接模仿。FERM分为特征编辑(feature editing，FE)和特征重建(feature reconstruction，FR)两个部分，如图3所示。

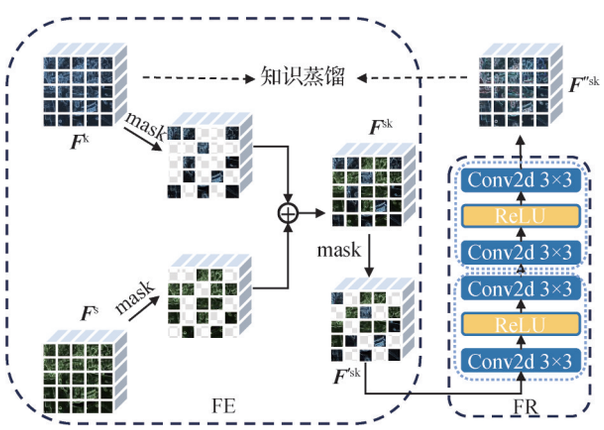


图3 FERM，包含FE和FR两个部分

Fig. 3 FERM with FE and FR components

在特征编辑阶段，首先会通过一对互补的二值化掩码对学生特征和KEM的输出特征进行处理，随后将两者结合。这意味着教师特征不仅作为“目标”供学生模仿，而且作为一种先验知识加入到学生特征中成为“输入”的一部分。这么做的灵感来源于现实世界教师的教学方式：在学习较难任务时，事先给予学生部分提示信息。先验知识的加入可以有效缩小教师和学生模型的差距，改善强教师教弱学生导致蒸馏性能不佳的问题。然后再通过随机像素屏蔽的方式删除此特征图中的边缘、轮廓等细节信息，迫使其在特征重建阶段恢复这些信息。特征重建部分由简单的卷积块堆叠而成，期望在此过程中学生特征可以得到良好的反馈信息，从而使得学生网络生成更好的特征。FERM的具体步骤如下：

1)获取KEM模块的输出特征F^{k}以及从学生网络backbone中提取特征图$\boldsymbol{F}^{\mathrm{s}}$。

2)通过一个二值化随机掩膜M_{p}来选择先验知识，具体为

$$M_{\mathrm{p}}(i,j) = \begin{cases} 1 & R_{(i,j)} < \alpha \\ 0 & \text{其他} \end{cases} \tag{10}$$

式中，α是控制先验知识比率的超参数。i和j代表特征图中像素的横坐标和纵坐标。

3)生成一个互补的掩膜M_{o}，具体为

$$M_{\mathrm{o}(i,j)} = 1 - M_{\mathrm{p}(i,j)} \tag{11}$$

4)将学生特征和KEM模块的输出特征分别通过各自的掩膜处理后进行融合得到$\boldsymbol{F}^{\mathrm{sk}}$，具体为

$$\boldsymbol{F}^{\mathrm{sk}} = \boldsymbol{F}^{\mathrm{k}} M_{\mathrm{p}}(i,j) + \boldsymbol{F}^{\mathrm{s}} M_{\mathrm{o}}(i,j) \tag{12}$$

5)再次使用式(10)的方法，通过随机二值化掩膜对$\boldsymbol{F}^{\mathrm{sk}}$进行处理，得到特征重建模块的输入特征$\boldsymbol{F}'^{\mathrm{sk}}$，具体为

$$M_{(i,j)} = \begin{cases} 0 & R_{(i,j)} < \beta \\ 1 & \text{其他} \end{cases} \tag{13}$$

$$\boldsymbol{F}'^{\mathrm{sk}} = \boldsymbol{F}^{\mathrm{sk}} \times M_{(i,j)} \tag{14}$$

式中，β是控制特征图$\boldsymbol{F}^{\mathrm{sk}}$中像素屏蔽比例的超参数。学生网络基于输入特征$\boldsymbol{F}'^{\mathrm{sk}}$重建$\boldsymbol{F}^{\mathrm{k}}$，特征重建模块由简单的卷积层、BN(batch normalization)层以及非线性激活层堆叠而成，单个卷积块的操作可以表示为

$$\varphi(\boldsymbol{F}'^{\mathrm{sk}}) = \mathrm{conv}(\mathrm{ReLU}(\mathrm{BN}(\mathrm{conv}(\boldsymbol{F}'^{\mathrm{sk}})))) \tag{15}$$

则特征重建部分的输出$\boldsymbol{F}''^{\mathrm{sk}}$可表示为

$$\boldsymbol{F}''^{\mathrm{sk}} = \varphi(\varphi(F'^{\mathrm{sk}})) \tag{16}$$

式中，conv表示3×3的卷积层。

1.3 目标函数

为了更好地融合KEM和FERM，重新设计了知识蒸馏的目标损失函数，此时的目标函数可以表示为

$$L_{\mathrm{dis}}(\boldsymbol{F}^{\mathrm{k}}, \boldsymbol{F}''^{\mathrm{sk}}) = \sum_{k=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(F_{k,i,j}^{\mathrm{k}} - F''^{\mathrm{sk}}_{k,i,j}\right)^2 \tag{17}$$

式中，C、H和W分别是特征图的通道数、高和宽。$\boldsymbol{F}^{\mathrm{k}}$表示KEM增强后的教师特征，$\boldsymbol{F}''^{\mathrm{sk}}$表示FERM模块的输出。由此可以得出学生网络训练的总损失函数，具体为

$$L_{\mathrm{all}} = L_{\mathrm{origin}} + \lambda \times L_{\mathrm{dis}} \tag{18}$$

式中，L_{origin}是目标检测任务的原始训练损失，λ是平衡蒸馏损失的超参数。

2 实验结果与分析

2.1 数据集与评价指标

实验数据集选取的是典型通用目标检测数据集

VOC2007(visual object classes 2007)，以及由东北大学发布的钢带表面缺陷数据集 NEU-DET(Northeastern University surface defect database)。VOC2007 共包含 9 963 幅图像，涵盖日常场景中 20 个常见的类别，参考 Faster R-CNN(Ren 等，2015)中的训练模式进行数据集划分，其中训练集 5 011 幅，测试集 4 952 幅，在测试集上评估所提方法的性能。NEU-DET 收集了热轧钢带 6 种典型的表面缺陷，包括：氧化铁皮压入(rolled-in scale)、斑块(patches)、开裂(crazing)、点蚀(pitted surface)、夹杂(inclusion)和划痕(scratches)。每个缺陷类别包含 300 个样本，共 1 800 幅图像。按照 8∶2 的方式将数据集划分为训练集、测试集进行训练。所有实验均采用在交并比(intersection over union，IoU)阈值为 0.5 时每类的标准平均精度(average precision，AP)和所有类别的平均精度均值(mean average precision，mAP)作为评价指标。

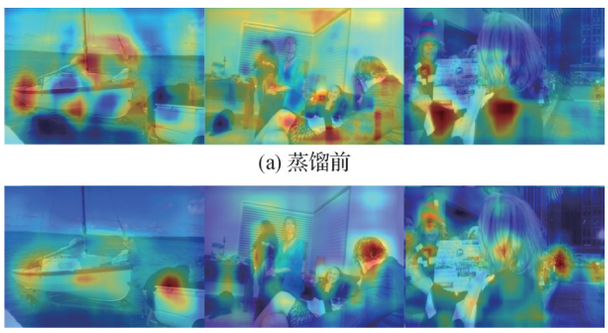
2.2 实验环境及设置

实验在 Ubantu 20.04 操作系统上进行，显卡配置为 Nvidia RTX 3090(24 GB)×4，cuda 11.3，cudnn 8.2，实验采用基于 Pytorch 1.1.3 的目标检测工具箱 MMDetection。为了验证所提方法的有效性，在 3 种不同类型的检测器上进行了实验，分别是单阶段检测器 RetinaNet(Lin 等，2017)、两阶段检测器 Faster R-CNN(Ren 等，2015)以及单阶段无锚框检测器 FCOS(fully convolutional one-stage object detection)(Tian 等，2019)。在单阶段检测器上实验超参数设置为 $\alpha = 0.5$，$\beta = 0.7$，$\lambda = 1 \times 10^{-6}$，两阶段检测器实验的超参数设置为 $\alpha = 0.3$，$\beta = 0.5$，$\lambda = 1 \times 10^{-7}$。训练过程中采用随机梯度下降(stochastic gradient descent，SGD)优化器，momentum 为 0.9，weight decay 为 0.000 1，初始学习率为 1E-3，采用余弦衰减策略训练 24 个 epoch。

2.3 实验结果

为了验证所提方法的有效性，本文将 ResNet101—ResNet50 作为教师—学生模型的骨干网络，以 RetinaNet(Lin 等，2017)为例，对检测器 Neck 层的特征图以及检测结果进行可视化，如图 4 和图 5 所示。

从图 4 可以看出，未蒸馏的检测器特征图注意力比较分散，对于背景投入了过多的关注；相比之下蒸馏后的检测器减少了对背景的关注，注意力更多地集中在人的面部等关键特征区域。反映在检测结果上的差别，如图 5 所示，第 1 幅图像中，蒸馏前的检测器将右下角的汽车尾部误识别为船；第 2 幅图像中没有识别到右下角被遮挡的人；第 3 幅图像中将左下角的手误识别为人，并且没有检测到右后方的盆栽。相比之下，使用本文方法蒸馏的检测器的检测框更加精准，目标召回率和检测置信度也更高。

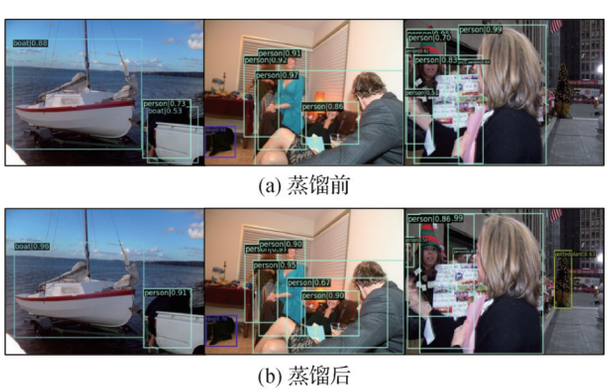


(a) 蒸馏前

(b) 蒸馏后

图 4 VOC2007 上 RetinaNet-R50 蒸馏前后特征图可视化对比

Fig. 4 Visual comparison of RetinaNet-R50 feature maps before and after distillation on VOC2007 dataset ((a) before distillation; (b) after distillation)



(a) 蒸馏前

(b) 蒸馏后

图 5 VOC2007 上 RetinaNet-R50 蒸馏前后检测结果对比

Fig. 5 Comparison of detection results before and after RetinaNet-R50 distillation on VOC2007 dataset ((a) before distillation; (b) after distillation)

继续在两阶段检测器 Faster R-CNN(Ren 等，2015)和无锚框检测器 FCOS(Tian 等，2019)上进行实验，并与当前主流的多种目标检测知识蒸馏算法进行对比，如 FGFI(fine-grained feature imitation)(Wang 等，2019)、GID(general instance distillation)(Dai 等，2021)、FGD(focal and global knowledge distillation)(Yang 等，2022)和 CWD(channel-wise knowledge distillation)(Shu 等，2021)。实验结果如表 1 所示，其中 S 和 T 分别表示学生和教师模型。从表 1 结果可以看出，KE-FERD 在 3 种类型目标检测

器上的表现都优于其他主流方法，学生模型的准确率都获得了较大的提升。为了便于实验结果的展示，表中只提供了部分VOC类别的检测结果。经过本文方法蒸馏之后，RetinaNet-Res50、Faster R-CNN-Res50和FCOS-ResNet50在VOC2007数据集上的mAP分别达到了82.8%、85.7%和75.9%，mAP相比基线分别获得了2.1、2.7和3.8的增益，高于当前性能最优的算法。

表1　VOC2007数据集上不同蒸馏算法的性能比较

Table 1　Performance comparison of different distillation algorithms on VOC2007 dataset

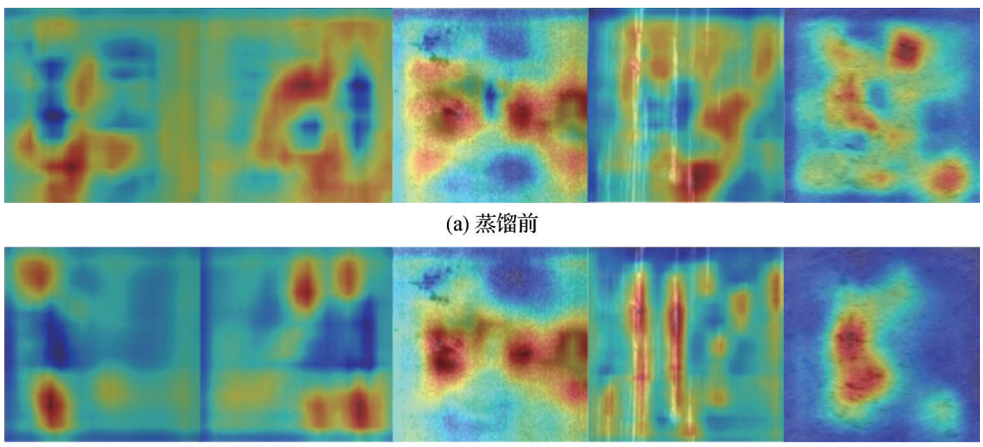
/%

教师	学生	mAP	AP					
			Aeroplane	Bicycle	Boat	Bus	Chair	Person
RetinaNet-Res101(T)	Teacher without student	83.0	89.4	88.6	74.8	82.2	71.7	87.4
	RetinaNet-Res50(S)	80.7	86.6	86.3	73.6	83.7	69.5	85.9
	FGFI	81.5(+0.8)	83.4	86.7	73.5	81.9	71.2	86.7
	GID	82.2(+1.5)	87.6	87.3	74.6	84.7	69.0	86.6
	FGD	82.3(+1.6)	86.7	87.2	74.3	83.8	72.5	86.7
	CWD	82.6(+1.9)	86.2	86.5	73.7	85.1	71.3	87.2
	本文	**82.8(+2.1)**	85.0	87.9	72.8	88.6	70.8	87.5
Faster R-CNN-Res101(T)	Teacher without student	86.2	90.2	89.0	80.8	89.6	78.2	89.2
	Faster R-CNN-Res50(S)	83.0	88.3	87.1	73.1	88.8	75.0	88.4
	FGFI	83.7(+0.7)	86.7	88.2	75.9	87.9	72.3	88.6
	GID	84.4(+1.4)	89.1	89.0	75.7	89.4	72.6	88.9
	FGD	84.4(+1.4)	88.2	88.1	76.4	89.3	73.5	87.3
	CWD	85.3(+2.3)	89.4	88.9	77.2	89.7	75.1	89.0
	本文	**85.7(+2.7)**	89.9	88.7	79.7	89.6	79.3	89.1
FCOS-ResNet101 (T)	Teacher without student	76.2	86.8	83.2	64.7	83.9	58.0	85.3
	FCOS-ResNet50 (S)	72.1	83.7	79.5	61.3	83.2	55.4	83.3
	FGFI	74.3(+2.2)	83.9	81.3	62.0	83.1	57.1	83.4
	GID	74.9(+2.8)	84.6	82.1	63.7	82.3	57.5	84.2
	FGD	75.2(+3.1)	84.3	81.8	61.7	83.4	56.6	84.0
	CWD	75.7(+3.6)	85.7	82.5	61.4	83.1	57.6	83.9
	本文	**75.9(+3.8)**	86.9	82.5	63.8	83.2	56.9	84.7

注：加粗字体表示不同目标检测器上mAP的最优结果，括号内为增益。

另外，为了证明本文所提出的方法对其他领域的数据集也具有泛化性，在NEU-DET数据集上进行同样的实验，与VOC2007上的实验保持相同的设置。再次以RetinaNet（Lin等，2017）为例对检测器Neck层的特征图以及检测结果进行可视化，如图6和图7所示。从图6可以看出，经过蒸馏后的检测器增强了关键位置特征的响应，而减少了对背景噪声的干扰。在工业领域，召回率是一个非常重要的指标，从图7的结果反映出蒸馏后的检测器召回率得到提升，误检和漏检的情况得到有效改善。

继续在两阶段检测器Faster R-CNN（Ren等，2015）和无锚框检测器FCOS（Tian等，2019）上进行实验，并与当前主流的多种目标检测知识蒸馏算法进行对比，如FGFI（Wang等，2019）、GID（Dai等，2021）、FGD（Yang等，2022）、CWD（Shu等，2021）。实验结果如表2所示，其中S和T分别表示学生和教

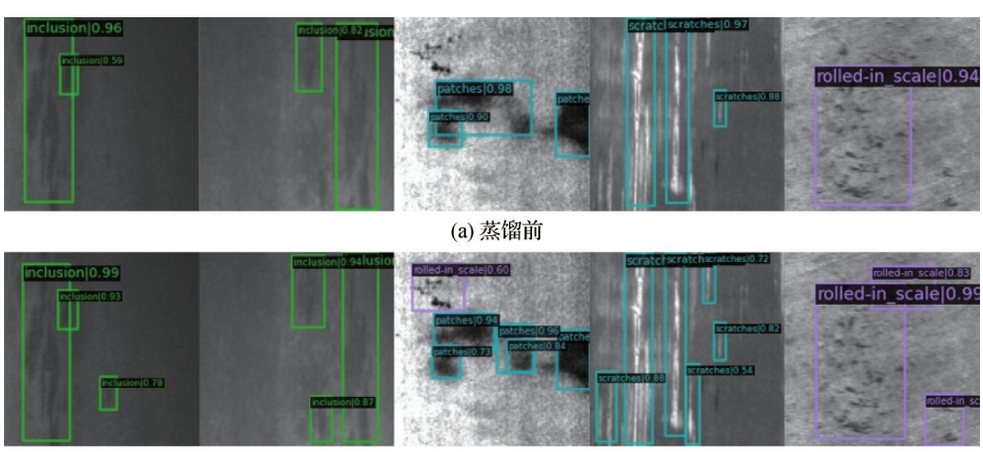


(a) 蒸馏前

(b) 蒸馏后

图 6　NEU-DET 上 RetinaNet-R50 蒸馏前后特征图可视化对比

Fig. 6　Comparison of feature map visualization before and after RetinaNet-R50 distillation on NEU-DET ((a) before distillation; (b) after distillation)



(a) 蒸馏前

(b) 蒸馏后

图 7　NEU-DET 上 RetinaNet-R50 蒸馏前后检测结果

Fig. 7　Detection results of RetinaNet-R50 before and after distillation on NEU-DET ((a) before distillation; (b) after distillation)

师模型。

从表 2 可以看出，本文提出的蒸馏方法对背景噪声干扰强、目标形态变化大的缺陷检测数据集也同样有效。经过蒸馏，以 ResNet50 为骨干网络的 RetinaNet、Faster R-CNN 和 FCOS 在 NEU-DET 数据集上的 mAP 分别达到 75. 7%、75. 3% 和 74. 5%，mAP 相比基线分别获得 2. 7、2. 6 和 2. 1 的增益，同样高于当前性能最优的算法。这说明本文提出的方法不仅能够在不同类型的检测器上使用，而且同时适用于典型目标检测场景和工业缺陷检测场景。

2. 4　消融实验与分析

为了进一步验证所提出方法的有效性，将 RetinaNet-Res101 作为教师模型，RetinaNet-Res50 作为学生模型在 NEU-DET 数据集上继续进行一系列消融实验并分析。

2. 4. 1　模块消融

本小节实验的目的是为了进一步明确所提方法各部分模块在蒸馏算法中的贡献，实验结果如表 3 所示。从表 3 可以看出，单独使用 KEM 对学生模型的提升有限，mAP 提升仅为 0. 5%，说明此时的教师和学生之间能力差距过大，以至于学生难以完全理解教师所传授的知识。此外，同时使用 FERM 中的 FE 和 FR 对比单独使用 FR 并没有很大的提升 (+0. 3)，这是因为此时对于没有经过 KEM 进行知识增强的教师所传递的知识，学生能较轻松地学习，所以此时在 FE 中加入的提示信息对学生的帮助不大。而当同时使用 KEM 和 FERM 时，mAP 却获得了 2. 7% 的大幅提升。这说明在 FE 中加入的先验知识

表2　NEU-DET数据集上不同蒸馏算法的性能比较

Table 2　Performance comparison of different distillation algorithms on the NEU-DET dataset

/%

教师	学生	mAP	AP					
			crazing	inclusion	patches	pitted surface	rolled-in_scale	scratches
RetinaNet-Res101(T)	Teacher without student	76.6	52.1	74.2	86.9	83.6	72.2	90.7
	RetinaNet-Res50(S)	73.0	48.2	68.5	84.9	83.9	69.0	83.2
	FGFI	74.0(+1.0)	**51.1**	**71.6**	88.3	84.4	71.9	76.9
	GID	74.4(+1.4)	48.4	69.8	87.7	83.0	72.6	85.0
	FGD	74.7(+1.7)	46.3	71.5	88.6	**85.9**	70.7	85.4
	CWD	75.2(+2.2)	47.9	71.3	**88.9**	85.6	71.3	86.4
	本文	**75.7(+2.7)**	48.6	70.2	88.3	85.6	**74.5**	**87.2**
Faster R-CNN-Res101(T)	Teacher without student	75.8	45.6	76.7	87.3	86.3	70.0	89.0
	Faster R-CNN-Res50(S)	72.7	42.7	74.9	83.5	79.8	68.2	87.3
	FGFI	73.5(+0.8)	44.1	74.4	87.8	84.1	62.3	88.5
	GID	74.1(+1.4)	43.7	75.3	88.8	83.1	62.9	90.5
	FGD	74.7(+2.0)	41.2	76.5	88.0	85.4	67.8	89.4
	CWD	74.9(+2.2)	43.5	71.5	87.6	86.0	71.0	89.5
	本文	**75.3(+2.6)**	43.8	76.0	87.4	83.8	70.8	89.8
FCOS-ResNet101 (T)	Teacher without student	74.6	46.8	78.1	89.6	85.0	63.0	85.2
	FCOS-ResNet50 (S)	72.4	41.7	74.9	85.5	81.7	64.9	85.8
	FGFI	73.1(+0.7)	42.7	75.6	85.7	84.3	67.3	82.8
	GID	73.4(+1.0)	43.1	76.3	87.1	80.9	68.3	84.9
	FGD	73.9(+1.5)	41.2	77.6	86.9	81.8	69.0	86.5
	CWD	74.2(+1.8)	41.9	77.5	87.4	82.9	68.6	87.1
	本文	**74.5(+2.1)**	42.8	77.8	87.5	84.0	66.0	88.8

注：加粗字体表示不同目标检测器上mAP的最优结果。

给学生模型提供了部分提示信息，从而降低了学生学习复杂知识的难度，解决了能力较差的学生无法理解复杂知识的问题。

表3　KE-FERD各模块消融实验

Table 3　KE-FERD ablation experiments by module

KEM	FE	FR	mAP/%
√	–	–	73.5(+0.5)
–	–	√	74.5(+1.5)
–	√	√	74.8(+1.8)
√	√	√	75.7(+2.7)

注："√"表示使用对应模块，"–"表示不使用对应模块。

2.4.2　特征重建部分的结构分析

图3所示的特征重建结构通过卷积块的堆叠来实现，本小节讨论了不同卷积块结构以及堆叠数量对检测精度的影响，结果如表4所示。

结果表明，当只有一个卷积块时精度最低，两个卷积块时精度最高，一个卷积块由卷积层、BN层以及非线性激活层组成。同时使用3×3的卷积核相比5×5的卷积核可以获得更低的计算量、更高的性能。

2.4.3　超参数分析

本文使用式(13)中的β来控制学生特征图中像素屏蔽的比例，本节讨论了不同的像素屏蔽比例对

表 4　卷积块结构及数量对精度的影响

Table 4　Effect of convolution block structure and number on accuracy

卷积块	卷积核	mAP/%
1	3 × 3	75.0
2	3 × 3	75.7
3	3 × 3	75.5
2	5 × 5	75.3

蒸馏效果的影响，结果如图 8 所示。从图 8 可以看出，当$\beta < 0.7$时，随着像素屏蔽比例的增加，学生网络可以学习到更多来自教师特征图的全局特征，使得精度逐渐提升。然而当β太大时，如$\beta = 0.9$，学生特征的屏蔽比例过高，导致特征图的语义信息太差，无法重建出完整的教师特征图，从而精度出现明显的下降。因此，为获得更好的性能表现，在保留学生特征图基本语义信息的前提下，应尽可能提高像素屏蔽比例。

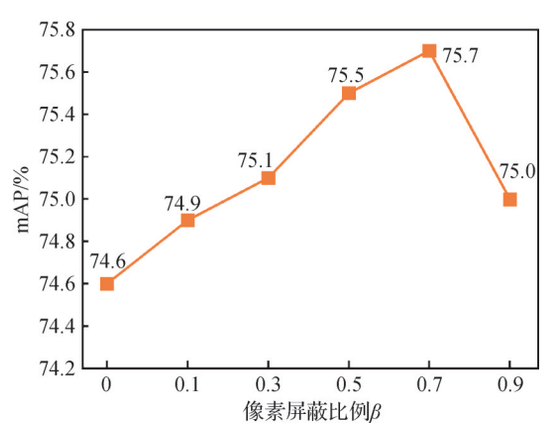


图 8　不同β取值对精度的影响

Fig. 8　Influence of different values of β on accuracy

3　结　论

现有的知识蒸馏方法只在对应特征层进行蒸馏，没有充分利用教师网络中隐藏的“暗知识”，并且当教师和学生模型能力差距过大时蒸馏效果不理想，由此提出了 KE-FERD，通过 KEM 利用空间注意力和通道注意力对教师特征进行自下而上和自上而下的多尺度特征融合，使学生网络可以跨层级学习到更加丰富的空间信息和语义信息；通过在 FERM 中引入教师部分特征作为先验知识，缩小了教师和学生模型之间的表征能力差距，降低了学生网络学习的难度。特征图可视化实验结果表明，本文提出的蒸馏方法对学生网络关键位置的特征响应进行了增强，从而减少了背景噪声造成的干扰。对比实验结果表明，本文方法的性能优于目前的性能最佳方法，且适用于单阶段和两阶段以及无锚框检测器，在通用目标检测数据集和工业缺陷检测领域数据集上均有较好的效果。所提出的方法仅针对检测器的主干网络特征进行蒸馏，如何充分利用教师网络颈部以及检测头部分存在的知识有待进一步研究。

参考文献（References）

Chen G B, Choi W, Yu X, Han T and Chandraker M. 2017. Learning efficient object detection models with knowledge distillation//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 742-751

Cho J H and Hariharan B. 2019. On the efficacy of knowledge distillation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 4793-4801 [DOI: 10.1109/ICCV.2019.00489]

Dai X, Jiang Z R, Wu Z, Bao Y P, Wang Z C, Liu S and Zhou E J. 2021. General instance distillation for object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 7838-7847 [DOI: 10.1109/CVPR46437.2021.00775]

Ding W L, Zhu F L, Zheng K and Jia X P. 2024. Classification of breast pathological images based on multiscale information interaction and fusion. Journal of Image and Graphics, 29(4): 1085-1099 (丁维龙, 朱峰龙, 郑魁, 贾秀鹏. 2024. 多尺度信息交互与融合的乳腺病理图像分类. 中国图象图形学报, 29(4): 1085-1099) [DOI: 10.11834/jig.221178]

Du Y L, Wu T Y, Chen K, Chen G and Song W. 2023. Small object detection for ocean eddies using contextual information and attention mechanism. Journal of Image and Graphics, 28(11): 3509-3519 (杜艳玲, 吴天宇, 陈括, 陈刚, 宋巍. 2023. 融合上下文和注意力的海洋涡旋小目标检测. 中国图象图形学报, 28(11): 3509-3519) [DOI: 10.11834/jig.220944]

Feng Y, Zhang S F and Wu X F. 2020. A light-weighted SSD network design for object detection. Journal of Signal Processing, 36(5): 756-762 (冯烨, 张索非, 吴晓富. 2020. 面向目标检测的SSD网络轻量化设计研究. 信号处理, 36(5): 756-762) [DOI: 10.16798/j.issn.1003-0530.2020.05.015]

Fu J, Liu J, Tian H J, Li Y, Bao Y J, Fang Z W and Lu H Q. 2019. Dual attention network for scene segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recogni-

tion (CVPR). Long Beach, USA: IEEE: 3141-3149 [DOI: 10.1109/CVPR.2019.00326]

Gao H, Tian Y L, Xu F Y and Zhong S. 2021. Survey of deep learning model compression and acceleration. Journal of Software, 32(1): 68-92 (高晗, 田育龙, 许封元, 仲盛. 2021. 深度学习模型压缩与加速综述. 软件学报, 32(1): 68-92) [DOI: 10.13328/j.cnki.jos.006096]

Ge Y X, Chen D P and Li H S. 2020. Mutual mean-teaching: pseudo label refinery for unsupervised domain adaptation on person re-identification [EB/OL]. [2023-12-20]. https://arxiv.org/pdf/2001.01526.pdf

Guo J Y, Han K, Wang Y H, Wu H, Chen X H, Xu C J and Xu C. 2021. Distilling object detectors via decoupled features//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 2154-2164 [DOI: 10.1109/CVPR46437.2021.00219]

He K M, Chen X L, Xie S N, Li Y H, Dollár P and Girshick R. 2022. Masked autoencoders are scalable vision learners//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 15979-15988 [DOI: 10.1109/CVPR52688.2022.01553]

Hinton G, Vinyals O and Dean J. 2015. Distilling the knowledge in a neural network [EB/OL]. [2023-12-20]. https://arxiv.org/pdf/1503.02531.pdf

Li F Y, Ye B and Qin C. 2023. Mutual attention mechanism-driven lightweight semantic segmentation network. Journal of Image and Graphics, 28(7): 2068-2080 (栗风永, 叶彬, 秦川. 2023. 互注意力机制驱动的轻量级图像语义分割网络. 中国图象图形学报, 28(7): 2068-2080) [DOI: 10.11834/jig.211127]

Lin T Y, Goyal P, Girshick R, He K M and Dollár P. 2017. Focal loss for dense object detection//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 2999-3007 [DOI: 10.1109/ICCV.2017.324]

Pham H, Dai Z H, Xie Q Z and Le Q V. 2021. Meta pseudo labels//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 11552-11563 [DOI: 10.1109/CVPR46437.2021.01139]

Ren S Q, He K M, Girshick R and Sun J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks//Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS). Montreal, Canada: MIT Press: 91-99

Shu C Y, Liu Y F, Gao J F, Yan Z and Shen C H. 2021. Channel-wise knowledge distillation for dense prediction//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 5291-5300 [DOI: 10.1109/ICCV48922.2021.00526]

Sun R Y, Tang F H, Zhang X P, Xiong H K and Tian Q. 2020. Distilling object detectors with task adaptive regularization [EB/OL]. [2023-12-20]. https://arxiv.org/pdf/2006.13108.pdf

Tian Z, Shen C H, Chen H and He T. 2019. FCOS: fully convolutional one-stage object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 9626-9635 [DOI: 10.1109/ICCV.2019.00972]

Wang T, Yuan L, Zhang X P and Feng J S. 2019. Distilling object detectors with fine-grained feature imitation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 4928-4937 [DOI: 10.1109/CVPR.2019.00507]

Woo S, Park J, Lee J Y and Kweon I S. 2018. CBAM: convolutional block attention module//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 3-19 [DOI: 10.1007/978-3-030-01234-2_1]

Xie Q Z, Luong M T, Hovy E and Le Q V. 2020. Self-training with noisy student improves ImageNet classification//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 10684-10695 [DOI: 10.1109/CVPR42600.2020.01070]

Yang Z D, Li Z, Jiang X H, Gong Y, Yuan Z H, Zhao D P and Yuan C. 2022. Focal and global knowledge distillation for detectors//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 4633-4642 [DOI: 10.1109/CVPR52688.2022.00460]

Zheng X P and Liang X. 2022. Deep capsule network based on pruning optimization. Chinese Journal of Computers, 45(7): 1557-1570 (郑香平, 梁循. 2022. 基于剪枝优化的深层胶囊网络. 计算机学报, 45(7): 1557-1570) [DOI: 10.11897/SP.J.1016.2022.01557]

作者简介

宋涛,男,副教授,主要研究方向为计算机视觉和图像处理。
E-mail: tsong@cqut.edu.cn
张景涛,通信作者,男,硕士研究生,主要研究方向为模型轻量化。E-mail:jtzhang134@qq.com
李沩沩,男,硕士研究生,主要研究方向为机器视觉。
E-mail:1821603900@qq.com
赵明富,男,教授,主要研究方向为智能信息处理、智能光纤感知理论与技术。E-mail:zmf@cqut.edu.cn
冉璐,女,硕士研究生,主要研究方向为小样本目标检测。
E-mail:kyrann@163.com
叶定兴,男,硕士研究生,主要研究方向为机器视觉。
E-mail:2567407957@qq.com
杨贻晨,男,硕士研究生,主要研究方向为计算机视觉。
E-mail:yyc230542@163.com
岳岱衡,男,研究研究生,主要研究方向为计算机视觉。
E-mail:ydh1998926@163.com