文章编号　1004-924X(2024)02-0237-15

多模态跨级特征知识转移下音频目标检测网络

刘诗蓓，陈　莹*

(江南大学 轻工过程先进控制教育部重点实验室，江苏 无锡 214122)

摘要：声音作为物体固有属性之一能为目标检测提供有价值的信息，当前仅通过监测环境声进行目标定位的方法鲁棒性较低，为解决这一问题提出了跨级特征知识转移下的多模态自监督目标检测网络。首先，针对教师网络和学生网络同级特征间学习能力有限的问题，设计了基于注意力融合的多教师跨级特征知识转移损失，通过注意力融合的方式融合学生的深层和浅层特征，更高效地学习对应的教师中间层特征，以提取更多的知识，同时结合KL散度，实现教师和学生网络中间层特征的对齐。此外，为了解决定位信息的缺失的问题，加入定位蒸馏损失，通过让学生的包围盒分布去拟合教师的包围盒分布的方式，来获取更多的定位信息。在多模态视听检测MAVD数据集中对网络进行训练，该网络的mAP值在IOU值为0.5，0.75和平均的情况下较基线网络分别有6.71%，14.36%和10.32%的提升。实验结果证明了该检测网络的优越性。

关　键　词：多模态；知识蒸馏；目标检测；自监督；深度学习

中图分类号：TP394.1；TH691.9　　**文献标识码**：A　　**doi**：10.37188/OPE.20243202.0237

Audio object detection network with multimodal cross level feature knowledge transfer

LIU Shibei, CHEN Ying*

(*Key Laboratory of Advanced Process Control for Light Industry* (*Ministry of Education*), *Jiangnan University*, *Wuxi* 214122, *China*)

* *Corresponding author*, *E-mail*: *chenying@jiangnan.edu.cn*

Abstract: As one of the inherent properties of objects, sound can provide valuable information for target detection. At present, the method of target positioning only by monitoring environmental sound is less robust. To solve this problem, a multi-modal self-supervised target detection network under cross-level feature knowledge transfer was proposed. First of all, in view of the teachers network and students at the same characteristics of network learning ability of the limited problem, design based on the integration of teachers across level knowledge transfer loss, through the way of attention fusion deep and shallow characteristics of students, more efficient learning to the corresponding teacher middle layer characteristics, to extract more knowledge, combined with KL divergence, realize the alignment of teachers and students network alignment. In addition, in order to solve the problem of missing localization information, localization distillation loss was added, and more localization information was obtained by fitting the distribution of the teacher. With the network trained in the multimodal audiovisual detection MAVD dataset, the mAP val-

收稿日期：2023-06-08；**修订日期**：2023-07-12.

基金项目：国家自然科学基金资助项目(No. 62173160)

ues improve by 6.71%, 14.36% and 10.32% from the baseline network at IOU values of 0.5, 0.75 and average, respectively. The experimental results demonstrate the superiority of this detection network.
Key words: multimodal; knowledge distillation; object detection; self-supervised; deep learning

1 引　言

目标检测是计算机视觉领域中的一项重要任务，其任务是从图像中找到目标物体，并判断该物体的类别和位置，其在人脸识别、自动驾驶、智能交通等方面取得了广泛的应用。近年来，目标检测取得很多突破，但仍面临许多挑战，例如光线不足，阴雨天、目标遮挡等，面对上述情况，常见的目标检测网络往往难以正确检测出目标物体的位置，这是由于它们大多依赖RGB图像进行目标检测，而这类图像对光照和天气的变化非常敏感。而声音作为物体的固有属性之一，包含了很多有价值的信息，人类可以通过对声音的感知来识别物体的类别和所在的位置。当视觉信息受限制时，声音所包含的信息对于目标检测能起到重要作用。

对于声音事件的定位和检测的实现，可使用多通道声音麦克风阵列，利用麦克风之间的信号音量差和到达时间差来推断声音发射对象相对于麦克风的位置[1]，而仅使用声音作为输入来进行目标定位的训练不仅鲁棒性低，且需要大量的劳动密集型手工注释。空间中的视听一致性表明[2]，一种模态的学习有望得到另一种模态在空间知识上的帮助，因此可以通过视听迁移学习的方法，利用知识蒸馏来避免昂贵且耗时的标记过程。Aytar等人[3]设计了一个师生网络，通过预先训练的教师模型来训练学生音频模型，并获取在未标记视频上的伪标签。Afouras等人[4]设计一个具有对比目标的自我监督框架，利用自监督的标签和预测包围盒来训练基于图像的对象检测器。Owens等人[5]通过自监督的方法，将环境声音作为监督信号来学习视觉表示。Gan等人[6]通过转移视觉教师中的知识训练立体声网络，该模型以立体声和包含相机姿态信息的元数据作为输入实现视觉框架上的目标检测和跟踪。Valverde等人[7]在Gan等人的基础上提出了一种自监督多模态蒸馏网络(Multi-Modal Distillation Network，MM-DistillNet)框架，结合多个模态，充分利用视觉和声音间的互补性和相关性，采用知识蒸馏的方式训练以音频为输入的学生网络。

中间层特征的使用能对知识蒸馏起到积极作用，但仅通过教师和学生同级特征间的知识提取，学生网络的学习能力有限，为更充分地实现视觉教师对音频学生的知识转移，本文在MM-DistillNet[7]框架的基础上改进，提出基于注意力融合的多教师跨级特征知识转移(Multi-teacher Cross-level Feature Transfer，MCFT)损失，区别于多教师对齐(Multi-Teacher Alignment，MTA)损失[7]的同级特征损失计算，MCFT损失采用自上而下不断堆叠的融合方式，通过注意力融合的方法将学生网络不同级别的中间层特征融合，得到跨级融合特征，去学习对应的教师网络浅层特征，提升学生网络的学习能力，同时加入了定位蒸馏(location distillation，LD)损失，进一步提升网络的定位能力。实验结果表明，本文提出的MCFT损失在MAVD数据集上，对于目标类别的检测精度和定位精度上均有提升。

2 相关工作

2.1 视听定位

在表示同一事件的视听流中，音频和视频流之间在时间和频率域上存在自然对应关系[8-9]，两者之间的互补性和相关性对于在视觉场景中结合声音实现目标定位来说很有价值。因此，音频和图像的结合使得能够使用多种模态来共同监督彼此[10-11]。Tian等人[12]等人提出了基于深度学习的方法来定位视频中的发声对象，Younes等人[13]利用强化学习针对复杂设置的鲁棒导航策略，实现在嘈杂和分散注意力的环境中对移动声源的捕捉。但这些方法的使用需要大量的有标记数据。

为实现在未标记视频中的定位声源，Chen等人[14]提出了一种自动背景挖掘技术，通过对比学习实现在图片中定位声源，Arandjelovic等人[15]通过使用视听通信作为目标函数从未标记的视

频中训练实现在图像中的发声对象的定位，Hu等人[16]以视听在语义上的一致性作为监督信号，使用K-means聚类方法实现声源定位。以上方法大多使用视听对双流输入，通过在输入或特征级别进行融合等方法来利用互补线索，这在具有挑战性的感知条件下能极大地提高声源定位和目标检测的性能，但各模态的融合也会使计算量增大。因此可以通过视听迁移学习的方法，利用视频中的图像和音频的自然共现线索的进行知识转移，不仅避免了劳动密集型的手工标记，且在测试过程中无需输入多个模态并进行融合，减少了目标检测过程的计算量和存储损耗。

2.2　基于中间层的知识蒸馏

知识蒸馏的使用在提升学生网络的性能同时不增加过多的训练开销。很多研究表明，除了最小化教师和学生网络最后一层分类器的输出之间的KL散度外[17]，从主干网络中提取中间层的特征表示对学生网络的学习有着积极作用[18-19]。

在知识转移上，Tian等人[20]基于对比学习在深度网络之间转移知识，Ahn等人[21]提出了一种基于变分信息最大化的知识转移框，通过学习最大化交互信息，估计激活在教师网络中的分布，激发知识的转移，Zagoruyko等人[22]尝试将教师网络的注意力图转移到学生网络中，结果表明注意力图的使用能在知识转移中起到积极作用。

大多数知识蒸馏方法对于中间层特征的使用侧重于同级别之间的特征转换或损失函数计算，而Pengguang Chen等人[23]的研究表明学生网络的深层特征对于教师网络的浅层特征有更强的学习能力。此外，除同时蒸馏所有中间层外，Aguilar G等人[24]提出了自下而上逐步匹配或自下而上不断堆叠的内部蒸馏方式，用以更有效地提取教师的中间层知识。

3　多模态目标检测网络

3.1　网络整体结构

在表示同一事件同一时间下的音频和图像对，在事件和频率与上存在自然对应关系，不同模态所包含不同信息具有互补性和相关性。如图1所示，RGB图像中包含了丰富的空间信息，在视觉上直观地表明的车辆位置，深度图像中包含了丰富的深度信息，直接反映了景物可见表面的几何形状，音频数据包含了丰富的时域和频域信息，通过频谱图来反映不同音频信号频率频谱随着时间而变化的视觉呈现。RGB和深度图像易受天气等诸多因素干扰，而音频数据虽然对天气干扰有较强的鲁棒性，但缺乏直观的空间和深度信息，易受环境噪声的影响，仅使用音频进行

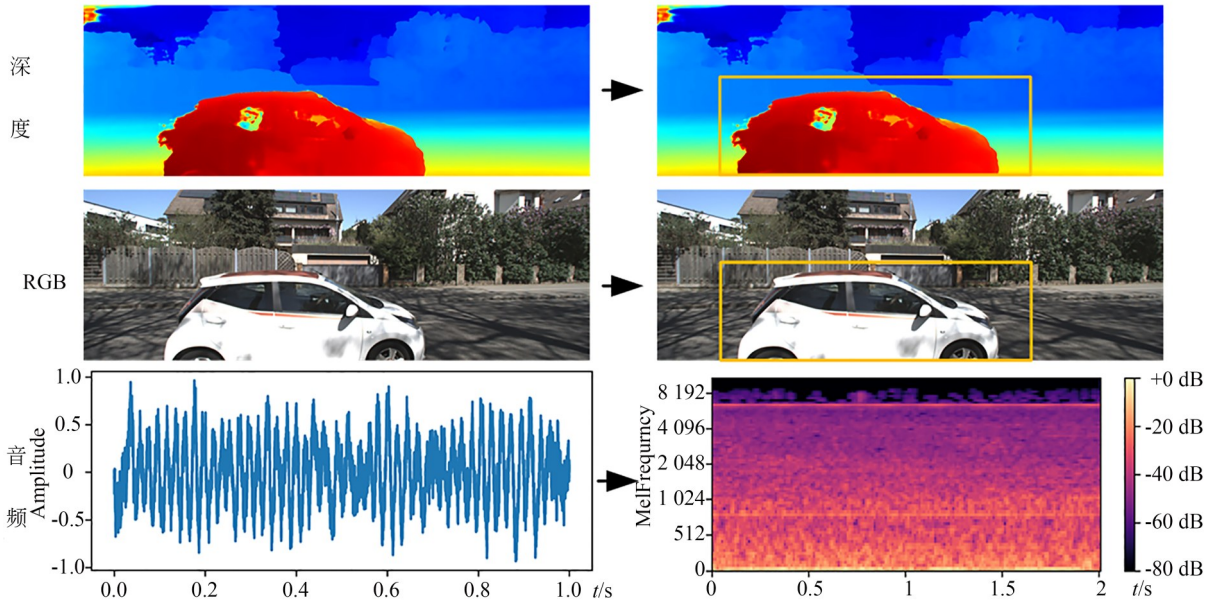


图1　RGB、深度和音频信息示意图

Fig. 1　Schematic of RGB, depth and audio information

目标检测鲁棒性低，此外由于音频在空间上的低分辨率，使得手动标记音频用于视觉域的目标定位极其困难。

根据空间中的视听一致性，一种模态的学习有望得到另一种模态在空间知识上的帮助，RGB和深度图像中所包含的丰富的空间和深度信息，可以弥补音频在空间和深度信息上的缺失，通过多模态融合的方式，虽然能在具有挑战性的感知条件下极大地提高了目标检测的性能，但音频存在的定位不准确性会对融合结果带来负面影响，且模态的增加也会使得人工标注需求增加，此外，模态的融合也会使得测试阶段的推理成本提高。由此考虑通过知识蒸馏的方式进行知识转移，使音频在保留其不受视觉限制的特点下，学习RGB和深度图像的空间和深度特征，在测试阶段仅通过音频学生即可实现在视觉空间上对车辆位置的定位，大幅提升了推理速度，此外伪标签的使用也避免了劳动密集型手工标记。

本文的网络结构如图2所示，以RGB图像和深度图像作为教师模态，从8通道单声道麦克风阵列获取的音频作为学生模态，学生网络以未标记的数据作为输入，从预先训练完成的教师网络中提取知识。该网络的目的是，学习从环境声谱图到边界盒坐标的映射，获取车辆在视觉空间中的位置。在该网络中，同一时间戳下的RGB图像、深度图像和音频分别输入到教师网络和学生网络中，每个预先训练的特定模态教师通过预测包围盒来指示车辆在各自模态空间中的位置。这些预测通过非极大抑制(Non Maximum Suppression，NMS)算法融合得到一个预测，作为训

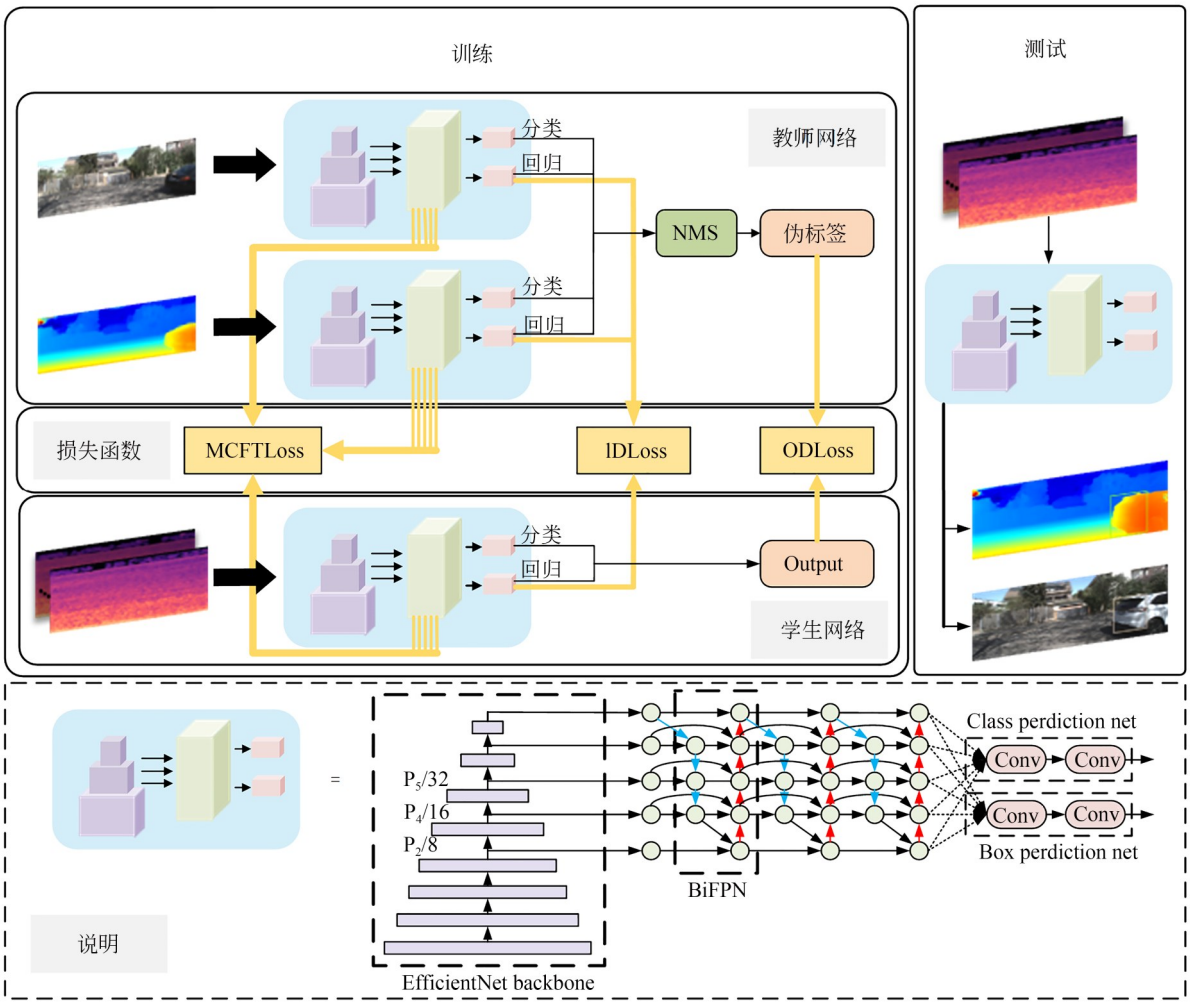


图2　多模态知识蒸馏目标检测网络

Fig. 2　Multimodal knowledge distillation target detection network

练学生网络的伪标签。

为更有效地利用图像和音频模态间的互补线索以及中间层特征包含的信息，利用多教师跨级特征知识转移损失（MCFT）来对齐学生和教师的中间表征并进行知识的提取。为获取更多的定位信息，提高定位精度，加入了定位蒸馏损失（LD）。最后利用该网络进行车辆的目标检测。研究表明采用教师和学生使用完全相同的体系结构可以提高学生网络的性能[25]。由此，综合考虑模型的性能和速度[7]选择 EfficientDet-D2[26]架构作为教师和学生的网络架构。该架构以 EfficientNet-B2[27]作为主干网络，以分辨率为 768×768 pixel 的图像作为输入，后接五个重复的双向特征金字塔（BiFPN）进行高效的多尺度特征融合，每个 BiFPN 都有 112 个通道，融合后的特征经过一个回归和分类器分支得到类别和位置的预测结果。

其中两个预先训练完成的教师模型由现有的已标记完成的公开数据集预先进行训练得到。RGB 教师模型由 COCO[28]，PASCAL VOC[29]，ImageNet[30]中的车辆数据训练得到，深度教师将 Argoverse[31]

数据集中的 3D 车辆包围盒映射到 2D 后训练得到。在最终的训练阶段，则通过 MAVD 数据集实现音频学生的训练，由预先训练完成的两个教师提供伪标签。

3.2　多教师知识的提取

为了将知识从视觉目标检测模型转移到音频模态中，本文使用了三种不同的损失函数来训练学生网络。

3.2.1　目标检测损失

目标检测损失（Object Detection，OD），在网络的最终预测时使用。两个教师模型会输出两个不同的包围盒位置的预测结果，通过 NMS 算法合并后得到一个统一的预测结果作为学生网络的伪标签。使用焦点损失函数（Focal Loss）来解决一阶段目标检测场景中前景与背景类别在训练时极端的不平衡。焦点损失公式如式（1）所示：

$$\boldsymbol{L}_{\text{focal}}=-\alpha\left(1-p_t\right)^{\gamma}\log\left(p_t\right), \tag{1}$$

其中：p_t 表示预测概率，反应与真实值类别的接近程度，α 是分配给困难样本的权重，用以平衡正负样本的重要性，γ 是聚焦参数，用以平滑地调整简单样本的权重。在训练中，与文献[7]相同，两个参数设置为 $\alpha=0.25$，$\gamma=2$。

为避免梯度爆炸，使用 Smooth L1 作为定位损失，Smooth L1 损失公式如式（2）所示：

$$\boldsymbol{L}_{\text{smooth}}=\begin{cases}0.5x^2, & if\ \ |x|<1\\ |x|-0.5, & \text{if\ \ other}\end{cases}, \tag{2}$$

其中，x 为定位预测值和真实值的差值。

两者相加即为目标检测损失：

$$\boldsymbol{L}_{\text{OD}}=\boldsymbol{L}_{\text{focal}}+\boldsymbol{L}_{\text{smooth}}. \tag{3}$$

3.2.2　多教师跨级特征知识转移损失

中间层特征包含很多分类和定位信息，利用中间层获取教师模态中包含的互补线索对于学生网络的学习有很大的作用。

MTA 损失[7]采用计算同级特征之间的 KL 散度来实现互补线索的获取，但同级特征之间学生的学习能力有限，当学习到一定程度时无法再获取更多的知识。这是由于同级的学生和教师之间存在着较大的知识差距，这好比小学学生和大学教授，在前期学生特征能从教师特征中学习到部分简单的基础知识，但到后期，学生网络无法理解教师特征中的抽象概念，难以继续进行学习。

为解决这一问题，提出了多教师跨级特征知识转移损失（MCFT），利用注意力机制，对学生的特征进行跨级融合，由融合后的学生特征去学习教师的浅层特征，这相比于同级之间的学习更有效。

学生网络的特征由浅到深也是一种学习的过程，深层的学生特征所学到的知识更为抽象，因此相比于学生的浅层特征更易理解教师浅层特征中的抽象内容。学生的由浅到深的特征就好比有着不同学习能力的学生，浅层特征侧重具体知识的学习，而深层特征注重于抽象知识，因此学生网络的浅层特征和深层特征共同学习教师网络的浅层特征，比仅使用学生的深层特征更稳定，学习的更全面更稳定。图 3 给出了中间层跨级融合和未跨级的特征热图，其中红色部位为高响应区域（彩图见期刊电子版），即网络检测的侧重区域，可以看到未跨级特征的高响应区域并不清晰，未能明确集中在目标所在位置，而跨级融合后特征的高响应区域则集中在检测目标上。

这是由于跨级融合特征采用注意力融合的方式来融合浅层和深层网络，学生网络不同层次的特征所包含的信息多样，特征所注意的区域不同，通过注意力地图可以更有效地聚合不同层次的特征，使得融合后特征能集中于检测目标位置。

MCFT 损失主要作用于 EfficientNet-B2 的 p3，p4，p5 层，该三层特征的输出分别为[batchsize，48，96，96]、[batchsize，120，48，48]，[batchsize，352，24，24]，该三层的输出经过五个重复的 BiFPN 特征网络，最终得到五个通道数均为 112 大小分别为 96×96，48×48，24×24，12×12，6×6 的不同的特征，这五个特征值由于经过了多尺度的特征融合包含了更丰富的分类和定位信息。MCFT 损失的计算主要基于这五个特征值。如图 4 所示，学生网络不同级别的特征，通过注意力融合模块（Attention-attention Fusion Module，AFM）融合生成新的特征。该融合过程不是学生网络的特征简单的由浅层到深层的两两融合：1⊕2→2⊕3→3⊕4→4，而是学生网络的特征由深层到浅层，自上而下不断堆叠融合的过程：4→3⊕4→2⊕3⊕4→1⊕2⊕3⊕4，这样的方式能更有效地利用学生的深层特征。

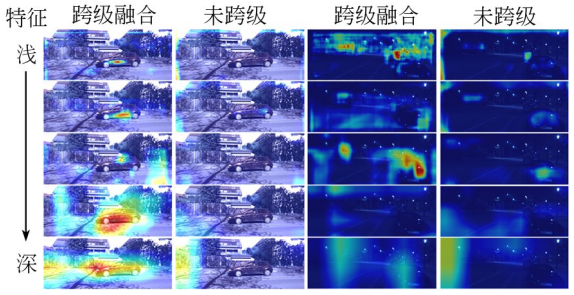


图 3　跨级融合和未跨级特征热力图

Fig. 3　Cross-level fusion and no cross-level feature heatmaps

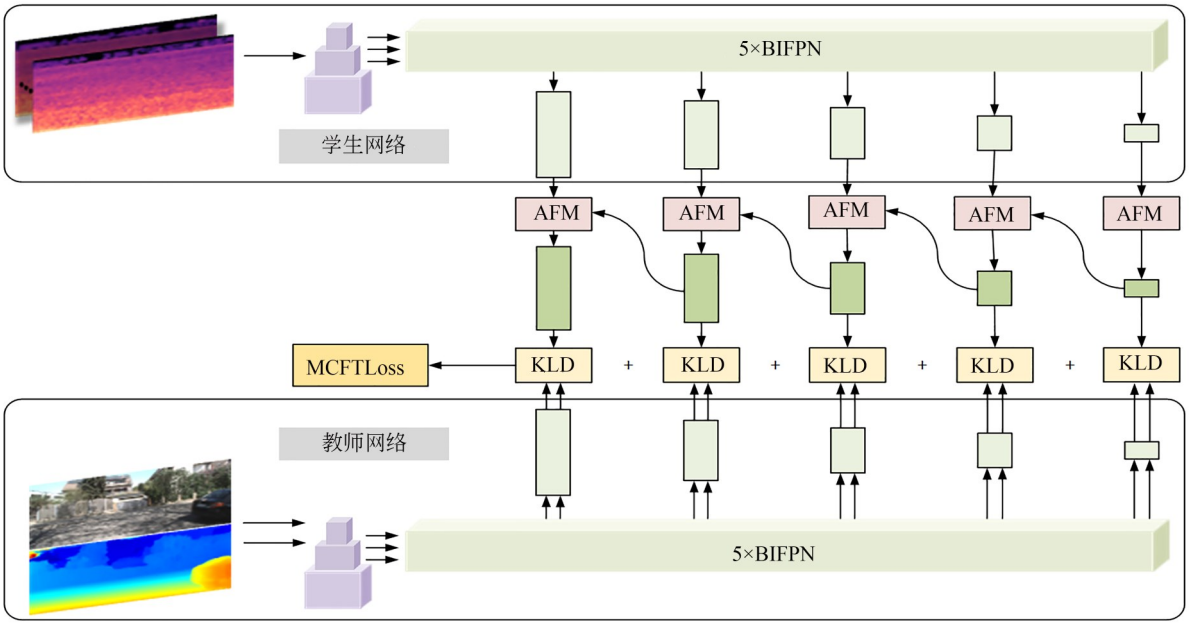
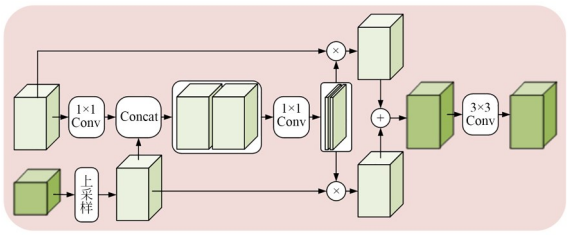


图 4　基于注意力融合的跨级特征知识转移损失

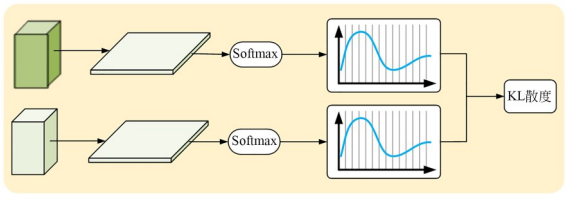
Fig. 4　Cross-level feature knowledge transfer loss based on attentional fusion

融合后的学生特征和对应的教师特征通过 KL 散度计算模块（KL Divergence Calculation Module，KLD），计算得到对应的 KL 散度值，将得到的 KL 散度求和，即为最终 MCFT 损失的值。

注意力融合模块（AFM）如图 5(a)所示，浅层特征经过一个 1×1 的卷积进行特征提取，将深层特征进行上采样使其和第一层特征大小一致，两者通道数均为 112，将两者进行通道堆叠后，经过一个 1×1 的卷积，通道数由 112×2 变为 1×2，生成两个相同大小的注意力图，将两个地图分别与对应特征相乘并相加后，得到融合特征，该特征最后经过一个卷积核大小为 3×3，填充为 1 的



(a) 注意力融合模块
(a) Attention fusion module



(b) KL散度计算模块
(b) KL divergence calculation module

图5　注意力融合模块(AFM)和KL散度计算模块(KLD)

Fig. 5　Attention fusion module(AFM) and the KL divergence calculation module(KLD)

卷积，最终得到融合后的学生特征。该过程可由公式(4)表示：

$$\begin{cases} \boldsymbol{F}_U^i = U\left(\boldsymbol{F}_A^{i+1}\right) \\ \boldsymbol{F}_D^i = f_{1\times1}\left(\boldsymbol{F}_S^i\right) \\ \boldsymbol{F}_C^i = f_{1\times1}\left(C\left(\boldsymbol{F}_U^i, \boldsymbol{F}_D^i\right)\right) \\ \boldsymbol{F}_A^i = f_{3\times3}\left(\boldsymbol{F}_D^i \cdot \boldsymbol{F}_C^i[0] + \boldsymbol{F}_U^i \cdot \boldsymbol{F}_C^i[1]\right) \end{cases}, \tag{4}$$

其中：$\boldsymbol{F}_S^i \in \boldsymbol{R}^{2n\times2n\times112}$表示第$i$层学生特征，即浅层特征，$\boldsymbol{F}_A^{i+1} \in \boldsymbol{R}^{n\times n\times112}$表示第$i+1$层融合后的学生特征，即深层特征，$\boldsymbol{F}_A^i \in \boldsymbol{R}^{2n\times2n\times112}$表示第$i$层融合后的学生特征，其中$i \in [4,3,2,1]$，$U(\cdot)$表示上采样操作，$f_{k\times k}(\cdot)$表示$k\times k$的卷积，$C(\cdot)$表示通道堆叠。

学生的融合特征是由网络的顶部至底部，由上至下计算的。当$i=5$时，即最深层特征，对其执行一个1×1卷积和一个卷积核大小为3×3，填充为1的卷积即可。公式如式(5)所示：

$$\boldsymbol{F}_A^5 = f_{3\times3}\left(f_{1\times1}\left(\boldsymbol{F}_S^5\right)\right). \tag{5}$$

KL散度计算模块(KLD)如图5(b)所示。在知识转移过程中，将教师网络的注意力图转移到学生网络能起到积极作用[22]，由此计算出融合后的学生特征和对应的教师特征的注意力地图，并标准化到[0,1]区间。学生和教师注意力地图计算公式如式(6)所示：

$$\begin{cases} \boldsymbol{Q}_S^i = f_{\text{nor}}\left(f_{\text{mean}}^r\left(\boldsymbol{F}_A^i\right)\right) \\ \boldsymbol{Q}_T^{i,j} = f_{\text{nor}}\left(f_{\text{mean}}^r\left(\boldsymbol{F}_T^{i,j}\right)\right) \end{cases}, \tag{6}$$

其中：$f_{\text{nor}}(\cdot)$表示标准化函数，$f_{\text{mean}}^r(\cdot)$表示对特征$\boldsymbol{F}$的r次幂求其在通道维度的平均，并将其展平成一维，$\boldsymbol{F}_A^i \in \boldsymbol{R}^{n\times n\times112}$表示融合后的第$i$层学生特征，$\boldsymbol{F}_T^{i,j} \in \boldsymbol{R}^{n\times n\times112}$表示第$j$个模态的第$i$层的教师特征，其中$i \in N$，$j \in M$，$N=[1,2,3,4,5]$，$M=[1,2]$。在训练中设置$r=2$。最后计算各教师模态下，各层学生注意力地图和教师注意力地图之间激活的分布的相似性，即KL散度。

最终MCFT损失定义为：

$$\boldsymbol{L}_{\text{MTCA}} = \sum_j^N \sum_i^M KL_{\text{div}}\left(S\left(\boldsymbol{Q}_S^j, \tau\right) \| S\left(\boldsymbol{Q}_T^{i,j}, \tau\right)\right), \tag{7}$$

其中：$KL_{\text{div}}(\cdot)$表示KL散度的计算，$S(\cdot,\tau)$表示温度为τ的softmax函数，通过该函数获取注意力地图的分布。在训练中设置$\tau=9.0$。

3.2.3　定位蒸馏损失

MCFT损失在获取丰富的分类信息时却损失了一部分定位信息，导致车辆包围盒的定位不够准确，因此加入定位蒸馏损失(LD)，以弥补损失的定位信息。

LD损失通过将包围盒的表示从四元表示转换成概率分布的形式，让学生的包围盒分布去拟合教师的包围盒分布。公式如式(8)所示：

$$\boldsymbol{L}_{\text{LD}} = \sum_j^M KL_{\text{div}}\left(S\left(\boldsymbol{R}_S, t\right) \| S\left(\boldsymbol{R}_T^j, t\right)\right), \tag{8}$$

其中：$\boldsymbol{R}_S$表示学生预测包围盒位置，$\boldsymbol{R}_T^j$表示第j个模态的教师预测包围盒位置。$S(\cdot,t)$表示温度为t的softmax函数，通过该函数将学生和教师的包围盒位置转换成概率的分布。在训练中设置$t=10.0$。

3.2.4　损失函数

最终的损失函数是对目标检测损失，多教师跨级特征知识转移损失和定位蒸馏损失进行加权和：

$$\boldsymbol{L}_{\text{total}} = \delta * \boldsymbol{L}_{\text{TD}} + \beta * \boldsymbol{L}_{\text{MTCA}} + \lambda * \boldsymbol{L}_{\text{LD}}, \tag{9}$$

其中，δ，β，λ三个值作为超参数用于平衡损失。

4　实验结果和分析

4.1　实验配置

本文网络在深度学习框架 Pytorch 下完成，训练和测试所使用的环境为 Ubuntu 18.04，CUDA11.0，Python3.6，硬件配置为两张 RTX 3090 显卡。

在训练中，采用 ReduceLROnPlateau 策略动态更新学习率，最多训练 50 个 epoch，初始学习率设为 0.000 1，权值衰减为 0.000 5，动量为 0.9，批处理大小为 8，采用 Adam 优化器。

对于损失函数的计算，对各参数设置为：$\delta=1.0$，$\beta=0.005$，$\lambda=0.25$，在该参数设置下，目标检测性能最优，具体分析见 4.4 中的表 4 和表 5。

对于输入的图像，取相同时间戳下的 RGB 和深度图像作为教师网络的输入，RGB 和深度图像的原始分辨率为 1 920×650 pixel，将其调整为 768×768 pixel 为作为 EfficientDet-D2 架构的输入。

对于输入的音频，如图 6 所示，以图像时间戳为中心，提取该时间戳前 0.5 s 和后 0.5 s 的音频片段，得到 8 个时长为 1 s 的单声道麦克风的环境声音片段。将这 8 个声音片段通过短时傅里叶变换（STFT）来得到声谱图，综合考虑频域分辨率以及算计成本，选择 1 024 作为 FFT 窗口的大小，由此生成 8 个 513×173 的声谱图，之后经过 80 个梅尔滤波器，在梅尔频率尺度上重采样得到 8 个 80×173 的梅尔频谱图，最后通过双立方插值法，将 8×80×173 的梅尔频谱图调整为 8×768×768 pixel，调整后的频谱图矩阵的 w/h 对应频率和时间，以此作为学生网络的输入，对其卷积操作能提取频谱图的频域特征。

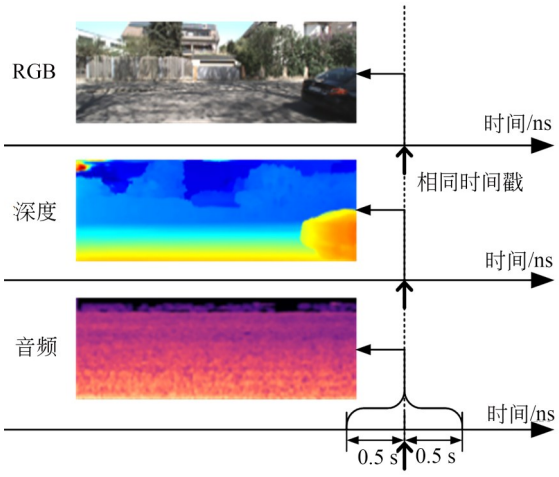


图 6　图像和音频的选取示意

Fig. 6　Selection diagram of image and audio

4.2　数据集和评价指标

4.2.1　数据集

本文使用公开的多模态视听检测（MAVD）数据集，该数据集记录了高交通密度、高速公路行驶和多个交通灯等不同场景以及传统城市驾驶、有轨电车附近和通过隧道等不同噪音状况下的各模态数据，并提供了静止状态和行驶状态下记录数据，用于训练和测试的数据集总共包含 24 589 张白天静态图像、26 901 张夜间静态图像、26 357 张白天行驶图像和 35 436 张夜间行驶图像，总计 113 283 个同步多声道音频、RGB、深度图和红外图。该数据集使用一个 RGB 立体摄像机装置、一个热立体摄像机装置和八阵列单声道麦克风，音频以 1 通道 Microsoft WAVE 格式记录和存储，采样率为 44 100 Hz，所有数据都通过 GPS 时钟相互同步。由于该数据集是在实际环境中采集得到，存在一定的噪声数据，如图 7 所示，存在出现车辆不完整，车辆密集，小目标，复杂环境，光线昏暗，车辆模糊等情况，并不是理想化的数据集，因此对于实际应用有较大的使用价值。

图 7　MAVD 数据集示例图像

Fig. 7　Example images of MAVD dataset

数据集中的音频文件以 MP3 的格式存储，由于 libraosa 读取 MP3 文件的速度较慢，因此需在训练前将 MP3 格式的音频文件转换成声谱图，将其存储成 pkl 文件的形式，在训练和测试时将转换得到的 pkl 文件作为音频的输入，以缩短数据读取时间。

在数据分割上，本文采用60/20/20%的方式对，分别对应训练、验证和测试。

4.2.2　评价指标

本文主要采用以下两个评价指标作为目标检测网络性能的优劣评判依据：

平均精度均值(mAP)：指在每个类别的精度和召回率曲线下的插值区域的类别上的平均值。本文计算了IoU为0.5和0.75时的mAP，以及IoU阈值从0.5到0.95间隔为0.05时的平均mAP。

Gan[6]等人提出的中心距离：使用预测盒的中心点来测量x和y坐标上的定位精度。中心距离的计算公式如式(10)所示：

$$\begin{cases} CD_x = \dfrac{1}{K}\sum_i^K \left| P_i^x - G_i^x \right| / w \\ CD_y = \dfrac{1}{K}\sum_i^K \left| P_i^y - G_i^y \right| / h \end{cases}, \tag{10}$$

其中：(P^x, P^y)表示最近的预测包围盒的中心点，(G^x, G^y)表示真实包围盒的中心点，K表示真实包围盒的总数，w和h表示图像的宽度和高度。

4.3　定量结果

为评估本文提出方法的有效性，在MAVD对算法进行了评估，本文算法和基线网络在不同教师模态下的比较结果在表1中给出，其中粗体表示最优结果，加下划线表示次优结果。可以看出本文方法在各IOU值下的mAP值均大于基线网络。在单RGB教师模态下，MCFT损失相较于MTA损失在IOU值为平均、0.5和0.75的情况下mAP值分别提升了6.12%，7.80%，6.78%，中心距离*CDx*和*CDy*分别降低了0.68和0.41。在单深度教师模态下，MCFT损失相较于MTA损失在IOU值为平均、0.5和0.75的情况下mAP值分别提升了7.76%，9.44%，7.95%，中心距离*CDx*和*CDy*分别降低了1.28和0.61。在RGB和深度的双教师模态下，本文方法在IOU值为平均、0.5和0.75情况下mAP值分别为62.23%，82.63%和61.49%，相较于MTA损失分别有10.32%，6.71%和14.36%的提升，中心距离*CDx*和*CDy*分别降低了0.12和0.06，说明本文方法在明显提升目标识别精度上的同时，并未降低定位精度。

表1　本文方法和基线网络在不同教师模态下的结果比较

Tab.1　Results comparison of the method and the baseline network under different faculty modes

模型	教师模态		mAP值(越大越好)			中心距离(越小越好)	
	RGB	深度	mAP@Avg	mAP@0.5	mAP@0.75	*CDx*	*CDy*
StereoSoundNet[6]	√	—	44.05	62.38	41.46	3.00	2.24
Baseline[7]	√	—	51.45	69.22	49.07	2.97	1.72
	—	√	40.28	54.09	38.45	6.08	3.28
	√	√	51.91	75.92	47.13	<u>2.07</u>	<u>1.11</u>
Ours	√	—	<u>57.57</u>	<u>77.02</u>	<u>55.85</u>	2.29	1.31
	—	√	48.04	63.53	46.40	4.80	2.67
	√	√	**62.23**	**82.63**	**61.49**	**1.95**	**1.05**

为评估本文方法的实时性，选取了Faster R-CNN，SSD，Yolov3，Yolov5这几个经典的目标检测网络进行了比较，利用每秒检测帧数(frame per second，FPS)作为实时性评价指标，所有模型均在一张RTX 3090显卡上进行，用随机数的方式生成对应大小的输入数据，每个网络均经过多次测试取其平均值，最终结果如表2所示。可以看到本文方法相较于YOLOv3和SSD实时性较差，相较于Fast R-CNN，Yolov5-x实时性较好。这是由于本文使用的网络结构为EfficientDet-D2，相较于以实时性为特点SSD和YOLO网络性能较差，但在同样以EfficientNet-B2作为骨干网络的情况下，本文方法的实时性较高。

在检测精度上EfficientDet的表现更优秀，在尽可能减少计算量的情况下仍有较高的精度，对于鲁棒性较低的视听定位任务，EfficientDet的高鲁棒性和高准确度更为合适。如图8所示，SSD在小目标检测上能力较弱，YOLO在大目标检测

表 2　本文方法与经典目标检测网络实时性比较

Tab. 2　This paper compares the method with classical object detection networks

模型	FPS/(FPS)	模型	FPS/(FPS)
Faster R-CNN VGG16	18.41	Yolov3-m	94.81
Faster R-CNN ResNet	13.15	Yolov3-l	66.89
Yolov5-x(EfficientNet-B2)	43.82	Yolov5-s	118.17
SSD300(EfficientNet-B2)	44.41	Yolov5-m	93.20
SSD300	121.39	Yolov5-l	67.04
SSD500	84.16	Yolov5-x	48.33
Yolov3-s	96.90	Ours	49.91

图 8　不同网络架构下目标检测能力比较

Fig. 8　Comparison of object detection capability under different network architecture

上能力较弱，EfficientDet网络在检测能力最为优秀。此外，EfficientDet的骨干网络EfficientNet具有较好的特征提取能力，BiFPN网络能更有效地融合骨干网络输出的不同尺度的特征，对于知识蒸馏过程中对于中间层特征的使用更有价值。

4.4　消融实验

为验证MCFT损失和LD损失的有效性，对该两种损失进行了消融实验，结果如表3所示。使用了LD损失后的M2模型比未使用任何损失的M1模型在mAP值为平均、0.5和0.75时分别有3.28%，4.63%和4.83%的提升，在中心距离CDx和CDy上分别降低了0.18和0.1，使用了MCFT损失和LD损失的M4模型结果比仅使用MCFT损失的M3模型在中心距离CDx和CDy上分别降低了0.03和0.03，在mAP值为0.5和0.75时分别提升了0.04%和0.11，证明LD损失能提升目标检测的定位精度。使用了MCFT损失的M3模型比未使用任何损失的M1模型在mAP值为平均、0.5和0.75时分别有9.71%，10.18%和11.34%的提升，在中心距离*CDx*和*CDy*上分别降低了0.71和0.43，证明MCFT损失对于教师中间层知识提取的有效性。

表 3　两种损失的消融研究

Tab. 3　Ablation studies for both losses

模型	损失		mAP值			中心距离	
	MCFT Loss	LD Loss	mAP@Avg	mAP@0.5	mAP@0.75	*CDx*	*CDy*
M1	—	—	52.68	72.05	50.04	2.69	1.51
M2	—	√	55.96	76.68	54.87	2.51	1.41
M3	√	—	**62.39**	82.23	61.38	1.98	1.08
M4	√	√	62.23	**82.63**	**61.49**	**1.95**	**1.05**

为获取更性能最优的超参数，分别对δ，β和λ三个超参弧数进行了两组消融实验。表4给出了δ为1.0时，不同β设置下的网络检测精度，表5给出了δ为1.0，β为0.005时，不同λ设置下的网络检测精度。如表4所示，在仅使用MCFT损失的情况下，令$\delta=1.0$，当$\beta=0.005$时模型mAP值最高，且中心距离最小。如表5所示，在加入LD损失的情况下令$\delta=1.0$，$\beta=0.005$，当$\lambda=0.25$时模型mAP值最高，且中心距离最小。

因此在训练时令$\delta=1.0$，$\beta=0.005$，$\lambda=0.25$，此时目标检测精度最优。

表4　损失函数中超参数δ和β的消融研究

Tab. 4　Ablation study of loss parameters δ and β

超参数		mAP值			中心距离	
δ	β	mAP@Avg	mAP@0.5	mAP@0.75	*CDx*	*CDy*
1.0	0.003	52.88	72.85	50.77	2.65	1.57
1.0	0.005	**62.39**	**82.23**	**61.38**	**1.98**	**1.08**
1.0	0.008	53.86	72.49	51.74	2.81	1.61
1.0	0.01	50.43	69.12	48.44	3.11	1.80
1.0	0.03	51.55	69.56	49.82	3.06	1.75
1.0	0.05	59.29	78.97	57.52	2.25	1.25
1.0	1.0	49.97	67.24	47.87	3.22	1.82

表5　损失函数中超参数δ，β和λ的消融研究

Tab. 5　Ablation studies of loss parameters δ, β and λ

超参数			mAP值			中心距离	
δ	β	λ	mAP@Avg	mAP@0.5	mAP@0.75	*CDx*	*CDy*
1.0	0.005	0.005	50.13	66.40	48.27	3.52	2.06
1.0	0.005	0.06	51.17	70.89	48.76	2.87	1.65
1.0	0.005	0.01	52.22	71.67	49.80	2.86	1.64
1.0	0.005	0.25	**62.23**	**82.63**	**61.49**	**1.95**	**1.05**
1.0	0.005	0.3	50.95	70.24	48.97	2.92	1.71
1.0	0.005	1.0	55.79	78.13	53.40	2.28	1.29

为验证不同融合方式以及不同损失函数计算方式对实验结果的影响，进行如图9所示的消融实验，分别针对是否跨级，是否融合，融合方式以及损失函数计算方式进行消融研究。融合方式包括两两融合和堆叠融合，损失函数计算方式包括KL散度计算和L2距离计算。

图9(a)的不跨级不融合为学生和教师网络同级特征间的损失计算，图9(b)的跨级融合为学生网络的深层特征和教师网络浅层特征间的损失计算，图9(c)的跨级两两融合为学生网络浅层特征和该浅层特征后一层的深层特征通过注意力融合得到的融合后特征与教师网络的浅层特征间的损失计算，图9(d)的跨级堆叠融合为学生网络由深到浅不断堆叠融合，浅层的学生网络特征通过注意力融合的方式融合了该浅层前一层的融合后特征，由得到的融合后特征和教师网络的浅层特征进行损失计算。

实验结果如表6所示，结果表明，对于不同的损失计算方式，使用KL散度得到的目标检测精度较使用L2距离得到的精度高，中心距离也相较于使用L2距离得到的更小，L2距离通过计算学生和教师特征间的距离，拉近学生和教师在特征层上的相似度，而KL散度通过计算学生和教师特征间概率分布的差异来拟合学生和教师的中间层，相较于单纯的相似度计算，KL散度能通过拟合学生和教师间注意力图激活的分布来对

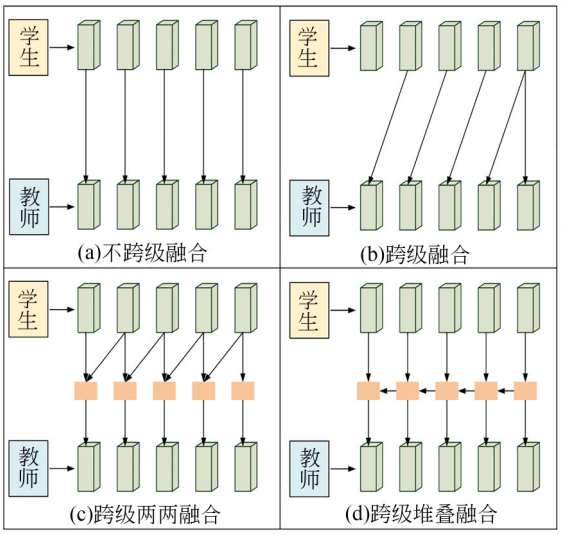


图 9　不同融合方式示意图

Fig. 9　Schematic diagram of different fusion modes

齐并拉近不同模态的教师和学生网络的特征，对于目标间层精度的提升有着更积极的作用。

对于跨级特征实验结果的影响，可以看到使用跨级特征对于目标检测的精度有较大提升，但同时也会导致中心距离增大，说明跨级特征在提取更多的分类信息的同时会损失一部分定位信息。

对于三种不同的融合方式，可以看到堆叠融合的目标检测精度最高，且中心距离最小，其次是两两融合，不融合的精度最低且中心距离最大，这说明学生浅层特征和深层特征融合后的特征相较于单纯的学生深层特征能从教师特征中学到更多的分类和定位信息，而堆叠融合相较于简单的两两融合能使学生的浅层特征融合更多的深层特征，其融合后特征的学习能力更为优秀。

表 6　不同融合方式以及损失计算方式的消融研究

Tab. 6　Ablation studies with different fusion methods and loss calculation methods

方法			mAP 值			中心距离	
跨级	融合方式	损失计算方式	mAP@Avg	mAP@0.5	mAP@0.75	*CDx*	*CDy*
—	—	KL	51.91	75.92	47.13	2.07	1.11
—	—	L2	52.68	72.05	50.04	2.69	1.51
√	—	KL	58.36	78.02	56.97	2.31	1.27
√	—	L2	56.13	75.33	54.74	2.67	1.48
√	两两融合	KL	62.15	81.84	61.13	2.04	1.13
√	两两融合	L2	58.68	77.97	56.44	2.28	1.28
√	堆叠融合	KL	**62.39**	**82.23**	**61.38**	**1.98**	**1.08**
√	堆叠融合	L2	61.74	80.54	60.45	2.05	1.11

4.5　定性评估

LD 损失的加入，目的是提高定位精度。图 10 中分别显示了无 LD 损失和有 LD 损失时的预测结果，其中红色框为预测值绿色框为真实值（彩图见期刊电子版）。如图 10 所示，当未加入 LD 损失时，预测车辆包围框与真实值相比有较大的偏差，加入 LD 损失后偏差变小，说明 LD 损失起到了提升定位精度的作用。

MTA 损失由于其仅作用在同级学生和教师的中间层特征中，对教师网络的学习能力有限，由图 11 中 MTA 损失的 Loss 曲线可看出，MTA 损失在下降到一定程度后损失大小趋于不变，说明学生网络在学习到一定程度后难以继续从教师网络中获取知识，而 MCFT 损失的 Loss 曲线

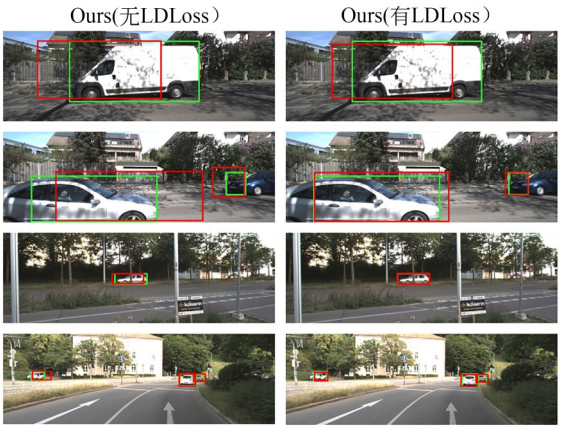


图 10　定性比较有无 LDLoss 时车辆检测能力

Fig. 10　Qualitative comparison of vehicle detection capability with or without LDLoss

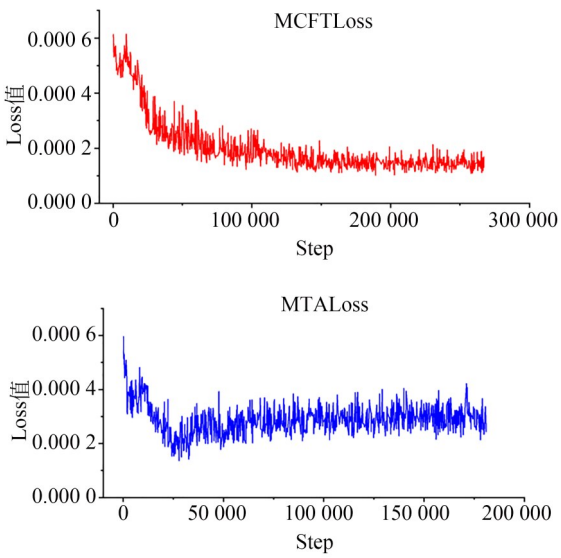


图 11　MTALoss 和 MCFTLoss 的 Loss 曲线

Fig. 11　Los curves for MTALoss and MCFTLoss

则处于不断下降的过程，说明 MCFT 损失对于教师知识的学习相较于 MTA 损失更稳定更持久更有效。

图 12（彩图见期刊电子版）中显示了由 RGB 和深度两个教师预测生成的伪标签（绿色框所示），基线网络和本文方法的车辆预测结果（红色框所示）。如图 12 第一列所示，RGB 和深度教师中都能检测出该车辆，但在基线网络中未能检测出；第二列中 RGB 教师检测出三辆车，深度教师检测出两辆车，基线网络仅测出一辆车；第三列中 RGB 教师检测出四辆车，深度教师检测出两辆车，而基线网络未检测出车辆；第四列中 RGB 教师检测出五辆车，深度教师检测出四辆车，而基线网络仅检测出两辆车，而本文方法在四个场景中分别检测出了一辆、三辆、四辆和五辆，且在车辆定位上也与伪标签并无过多的偏差，相较与基线网络，本文方法检测出了更多的车辆，证明了 MCFT 损失在教师中间层知识的提取上比 MTA 损失更充分。

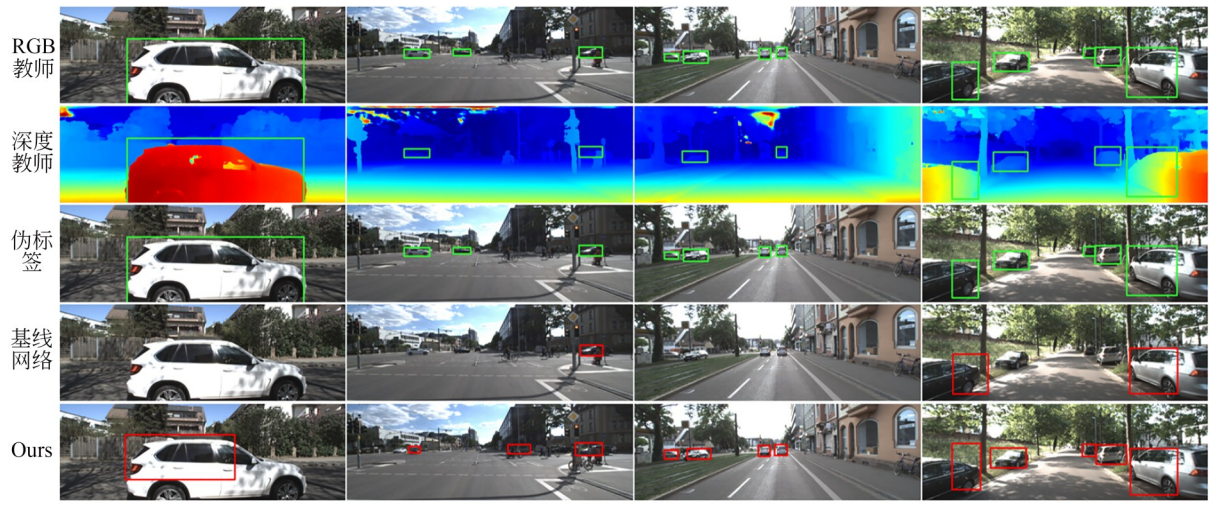


图 12　定性比较基线网络和本文方法的车辆检测能力

Fig. 12　Qualitatively compares the vehicle detection capabilities of the baseline network and the method presented in this paper

5　结　论

本文提出了基于声音定位的多模态跨级特征知识转移知识蒸馏目标检测网络，通过知识蒸馏的方式令音频网络学习以 RGB 和深度图像作为输入的教师网络的知识。设计了基于注意力融合的多教师跨级特征对齐损失，通过注意力融合的方式，融合学生的深层和浅层特征以拥有更强的学习能力，从而实现更有效且稳定的学习。加入定位蒸馏损失，用学生网络的包围盒分布去拟合教师网络的包围盒分布，以获取更多的定位信息。在公开的 MAVD 数据集中，在 IOU 值为 0.5，0.75 和平均情况下 mAP 值分别为 82.63%，61.49% 和 62.23%，相较于基线网络分别有 6.71%，14.36% 和 10.32% 的提升。

参考文献：

[1] ADAVANNE S, POLITIS A, NIKUNEN J, *et al*. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks[J]. *IEEE Journal of Selected Topics in Signal Processing*, 2019, 13(1): 34-48.

[2] WEI Y, HU D, TIAN Y, *et al*. Learning in Audio-Visual Context: a Review, Analysis, and New Perspective [EB/OL]. [2022-08-20]. https://arxiv.org/abs/2208.09579

[3] AYTAR Y, VONDRICK C, TORRALBA A. SoundNet: learning sound representations from unlabeled video[C]. *Advances in neural information processing systems*, 2016: 892-900.

[4] AFOURAS T, ASANO Y M, FAGAN F, *et al*. Self-supervised object detection from audio-visual correspondence[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA*. IEEE, 2022: 10565-10576.

[5] OWENS A, WU J J, MCDERMOTT J H, *et al*. *Ambient Sound Provides Supervision for Visual Learning* [M]. Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 801-816.

[6] GAN C, ZHAO H, CHEN P H, *et al*. Self-supervised moving vehicle tracking with stereo sound [C]. 2019 *IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South)*. IEEE, 2019: 7052-7061.

[7] VALVERDE F R, VALERIA HURTADO J, VALADA A. There is more than meets the eye: self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge[C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA*. IEEE, 2021: 11607-11616.

[8] AFOURAS T, CHUNG J S, SENIOR A, *et al*. Deep audio-visual speech recognition [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(12): 8717-8727.

[9] LI Y D, LIU H, TANG H. Multi-modal perception attention network with self-supervised learning for audio-visual speaker tracking[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, 36(2): 1456-1463.

[10] ZÜRN J, BURGARD W. Self-supervised moving vehicle detection from audio-visual cues[J]. *IEEE Robotics and Automation Letters*, 2022, 7(3): 7415-7422.

[11] ZÜRN J, BURGARD W, VALADA A. Self-supervised visual terrain classification from unsupervised acoustic feature learning[J]. *IEEE Transactions on Robotics*, 2021, 37(2): 466-481.

[12] TIAN Y P, SHI J, LI B C, *et al*. *Audio-Visual Event Localization in Unconstrained Videos*[M]. Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 252-268.

[13] YOUNES A, HONERKAMP D, WELSCHEHOLD T, *et al*. Catch me if you hear me: audio-visual navigation in complex unmapped environments with moving sounds[J]. *IEEE Robotics and Automation Letters*, 2023, 8(2): 928-935.

[14] CHEN H L, XIE W D, AFOURAS T, *et al*. Localizing visual sounds the hard way [C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA*. IEEE, 2021: 16862-16871.

[15] ARANDJELOVIC R, ZISSERMAN A. Objects that sound[C]. *Proceedings of the European conference on computer vision (ECCV)*, 2018: 435-451.

[16] HU D, QIAN R, JIANG M Y, *et al*. Discriminative Sounding Objects Localization via Self-Supervised Audiovisual Matching[EB/OL]. 2020: *arXiv*: 2010.05466. http://arxiv.org/abs/2010.05466.pdf

[17] HINTON G, VINYALS O, DEAN J. Distilling the Knowledge in a Neural Network [EB/OL]. 2015: *arXiv*: 1503.02531. http://arxiv.org/abs/1503.02531.pdf

[18] ROMERO A, BALLAS N, KAHOU S E, *et al*. FitNets: Hints for Thin Deep Nets [EB/OL]. 2014: *arXiv*: 1412.6550. http://arxiv.org/abs/1412.6550.pdf

[19] TUNG F, MORI G. Similarity-preserving knowledge distillation[C]. 2019 *IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South)*. IEEE, 2019: 1365-1374.

[20] TIAN Y L, KRISHNAN D, ISOLA P. Contrastive Representation Distillation [EB/OL]. 2019: *arXiv*: 1910.10699. http://arxiv.org/abs/1910.10699.pdf

[21] AHN S, HU S X, DAMIANOU A, *et al*. Variational information distillation for knowledge transfer [C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *Long Beach*, *CA*, *USA*. IEEE, 2019: 9155-9163.

[22] ZAGORUYKO S, KOMODAKIS N. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer [C]. *International Conference on Learning Representations* (*ICLR*), 2017.

[23] AGUILAR G, LING Y, ZHANG Y, *et al*. Knowledge distillation from internal representations [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34(5): 7350-7357.

[24] CHEN P, YU D. Improved Faster RCNN approach for vehicles and pedestrian detection[J]. *International Core Journal of Engineering*, 2020, 6(3): 119-124.

[25] FURLANELLO T, LIPTON Z C, TSCHANNEN M, *et al*. Born Again Neural Networks[EB/OL]. 2018: *arXiv*: 1805.04770. http://arxiv.org/abs/1805.04770.pdf

[26] TAN M X, PANG R M, LE Q V. EfficientDet: scalable and efficient object detection [C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *Seattle*, *WA*, *USA*. IEEE, 2020: 10778-10787.

[27] TAN M, LE Q. Efficientnet: rethinking model scaling for convolutional neural networks [C]. *International Conference on Machine Learning*, 2019: 6105-6114.

[28] LIN T Y, MAIRE M, BELONGIE S, *et al*. *Microsoft COCO*: *Common Objects in Context*[M]. Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.

[29] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, *et al*. The pascal visual object classes (VOC) challenge [J]. *International Journal of Computer Vision*, 2010, 88(2): 303-338.

[30] DENG J, DONG W, SOCHER R, *et al*. ImageNet: a large-scale hierarchical image database [C]. 2009 *IEEE Conference on Computer Vision and Pattern Recognition*. *Miami*, *FL*, *USA*. IEEE, 2009: 248-255.

[31] CHANG M F, LAMBERT J, SANGKLOY P, *et al*. Argoverse: 3D tracking and forecasting with rich maps [C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *Long Beach*, *CA*, *USA*. IEEE, 2019: 8740-8749.

作者简介：

刘诗蓓(1998－)，女，浙江舟山人，硕士研究生，2021年于江南大学获学士学位，主要研究方向为模式识别。E-mail: 1034170406@stu.jiangnan.edu.cn

通讯作者：

陈　莹(1976－)，女，浙江丽水人，博士，教授，博士生导师，2005年于西安交通大学获得博士学位，主要研究方向为机器视觉、信息融合、模式识别。E-mail: chenying@jiangnan.edu.cn