DOI：10. 11714/acta. snus. ZR20260045

基于自监督视觉模型的特种车辆细粒度检测方法*

李熙莹[1,2]，吴浩[1,2]，盘婳燕[3]，李锦[3]，朱奕仰[1,2]，胡伟鹏[1,2]

1. 中山大学智能工程学院，广东 深圳 518107
2. 广东省智能交通系统重点实验室，广东 深圳 518107
3. 广东省公安厅科技信息化总队，广东 广州 510050

摘　要：本文结合通用检测与原型库匹配技术，提出了一种基于自监督视觉表征的域适应检测框架。该方法利用DINOv3的预训练特征作匹配与筛选，从而在少样本标注且无需微调的情况下定位并区分目标车辆。首先，构建了一个多视角合成的数据增强前置模块，生成多角度样本以对齐俯视监控场景，弥补跨域场景下的视角缺失。然后，设计了一种类间聚类与原型匹配方法，通过聚类算法挖掘数据模态，构建包含多种形态的真实原型库以解决类内差异大的问题；在此基础上引入全局与局部联合表征，结合图像不同网络层级中语义与纹理的细节，实现对目标车辆的细粒度判别。实验表明，该方法在少样本条件下有效克服了视角域偏移导致的检测能力降低；相比传统方法，该方法提升了检测召回率，显著降低了由非目标车辆引起的误报，验证了该域适应框架在特种车辆监管场景中的有效性与鲁棒性。

关键词：特种运输车辆；自监督视觉模型；域适应检测

中图分类号：U491　　**文献标志码：**A　　**文章编号：**2097－0137(XXXX)XX－0001－11

Fine-grained detection method for special vehicles based on self-supervised vision models

Li Xiying[1,2], *Wu Hao*[1,2], *Pan Huayan*[3], *Li Jin*[3], *Zhu Yiyang*[1,2], *Hu Weipeng*[1,2]

1. School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518107, China
2. Guangdong Provincial Key Laboratory of Intelligent Transportation Systems, Shenzhen 518107, China
3. Science, Technology and Informatization Corps of Guangdong Provincial Public Security Department, Guangzhou 510050, China

Abstract: This paper proposes a domain adaptation detection framework based on self-supervised visual representations, integrating general object detection with prototype library matching techniques. By leveraging the pre-trained features of DINOv3 for matching and screening, the proposed method can localize and distinguish target vehicles under few-shot conditions without the need for fine-tuning. First, a data augmentation front-end module based on multi-view synthesis is constructed to generate multi-angle samples. This aligns with overhead surveillance scenes, compensating for viewpoint deficiency in cross-domain settings. Subsequently, an inter-class clustering and prototype matching method is designed. By mining data modalities via clustering algorithms, a real-world prototype library encompassing various morphologies is constructed to address the issue of large intra-class variations.

* **收稿日期：**2026-02-09　　**录用日期：**2026-04-01　　**网络首发日期：**XXXX－XX－XX
基金项目：国家自然科学基金(U21B2090)；
广东省基础与应用基础研究基金(2022A1515010361，2024YY27)
作者简介：李熙莹(1972年生)，女；**研究方向：**图像处理；E-mail：stslxy@mail.sysu.edu.cn
通信作者：胡伟鹏(1993年生)，男；**研究方向：**计算机视觉；E-mail：huwp7@mail.sysu.edu.cn

全文阅读
ZR20260045

Building upon this, a joint global and local representation is introduced, which integrates semantic and textural details from different network layers to achieve fine-grained discrimination of target vehicles. Experimental results demonstrate that under few-shot conditions, the proposed method effectively overcomes the degradation in detection performance caused by viewpoint domain shift. Compared with traditional approaches, it improves the detection recall rate and significantly reduces false positives triggered by non-target vehicles, validating the effectiveness and robustness of the proposed domain adaptation framework in special vehicle monitoring scenarios..

Key words: special transport vehicle; self-supervised visual model; domain adaptive detection

在交通监控检测任务场景下，车辆检测存在尺度变化大、场景复杂等典型而又难以解决的情况。对于尺度变化这一问题，研究者在特征金字塔网络FPN的基础上进行了很多方法上的改进（Lin et al.，2017），例如：加入路径聚合网络PANet（Liu et al.，2018）、双向融合等方法来改进不同尺度下模型的特征表达语义。Woo et al.（2018）引入注意力机制CBAM，利用注意力机制更好地学习局部细节信息。Hu et al.（2018）将SE-Block模块嵌入骨干中，以抑制背景中噪声，突出检测对象的细节特征。上述研究在通用车辆检测中效果不错（李经宇等，2021），但在特种车辆检测中存在高度相似且难于区分的关键信息点。而且，这些关键信息点（如菱角、圆形罐体等）占车辆外观图像像素值比例较小，导致现有方法并不适用。对于车辆重识别任务，一些细节感知方法可以有效地提取外观相似的车辆特征（邱铭凯等，2021），但是这些有监督检测方法过多地依赖海量注释数据的强监督信号拟合细节信息，在样本稀少的特种车辆检测问题中会因细节信息不充分而降低模型的准确率。

大模型的发展使得基于视觉-语言多模态预训练的开放词汇检测逐渐变为主流（薛天朗等，2025）。GLIP、GroundingDINO、YOLO-World等代表性的开放词检测方法，通过文本编码器把检测任务转化为图像-文本的对齐问题，可实现对非见类别的零样本定位、检测（Li et al.，2022；Liu et al.，2023；Cheng et al.，2024）。但危化品车与普通货车的不同点集中体现在罐体的具体形状、危化品标识等细节上，这些“难以描述”的特征难以利用自然语言文字来表达和对齐；且，训练高鲁棒性的图像-文本对齐检测器仍需依赖大量的标注数据，这使其难以适应特种车辆数据稀缺的现实场景。图像提示直接利用支撑集图像作为查询向量，能规避选定检测目标语言描述的模糊性。不同于靠图文对齐的训练策略，以DINO为代表的自监督学习（Self-Supervised Learning）对图像的自掩码重建和自蒸馏训练，在无监督条件下也能够学习到更为抽象的视觉特征（Oquab et al.，2023）。因此，挖掘DINOv3模型在细粒度特征上的表征能力，成为在缺乏数据条件下检测难以描述的目标任务的关键路径。

数据增强也是解决跨域泛化问题的有效手段。MixUp（Zhang et al.，2017）、Mosaic（Bochkovskiy et al.，2004）、CutMix（Yun et al.，2019）将一个图像内的采样改造成多样性样本，而Copy-Paste（Ghiasi et al.，2021）、RandAugment（Cubuk et al.，2020）等采样方法进一步增加了目标实例样式的组合，但这些方法都是内部重组，并不能生成几何一致、成像角度不同的样本。生成式人工智能，比如：StableDiffusion结合ControlNet的结构控制，能很好地使用轮廓和姿态生成可控的图像（Rombach et al.，2022；Zhang et al.，2023）。针对视角跨域问题，Zero123（Liu et al.，2023）、DreamFusion（Poole et al.，2022）、MVDream（Shi et al.，2023）、DreamGaussian（Tang et al.，2023）等方法学习3D先验，可实现用一张图片对该对象任意视角的合成，但直接利用生成模型做数据增强并不能很好的完成数据增强的目的，生成的车辆纹理可能偏离真实风格。

综上所述，现有特种车辆检测方法面临三大挑战：（1）现有的通用检测器过于依赖大数据，难以在少样本下学习细节；（2）文本提示有语言的模糊性，难以判断标识细节或者描述细节以及各种复杂的大件货物；（3）传统方法只能实现像素距离内的扰动和组合，解决不了仰视监控下视角不同的待检测对象成像变化的数据扩充问题。因此，通过挖掘自监督预训练模型的泛化能力，本文提出了一种基于多尺度特征的原型匹配检测框架。首先，借助视角可控的扩散模型合成多视角数据，物理对齐源域和目标域的几何分布，实现跨域域间差异的降低；然

后，使用基于聚类算法的原型记忆库，替代易平滑的特征均值原型法，完整保存数据的精细差异性；最后，设计深浅层特征分离策略，融合浅层特征和深层特征信息，在无需微调的前提下实现了开放场景中特种车辆的准确识别。

1　基于自监督视觉模型的特种车辆细粒度检测方法

易获得的训练数据和实际场景下数据分布的差异是图像识别的最大难点。从图 1 可以发现，危化品运输车与普通货车的区别在于罐体形状、车身菱形警示标志等局部特征，而大件运输车因所载货物种类多，类间形态差异大而易造成漏检、误检。源域训练数据以路面平视图像为主，而目标域监控视频为俯视视角，二者的视角偏差给目标检测网络的特征提取带来了非常大的困难。

本文提出了一种基于自监督视觉模型（DEIM，detection-oriented image modeling）的免训练域适应检测框架，该框架分为了离线原型库构建与在线级联检测两个核心阶段，如图 2 所示。在第 1 阶段，先用扩散模型合成对应的俯视样本，使源域物理分布和目标域对齐。然后，使用自监督视觉模型 DEIM（Tang et al.，2023）对目标实例进行多尺度特征解耦，再用聚类算法从数据中挑选出最具表征性的真实样本加入到原型记忆库中，从而充分涵盖特种车辆的各种模态外观特征。第 2 阶段为在线级联检测和匹配，即在线推理：用一个通用检测器快速定位感兴趣区域 ROI（ROI，region of interest）并去除背景信号干扰；用前述提取到的两个层级解耦特征来计算其与库原型的双流加权相似度，最后采用检索匹配算法进行精确分类。

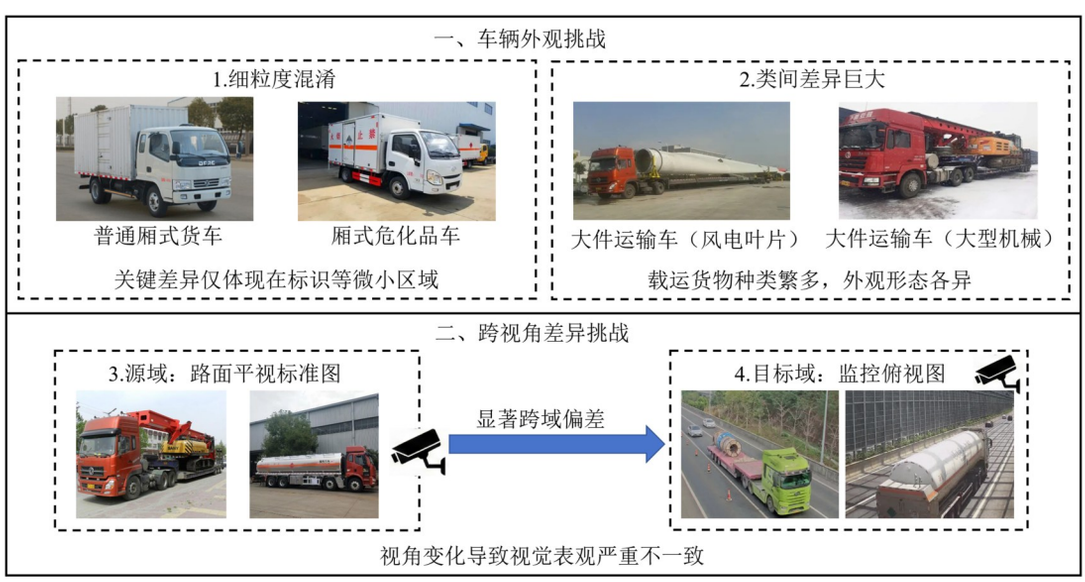

图 1　特种车辆细粒度外观混淆与跨域视角差异

Fig. 1　Illustration of fine-grained appearance confusion and cross-domain viewpoint shift

1.1　多视角合成模块

由于公开数据集中所见的数据多为路面平视采集，而监控场景多为俯视视角，因此本文引入了 Zero123 框架（Liu et al.，2023）来做三维视角合成，其目的就是从物理空间层面直接实现源域到目标域的几何对齐。不同于传统生成模型，Zero123 利用大规模 3D 数据集预训练的条件扩散模型，学习输入图像 x 与目标视角变化量 (R, T) 之间的映射关系，能将稀缺的平视标注样本合理、可控地转化为几何一致性极好的俯视监控样本。

鉴于源域路面数据集的原始拍摄角度主要在 0°~40°范围内，本文为了模拟真实路侧监控摄像头的架设角度，将目标视角增量参数设定为球坐标系下的相对变化量，在俯仰角（$\Delta\theta$）维度，将增量统一固定为 30°，通过与原始视角叠加，有效覆盖了监控场景中 30°~70°的常见俯视盲区；在方位角（$\Delta\phi$）维度，为充分模拟车辆在复杂路网中任意行驶方向多面的特征，选取了 30°、150°与 270°三个离散增量值进行增强。由此生成了不同俯视视角的合成图像序列，如图 3 所示。由此，本文方法在不增加采集成本的前提下，对目标域成像范围进行了扩充，提升了跨视角鲁棒的几何表达能力。

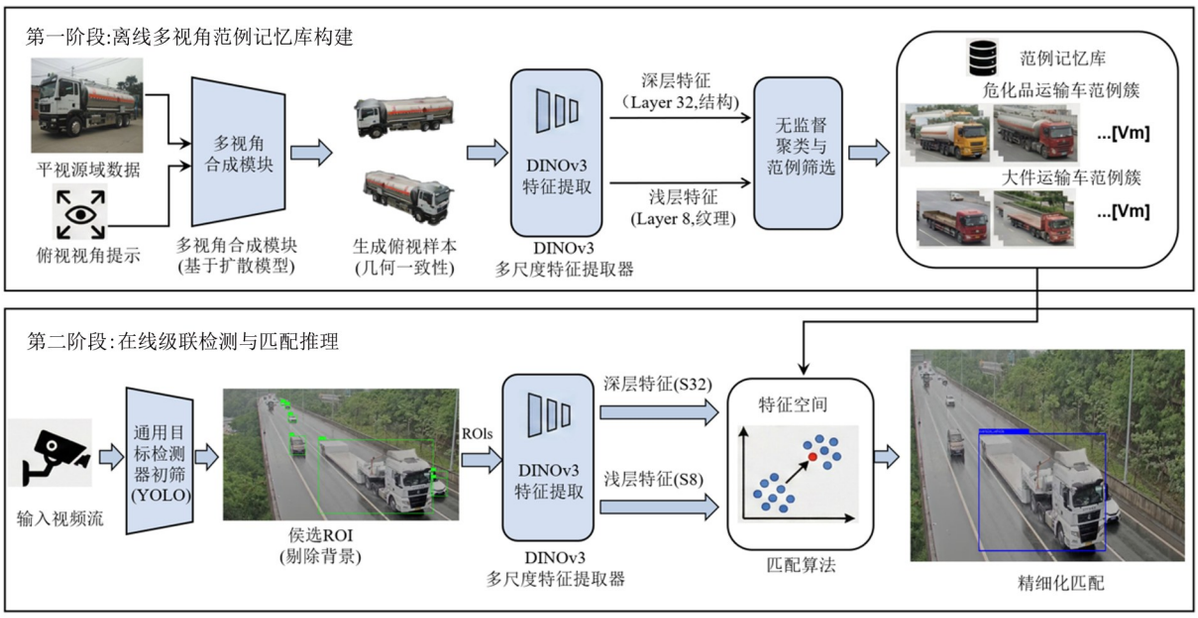

图2 总体框架图

Fig. 2 Overall framework diagram

图3 多角度生成车辆图像示例

Fig. 3 Examples of multi-view vehicle generated images

1.2 基于自监督视觉模型的多尺度特征编码网络

为了在无监督条件下合理地兼顾特种车辆的全局拓扑结构及局部细粒度纹理，本文构造了一个以自监督视觉模型（SSVMs，self-supervised vision models）为核心的多尺度特征编码网络。

1.2.1 自监督表征学习与特征金字塔构建 自监督模型的训练方式不同于依赖大规模标注数据的传统全监督检测模型，本文选用DINOv3作为特征提取的主干网络。该模型基于掩码图像建模（MIM，masked image modeling）与判别式自监督学习的联合训练，通过在海量无标注数据中重建被掩盖的图像区域和对比全局视图，使模型学习具有强泛化能力的视觉表征。这种训练机制使得DINOv3能够在缺乏特种车辆标签的情况下，依然能捕捉到车体、部件结构等更抽象的本质语义信息，为后续的聚类算法与检索提供所需的特征基础。

DINOv3是以标准的视觉Transformer（ViT）架构为基础设计的，因而有各向同性（Isotropic）特性：图像在初始切块（Patch Embedding）阶段空间分辨率即被固定，一般采用14或16的下采样步长（Stride），之后所有深层网络中所用特征尺度都单一不变。历代DINO模型，在相同模型规模下，所输出的特征尺度完全一致，因此高频细节信息诸如危化品标牌文字、细微裂纹在深层网络中容易被平滑或隐匿，不能直接满足对细粒度特征的严格要求。为了在不破坏DINOv3预训练权重完整性的前提下解决这一问题，本文引入DEIM框架所提出的空间调优适配器（STA，spatial tuning adapter）多尺度特征交互机制。该机制实质上是一个高效的特征适配器，可从DINOv3中间层显式、可靠地提取步长分别为8、16和32的多层级特征序列，记为$S8$，$S16$和$S32$。所设计的多尺度提取器网络结构，如图4所示。

1.2.2 尺度感知融合与特征解耦 由于深层抽象过程不可避免地会丢失高频细节，为了实现深层语

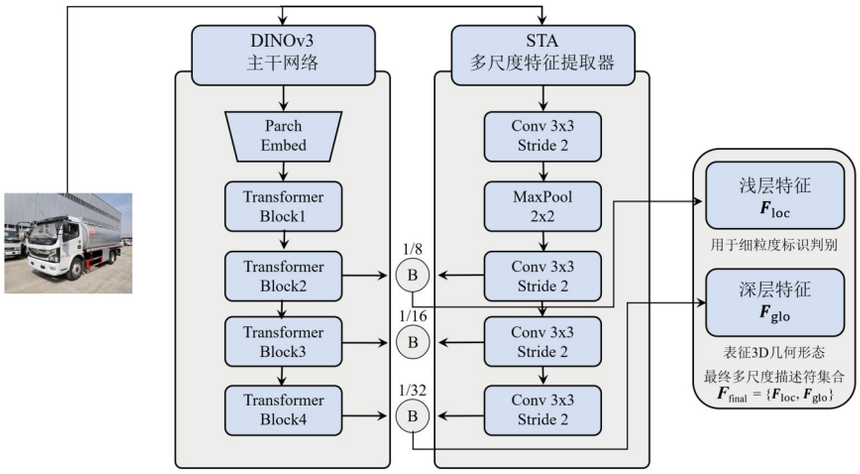

图4　多尺度特征提取器网络结构

Fig. 4　Architecture of the multi-scale feature extractor

义与浅层细节的良好互补，引入尺度感知师生对齐（STA）策略，设计了并行的空间细节分支：用卷积层的归纳偏置提取局部特征，又用预训练时所引入的多尺度对齐损失对其予以显式约束。即：

$$\mathcal{L}_{\text{align}}=\sum_{k\in\{S8,S16,S32\}}\lambda_k\left\|\mathcal{P}_k(Z^k_{\text{student}})-\text{StopGrad}(Z^k_{\text{teacher}})\right\|_2,$$

其中 Z^k 表示第 k 个尺度的特征映射。这一机制使得网络在对应小尺度 k 的浅层保留了局部纹理的信息。基于不同网络层级特征得到的车辆外观表征的差异性，本文选取有最大互补性的两层网络输出构建多尺度特征 $\mathcal{F}_{\text{final}}$。为了论证选择 $S8$（浅层）和 $S32$（深层）特征的合理性，用两张高相似度、易混淆的厢式货车和厢式危化品车辆（如图5所示）进行了敏感性分析，结果如表1所示。

图5直观地展示了高度聚焦局部纹理的 $S8$ 和倾向于整体结构轮廓的 $S32$ 特征在处理易混淆样本时的互补作用。选取两组典型的困难样本对：(a)外观极相似的普通厢式货车与危化品厢式车；(b)结构类似的混凝土搅拌罐车与危化品罐车。可视化结果显示，$S8$ 特征高度聚焦于车身表面的局部纹理。$S32$ 特征更倾向于整个车辆的结构轮廓。在区分罐车时，$S32$ 能够捕捉到搅拌车与危化品罐体在几何形状上的宏观差异，忽略了局部干扰。提取的空间特征集合为

$$\mathcal{F}_{\text{final}}(x)=\left\{\boldsymbol{F}_{\text{loc}}(x),\boldsymbol{F}_{\text{glo}}(x)\right\}\in\mathbb{R}^{H_1\times W_1\times D_{\text{loc}}}\times\mathbb{R}^{H_2\times W_2\times D_{\text{glo}}}.$$

表1　不同尺度层级相似度的敏感性分析[1)]

Table1　Sensitivity analysis of similarity across different scales

特征层级	原生 DINOv3	微调 DEIM	说明
$S8$（浅层）	0.995 1	**0.647 1**	主要负责捕获细粒度差异
$S16$（中层）	0.995 6	0.826 5	判别力与鲁棒性的折中，冗余度较高
$S32$（深层）	0.965 8	**0.847 2**	主要负责表征跨域不变的结构信息

1）相似度越低，代表特征对不同样本的区分度越高。

1.3　动态原型库与多原型投票推理机制

1.3.1　离线阶段：多视角联合特征聚类与建库

不同于传统方法仅使用单一视角的原始图像，本文引入1.1节的多视角合成模块得到多视角图像，构建视觉描述集合 $\Omega_{\text{aug}}=\{x_{\text{orig}},x_{v1},x_{v2},x_{v3}\}$。通过多尺度特征提取器提取样本的深层语义特征 $\boldsymbol{F}_{\text{glo}}$ 与浅层纹理特征 $\boldsymbol{F}_{\text{loc}}$，并经全局平均池化（GAP）压缩为特征向量。对两组向量执行独立的 L_2 归一化：

$$\hat{\boldsymbol{v}}_{\text{loc}}=\frac{\text{GAP}(\boldsymbol{F}_{\text{loc}})}{\|\text{GAP}(\boldsymbol{F}_{\text{loc}})\|_2},\quad \hat{\boldsymbol{v}}_{\text{glo}}=\frac{\text{GAP}(\boldsymbol{F}_{\text{glo}})}{\|\text{GAP}(\boldsymbol{F}_{\text{glo}})\|_2}.$$

为了自适应地平衡不同类别在语义与纹理上的依赖差异，引入加权参数 λ 构建联合特征向量：

$$\boldsymbol{v}_\lambda=\text{Normalize}\left(\left[\lambda\hat{\boldsymbol{v}}_{\text{glo}},(1-\lambda)\hat{\boldsymbol{v}}_{\text{loc}}\right]\right).$$

本文采用余弦轮廓系数作为优化目标，通过网格搜索确定最优权重 λ^*，使得生成的特征空间在同一类内更加紧凑，且不同类别间更加有区分性。在

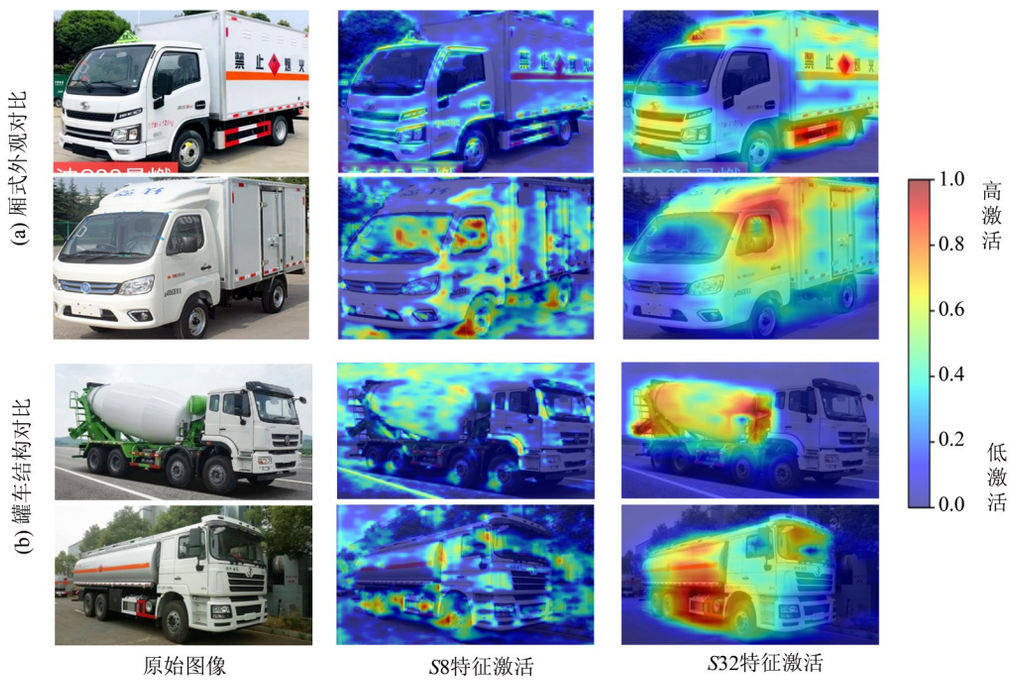

图5 危化品车与大件运输车的聚类结果可视化

Fig. 5 Visualization of clustering results for hazardous chemical and large transport vehicles

基于球面聚类的最优特征空间中，采用SphericalK-Means算法对每一类别的特征集合进行聚类。算法将每类数据划分为K个簇，每个簇中心$\boldsymbol{\mu}_k$代表该类别下一种代表性的外观和各种视角。为了节约推理开销，本文仅保留聚类产生的模态中心集合$\mathcal{P}=\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_N\}$作为动态原型库。这种策略将存储空间从$O(N_{\text{samples}})$显著压缩至$O(N_{\text{clusters}})$，在保留多视角多样性的同时实现了存储轻量化。每个星形的簇表示一个簇类，用一张原型图可视化表示；如图6所示，危化品车与大件运输车可分别聚类为35和42个模态簇。

1.3.2 在线阶段：由粗到精的层级化匹配算法

首先，将查询向量$\boldsymbol{v}_q$与原型库$\mathcal{M}$中所有聚类簇的中心向量$\boldsymbol{\mu}_k$进行余弦相似度计算，确定目标所属的模态子空间，检索出相似度最高的一个候选主簇：

$$k^*=\arg\max_k\left(\boldsymbol{v}_q\cdot\boldsymbol{\mu}_{k^*}\right),$$

其中k^*为选中簇的索引，该步骤能够迅速排除外观差异巨大的干扰类别，簇中心$\boldsymbol{\mu}_{k^*}$代表了该模态的平均特征，但其往往因平均化操作而丢失不同视角的几何细节。为了解决这一问题，本文选中当前簇后，在内部遍历其中存储的全量范例集合$\mathcal{E}_{k^*}=\{e_1,e_2,\cdots,e_N\}$。计算$\boldsymbol{v}_q$与簇内每一个具体范例$e_j$的相似度，并选取最大响应值作为最终的匹配置信度$S_{\text{final}}=\max_{e_j\in\mathcal{E}_{k^*}}(\boldsymbol{v}_q\cdot e_j)$。

上述流程可使查询样本在一个极为罕见的、离簇中心较远的俯视角度情况下被匹配出来，只需要簇内包含由Zero123生成的对应视角的合成范例，二者就能产生相对鲁棒的响应。这种机制显著提升了长尾场景下的匹配分数，确保了最终分类结果的高置信度与鲁棒性。匹配过程如图7所示。

2 实 验

2.1 数据集和评估指标

为了构建特种车辆的基础特征库，本文筛选并构建了一个源域数据集。该数据集主要包含由互联网采集的路面平视视角的车辆图像。车辆图像分为危化品运输车和大件运输车2个核心类别。为了验证模型的跨域泛化能力，收集了真实路侧高位监控视频帧，以构建目标域测试集。数据集组成如表2所示。

考虑到特种车辆监管对误检和漏检这两个指标的敏感性，采用精确率P衡量模型抗误判的能力。精确率反映了检测结果中真实目标所占的比例。采用召回率R衡量模型的覆盖能力，召回率表示模型能否在跨域俯视视角下检出所有目标。mAP@0.5是IoU阈值为0.5时的平均精度均值，用于评估检测器的综合定位与分类性能。

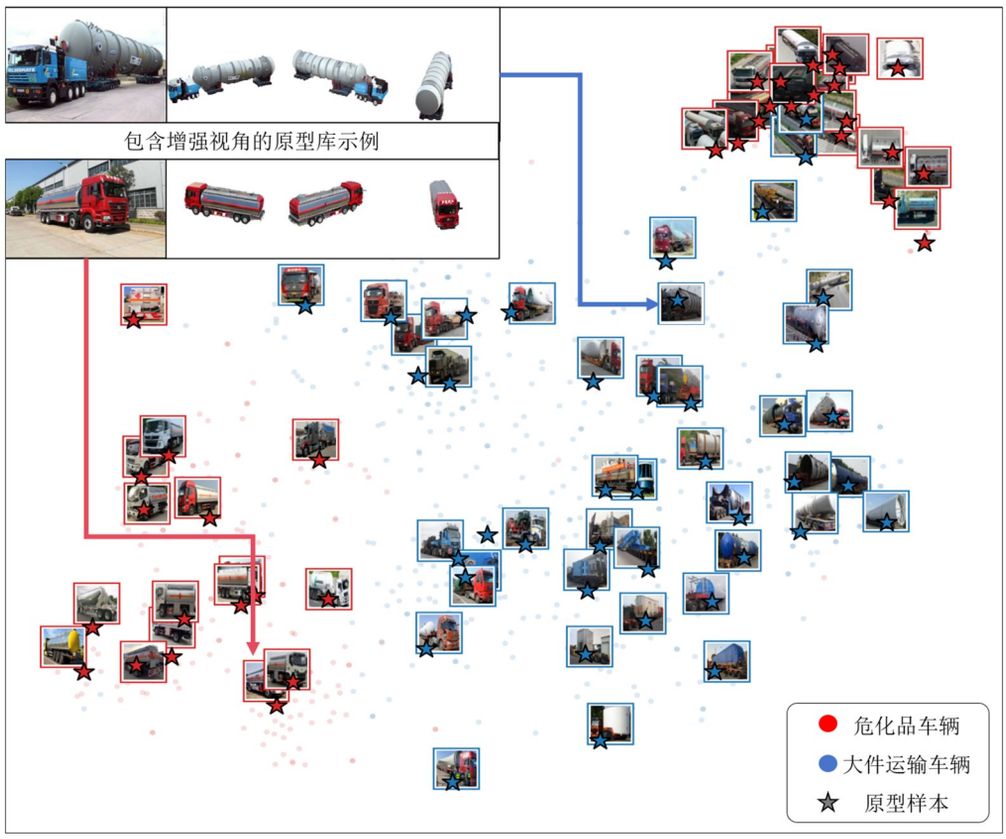

图 6　危化品车与大件运输车聚类结果可视化

Fig. 6　Visualization of clustering results for hazardous chemical and large transport vehicles

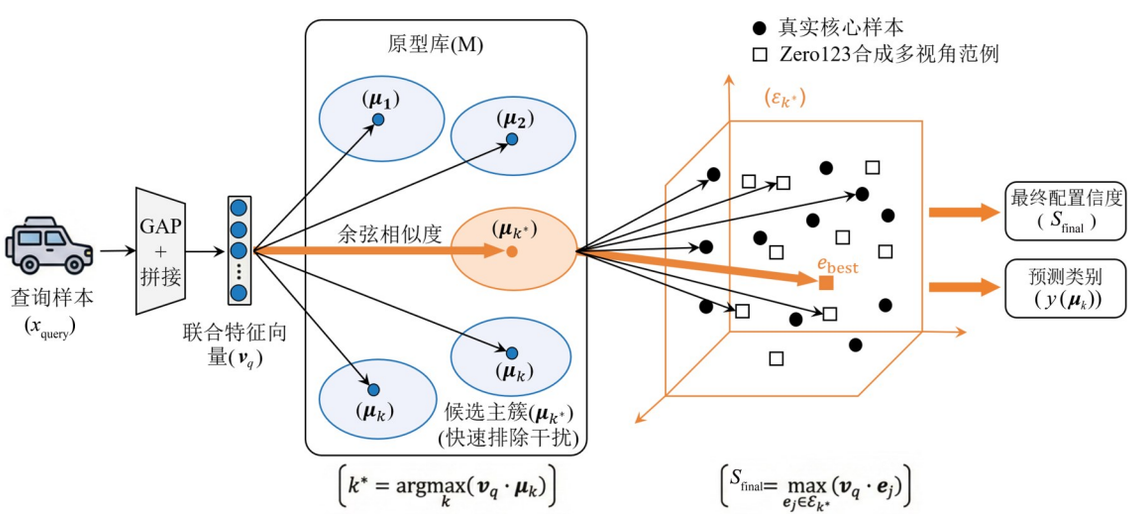

图 7　在线推理阶段的匹配过程

Fig. 7　The matching process in the online inference phase

2.2　实验设计

选用轻量级 DINOv3-Tiny 作为冻结的特征提取主干，加载 DEIMv2-S 适配器权重以显式重构 {*S*8, *S*16, *S*32} 多尺度特征金字塔。在离线建库阶段采用全量建库策略，用源域中全部 1 187 张图像作为支撑集，再用 Zero123 生成俯视增强样本，由此构造出包含多视角几何信息的完备原型库，进而用自适应聚类方法把危化品车和大件运输车分别聚类为 35 和 42 个模态簇。在线推理阶段，为了最大化召回率，采用预训练的 YOLO-World-L 作为前置检测器，使用包含“truck”，“tanker”等广义父类的提示词集及 0.1 的低置信度阈值生成候选区域。分数

表2 数据集组成

Table 2 Dataset composition

类别	源域实例数量	目标域实例数量	备注
危化品车	687	518	—
大件运输车	500	520	—
背景	—	500	无实例负样本

阈值设定为 $S_{\text{final}} = 0.60$。所有实验均在 NVIDIA RTX 3090 GPU 上基于 PyTorch 框架完成。

2.3 结果分析

将本文方法 SAFE（Self-supervised Adaptive Feature Extraction for Special Vehicles）与全监督检测器（YOLOv8, RT-DETR）、跨域/元学习检测器（YOLO-G, MetaR-CNN）、开放词汇检测器（Grounding DINO, YOLO-World）及视觉提示检测器（NTTT, YOLOE）在目标域进行对比，结果如表3所示。

表3 不同检测器在目标域的对比[1)]

Table 3 Comparison of different detectors in the target domain

方法		提示/适应模态	训练设置	*P*/%	*R*/%	mAP@0.5/%
全监督	YOLOv8-L	类别ID	全量微调	54.2	46.8	48.7
	RT-DETR-L	类别ID	全量微调	61.3	35.6	32.1
跨域目标检测	YOLO-G	类别ID	目标域无监督	**63.5**	49.7	**51.6**
	Meta R-CNN	类别ID	Meta-Finetune	58.4	38.1	43.2
开放词汇	Grounding DINO	文本描述	Zero-Shot	29.4	66.8	37.3
	YOLO-World	文本描述	Zero-Shot	34.7	**67.5**	41.6
视觉提示	NTTT	视觉	免训练	60.8	40.2	38.5
	YOLOE	视觉	免训练	54.6	38.7	35.2
本文方法	SAFE	视觉	免训练	**67.5**	**68.6**	**55.8**

1）表中数据均为各模型在本文构建的特种车辆目标域测试集上的评估结果。

实验结果表明，全监督模型和跨域模型尽管已经在源域上充分微调，但在目标域的召回率很低，比如 RT-DETR-L 的召回率只有 35.6%。这表明模型过度学习了源域的数据，很难适应不同域的差别。跨域模型中的 YOLO-G 模型受特种车辆自身的类间分布影响，导致性能下降，其 mAP@0.5 最高也只有 52%。开放词汇模型的召回率达到了 67%，但精确率较低（只有 35%），可以认为：文本提示对于细粒度检测具有较多歧义，难以有效区分外观相似的普通货车和特种车辆，导致累计了大量误报的故障。引入视觉提示后，相关模型的精确率有明显提高，但受限于固定的单一原型，召回率仍然很低（只有 40.2%）。相比于其他方法，本文方法能够达到各项指标综合效果最优。得益于浅层的纹理特征引入，本文模型的精确率可提高到 67.5%；由于记忆库中拥有更多的、完整视角的信息可以用来填补跨域的几何差异，本文模型的召回率提高到了 68.6%；本文模型的 mAP@0.5 为 55.8%，比其他方法高 4% 以上。因此，本文的方法具有有效性。

2.4 消融实验分析

2.4.1 核心组件贡献分析

对本文模型开展消融实验，实验结果见表4。基线模型的性能与之前提出的视觉提示方法接近，mAP@0.5 只有 35.8%。因为第1阶段使用了通用检测器的候选框，很好地保证了对潜在目标的全面检出，本文模型的召回率很高；而多视角合成的数据增强方法缓解了源域平视图像和目标域俯视监控之间的域差异，使得 mAP@0.5 达到了 51.7%；多尺度特征融合模块通过引入浅层纹理细节，增强了对外观相似类别的细粒度判别能力，将精确率从 46.5% 提升至 65.8% 以上，显著减少了误检。当双模块协同工作时，模型实现了精确率（67.5%）、召回率（68.6%）和 mAP@0.5（55.8%）的综合性能最佳，较基线模型提升近 20%。

2.4.2 超参数敏感性分析

为验证所选超参数的合理性并探究其对模型性能的影响，在目标域数据集上对特征融合权重 λ 和原型数量 K 进行敏感性实验，结果如图8所示。

λ 为调节深层语义与浅层纹理的权重系数。如图8(a)所示，检测性能呈显著“倒U型”，在 $\lambda = 0.65$ 处达峰值（55.8%）。λ 过低（<0.4）时会削弱匹配能力，致使 R 跌破 40%；λ 过高（>0.8）则会丢失关键纹理，导致 P 骤降至 36.2%。在 $\lambda = 0.65$ 时，有效平衡

表4　消融实验结果

Table4　Ablation study results

多尺度模块	多视角增强	P / %	R / %	mAP@0. 5/ %
×	×	59. 2	36. 5	35. 8
×	√	46. 5	66. 2	51. 7
√	×	65. 8	47. 6	45. 3
√	**√**	**67. 5**	**68. 6**	**55. 8**

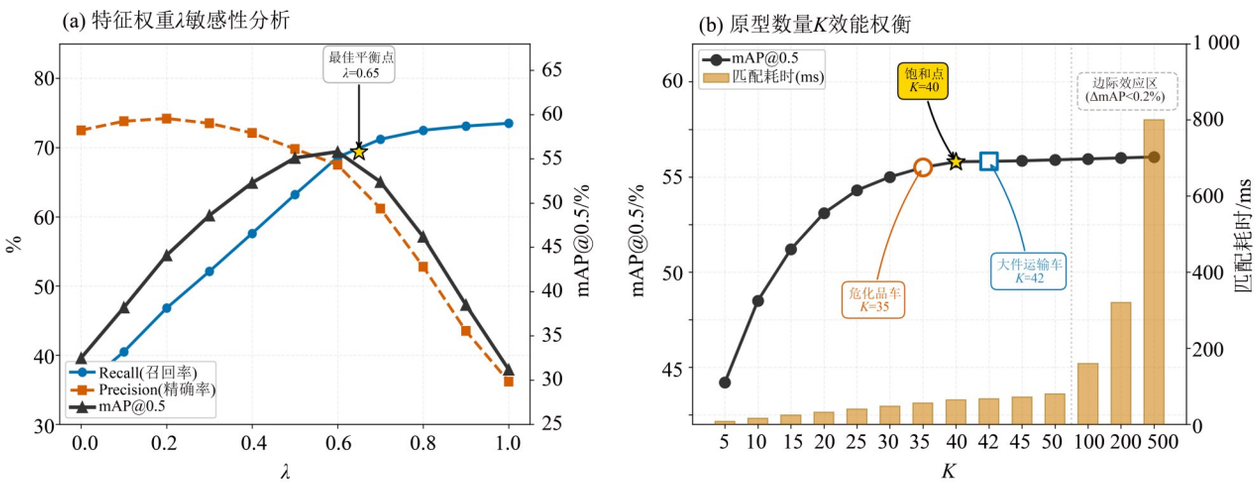

图8　超参数的敏感性分析

Fig. 8　Hyperparameter sensitivity analysis

了跨域性能的鲁棒性与细粒度判别力。

原型数量K决定了特征库的覆盖度。如图8(b)所示，mAP@0.5随K呈快速上升后饱和的趋势。K < 50时，细粒度原型显著增强；K > 50时性能进入饱和区，边际收益微弱(<0.2%)且推理耗时线性增长。因此，本文采用类别自适应策略，差异较大的大件运输车辆设K = 42，结构固定的危化品车设K = 35。而且，模型每次提取特征耗时40 ms，每次计算相似度耗时1.6 ms，所以本文方法在保证精度的同时实现了计算效率最优。

2. 5　可视化与定性分析

为了分析本文方法在实际路侧监控检测中的效果，由图9给出了不同类别目标的检测结果。图9中，黑色框为第1阶段检测到的目标候选框，彩色框是本文方法的检测结果。从图中可以看出，即使在复杂背景之下，模型对于特定外观特征的危化品运输车(红色框)不会出现检测错漏，对体型巨大且类内差异明显的大件运输车(黄色框)的识别结果同样准确。

图10展示了易误检、易混淆样本的检测结果。虽然厢式危化品车容易与普通物流厢式货车混淆，但本文方法通过车体细节将其正确识别为危化品车；对于常被误检为危化品车的粉罐车来说，模型也能够精准区分罐体的细节差异，实现对车辆的准确分类。在图10中，本文方法将易错车辆识别为普通车辆类别(绿色框)，有效避免了误报。

3　结　论

针对路侧交通监控场景下特种运输车辆的跨域检测问题，本文提出一种基于自监督视觉模型的免训练域适应检测方法。本文方法利用自监督视觉模型的泛化能力和其不同尺度网络输出特征的差异性进行算法设计，将检测准确率提升至了67.5%，有效避免了开放词汇检测模型的语义歧义问题；并通过生成的多视角合成数据对记忆库进行数据增强，在免微调模型的前提下使得模型保持了68.6%的检测召回率。实验结果表明：本文所提方法取得了检测实时性和精确性的平衡，显著优于现有的全监督及跨域检测方法。研究不仅验证了复杂场景下无需训练的原型匹配目标检测方法的有效性，也为缺乏目标域样本的特种车辆监管提供了一种高效、通用的解决方案。

图 9　复杂场景下的多类别车辆检测

Fig. 9　Multi-category vehicle detection in complex scenes

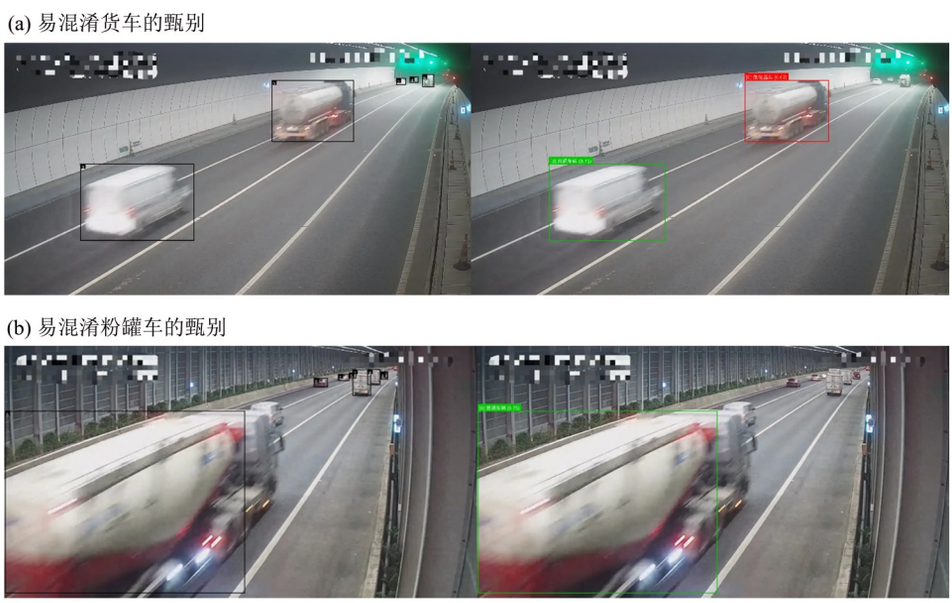

图 10　易混淆样本的检测与分类

Fig. 10　Classification and detection performance on fine-grained hard examples

参考文献：

李经宇，杨静，孔斌，等，2021. 基于注意力机制的多尺度车辆行人检测算法[J]. 光学精密工程，29(6)：1448-1458.

邱铭凯，李熙莹，2021. 用于车辆重识别的基于细节感知的判别特征学习模型[J]. 中山大学学报(自然科学版)，60(4)：111-120.

薛天朗，岳玉涛，2025. 基于文本-视觉多模态学习的交通目标识别与检索[J/OL]. 计算机应用与软件. https://www.shcas. net/cn/article/pdf/preview/cf7b460e-894b-4dbf-ba90-bc968e958623.pdf.

Bochkovskiy A，Wang C Y，Liao H M，2020. YOLOv4：Optimal speed and accuracy of object detection[PP/OL]. (2020-04-23)[2026-02-09]. https://doi. org/10.48550/arXiv.2004.10934.

Cheng T，Song L，Ge Y，et al，2024. YOLO-World：Real-time open-vocabulary object detection[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，WA，USA：16901-16911.

Cubuk E D，Zoph B，Shlensh J，et al，2020. RandAugment：Practical automated data augmentation with a reduced search space [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle，WA，USA：3008-3017.

Ghiasi G，Cui Y，Srinivas A，et al，2021. Simple copy-paste is a strong data augmentation method for instance segmentation [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Nashville，TN，

USA：2917−2927.

Hu J，Shen L，Sun G，2018. Squeeze−and−Excitation Networks［C］//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City，UT：7132−7141.

Huang S，Hou Y，Liu L，et al，2025. Real−time object detection meets DINOv3［PP/OL］.（2026−01−26）［2026−02−09］. https：//doi. org/10.48550/arXiv. 2509.20787. arXiv：2509.20787.

Li L H，Zhang P，Zhang H，et al ，2022. Grounded language−image pre−training［C］//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans，LA，USA：10955−10965.

Lin T Y，Dollar P，Girshick R，et al，2017. Feature pyramid networks for object detection［C］//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu，HI：936−944.

Liu R，Wu R，Van Hoorick B，et al，2024. Zero−1−to−3：Zero−shot one image to 3D object［C］//2023 IEEE/CVF International Conference on Computer Vision. Paris，France：9264−9275.

Liu S，Qi L，Qin H，et al ，2018. Path Aggregation network for instance segmentation ［C］//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City，UT：8759−8768.

Liu S，Zeng Z，Ren T，et al，2023. Grounding DINO：Marrying DINO with grounded pre−training for open−set object detection［PP/OL］.（2024−07−19）［2026−02−09］. https：//doi. org/10.48550/arXiv. 2303.05499. arXiv：2303.05499.

Oquab M，Darcet T，Moutakanni T，et al，2023. DINOv2：Learning robust visual features without supervision［PP/OL］.（2024−02−22）［2026−02−09］. https：//doi. org/10.48550/arXiv.2304.07193.

Poole B，Jain A，Barron J T，et al，2022. DreamFusion：Text−to−3D using 2D diffusion［PP/OL］.（2022−09−29）［2026−02−09］.https：//doi.org/10.48550/arXiv.2209.14988.

Rombach R，Blattmann A，Lorenz D，et al，2022. High−Resolution image synthesis with latent diffusion models［C］//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans，LA，USA：10674−10685.

Shi Y，Wang P，Ye J，et al，2023. MVDream：Multi−view diffusion for 3D generation［PP/OL］.（2024−04−18）［2026−02−09］.https：//doi.org/10.48550/arXiv.2308.16512.

Tang J，Ren J，Zhou H，et al，2023. DreamGaussian：Generative Gaussian splatting for efficient 3D content creation［PP/OL］.（2024−03−29）［2026−02−09］. https：//doi.org/10.48550/arXiv.2309.16653.0

Wei J，Wang Q，Zhao Z，2023.YOLO−G：Improved YOLO for cross−domain object detection［J］. Plos One，18（9）：e0291241.

Woo S，Park J，Lee J Y，et al，2018. CBAM：Convolutional block attention module［M］//Computer Vision−ECCV 2018. Cham：Springer International Publishing：3−19.

Yun S，Han D，Chun S，et al，2004. CutMix：Regularization strategy to train strong classifiers with localizable features［C］//2019 IEEE/CVF International Conference on Computer Vision. Seoul，Korea：6022−6031.

Zhang H，Cisse M，Dauphin Y N，et al，2017. Mixup：Beyond empirical risk minimization［PP/OL］.（2018−04−27）［2026−02−09］. https：//doi. org/10.48550/arXiv.1710.09412.0

Zhang L，Rao A，Agrawala M，2023. Adding conditional control to text−to−image diffusion models［C］//2023 IEEE/CVF International Conference on Computer Vision. Paris，France：3813−3824.

（责任编辑　王海蓉）