中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-17

论文引用格式:Liu Zhen, Yang Qinzhe, Liu Liqin, Liu Chenyang, Zou Zhengxia, Shi Zhenwei. Generation-detection unified method for giant panda object detection[J/OL]. Journal of Image and Graphics, XXXX:1-17. DOI: 10.11834/jig.260078.(刘祯, 杨沁哲, 刘丽芹, 刘辰阳, 邹征夏, 史振威. 面向大熊猫目标检测的生成-检测统一方法[J/OL]. 中国图象图形学报, XXXX:1-17. DOI: 10.11834/jig.260078.)[DOI:10.11834/jig.260078]

面向大熊猫目标检测的生成-检测统一方法

刘祯[1],杨沁哲[2],刘丽芹[3],刘辰阳[3],邹征夏[3],史振威[3*]

1. 北京航空航天大学国家卓越工程师学院,北京　100191; 2. 北京航空航天大学沈元学院,北京　100191; 3. 北京航空航天大学宇航学院,北京　100191

摘　要:**目的**　大熊猫作为全球生物多样性保护的旗舰物种,其在相机陷阱图像中的精确检测对生态评估与保护决策至关重要。然而,标注数据稀缺且预训练数据与野外图像存在域差异,限制了通用检测器在野外环境中的实用性。为此,本文提出一种集成生成模型与检测模型的统一生成-检测方法——PandaGenDet。**方法**　该方法通过生成模型合成图像以缓解数据资源的限制,并通过结构改进提升了检测模型在野外环境下的鲁棒性。具体而言,为生成模型设计了类别引导机制,增强生成图像的语义一致性。在检测模型中构建即插即用的图像增强器模块,调整野外图像至更适应检测器预训练权重的分布;进一步地,提出生成特征注入器,将生成模型中蕴含的多尺度语义先验迁移至检测网络。**结果**　实验表明,类别引导机制使生成图像的KID(kernel inception distance)从0.059改善至0.038,FID(fréchet inception distance)由147.00降至123.13;图像增强器使检测模型在大熊猫数据上的mAP(mean average precision)由88.8提升至89.7,mAR(mean average recall)由94.9提升至95.5;在此基础上,加入生成特征注入器模型的mAP达89.8,最终联合合成图像继续训练模型的mAP提升至90.1,并表现出良好的开放集检测能力。**结论**　PandaGenDet建立了一个从数据合成到目标检测的统一协同架构,通过数据级合成缓解样本稀缺、图像级增强缩小域间分布差异、特征级注入复用生成模型的语义表征,实现了三重维度的深度协同,显著提升了通用检测模型在复杂野外环境下的大熊猫检测性能。

关键词:目标检测;大熊猫;图像生成;合成数据;深度学习

Generation-detection unified method for giant panda object detection

Liu Zhen[1], Yang Qinzhe[2], Liu Liqin[3], Liu Chenyang[3], Zou Zhengxia[3], Shi Zhenwei[3*]

1. National College for Excellent Engineers, Beihang University, Beijing 100191, China; 2. Shen Yuan Honors College, Beihang University, Beijing 100191, China; 3. School of Astronautics, Beihang University, Beijing 100191, China

Abstract: **Objective** Giant pandas serve as a flagship species for global biodiversity conservation and play a key role in assessing ecosystem integrity and conservation effectiveness. Accurate and reliable detection of giant pandas in camera-trap images is thus essential for long-term wildlife monitoring, population assessment, and adaptive management of protected areas. In recent years, deep learning-based object detection algorithms have demonstrated remarkable success. However, directly deploying general-purpose detection models in wild scenarios remains challenging due to two fundamental issues. First, giant pandas are rare species, and acquiring large volumes of high-quality, finely annotated training data from the

收稿日期:2026-02-03;**修回日期**:2026-04-20

*通信作者:史振威　shizhenwei@buaa.edu.cn

基金项目:国家自然科学基金(62125102,62471014,62501026,U24B20177,624B2017);中央高校基本科研业务费专项资金项目

Supported by: National Natural Science Foundation of China(62125102,62471014,62501026,U24B20177,624B2017); Fundamental Research Funds for the Central Universities

wild is extremely costly, time-consuming, and often impractical. Second, there exists a substantial domain gap between commonly used pre-training datasets and unconstrained camera-trap images captured in natural habitats. To alleviate data scarcity and improve robustness in wild environments, we propose a unified generation-detection method termed PandaGenDet. **Method** Rather than treating data augmentation and detection as independent components, the core idea of PandaGenDet is to improve detection robustness through multi-level collaboration between generative and discriminative models. Specifically, PandaGenDet consists of three complementary components operating at different representational levels. First, at the data level, we introduce a class-conditioned image generation module equipped with a Category-guidance Mechanism. This mechanism explicitly incorporates semantic category information into the generative process, guiding the synthesis of panda images with improved semantic consistency and target realism, making them more suitable as high-quality supplementary training samples. Second, at the image level, we design an Image Enhancer module to reduce the domain discrepancy between wild camera-trap images and the visual priors learned from large-scale pre-training datasets. The Image Enhancer is implemented as a modular and easily integrable component that performs a learnable image-level mapping prior to detection. By adaptively reshaping low-level and mid-level image statistics, this module maps target-domain images to representations that are more compatible with the detector's pre-trained weights, without requiring any modification to the detector architecture. During training, the Image Enhancer and the detector are jointly optimized in an end-to-end manner, with all detector parameters fully fine-tuned from their pre-trained initialization. Third, at the feature level, we propose a Generative Feature Injector, which leverages the trained generative model as a multi-scale feature extractor. Hierarchical feature representations learned during the image generation process are extracted and injected into the detection backbone via a PSPNet (pyramid scene parsing network) and FPN (Feature Pyramid Network) fusion network. This design enables the detector to leverage rich semantic and structural priors embedded within the generative model, enabling the transfer of multi-scale semantic priors from the generative model into the detection network. Together, these mechanisms form a unified and extensible method for robust wildlife detection. **Result** We conduct extensive experiments using Grounding DINO, a modern open-set object detection model, as the detection backbone. Evaluations are performed on the giant panda subset of the LoTE-Animal (long time-span dataset for endangered animal) dataset, which contains challenging camera-trap images representative of real-world conservation scenarios. Experimental results demonstrate that the proposed Category-guidance Mechanism significantly improves generative quality. Specifically, KID (kernel inception distance) decreases from 0.059 to 0.038, while FID (fréchet inception distance) is reduced from 147.00 to 123.13, indicating that the synthesized images achieve higher fidelity and improved semantic consistency with real wild panda images. These improvements directly translate into more effective training data for detection. When the Image Enhancer is integrated into the Grounding DINO detector, notable gains in detection performance are observed. On the LoTE-Animal panda subset, mAP (mean average precision) increases from 88.8 to 89.7, while mAR (mean average recall) improves from 94.9 to 95.5, confirming the effectiveness of image-level domain adaptation. Further incorporating the Generative Feature Injector leads to additional performance improvements, with the detector achieving a mAP of 89.8, outperforming both the baseline and image-enhancer-only configurations. Finally, training the detector using a mixture of real images and high-quality synthetic images generated by the full PandaGenDet pipeline yields the best overall performance, achieving a final mAP of 90.1. Qualitative analyses further reveal that synthesized images exhibit more accurate panda poses, better integration with realistic environmental textures, and fewer semantic artifacts. Detection visualizations demonstrate high localization accuracy in challenging scenarios, including dense vegetation, low illumination, and partial occlusion. Furthermore, the final model demonstrates strong robustness in open-set detection, maintaining stable performance even when encountering object categories not present in the training dataset. **Conclusion** This study presents PandaGenDet, a unified collaborative framework from data synthesis to object detection for giant panda monitoring in complex wild environments. By integrating data-level synthesis, image-level enhancement, and feature-level injection in a unified manner, the proposed method effectively addresses two major bottlenecks in real-world wildlife detection: the scarcity of annotated data and the presence of severe domain gaps between pre-training and deployment scenarios. Extensive experiments on camera-trap datasets demonstrate that PandaGenDet substantially improves both synthetic image fidelity and detection accuracy, while also enhancing open-set robustness. Through a three-dimensional collaborative strategy—data-level synthesis, image-level enhancement, and

feature-level injection—PandaGenDet significantly improves the detection performance of general-purpose models in complex wild environments.

Key words: object detection; giant panda; image generation; synthetic data; deep learning

0 引 言

大熊猫作为全球生物多样性保护的旗舰物种，其种群动态不仅反映区域生态系统的健康状况，也是衡量自然保护区管理效能的重要指标(Chen 等，2020)。因此，构建高效可靠的监测体系对大熊猫保护至关重要。在众多野生动物监测技术中，相机陷阱凭借非侵入式、全天候数据采集等优势，已成为野生动物监测的主要手段(Nguyen 等，2017)。面对日益增长的相机陷阱图像数据，如何筛选出大熊猫图像成为亟待解决的问题。广泛应用于区域监控、自动驾驶等领域的目标检测技术(焦李成 等，2023)已被证明是一种低成本且可靠的解决方案，具有长期稳定工作和应用范围广等优点(Zhao 等，2021)。

现代目标检测方法大致可以分为两类：以 R-CNN(region-based convolutional neural network)(Girshick 等，2014；Girshick 等，2015)为代表的两阶段检测算法和以 YOLO(you only look once)(Redmon，2016)和 SSD(single shot multibox detector)(Liu 等，2016)为代表的单阶段检测算法。前者以高精度见长，而后者则具有检测速度快、适应部署需求的优势(罗会兰和陈鸿坤，2020)。得益于上述两种检测框架的发展，当前最先进的目标检测算法在 MS COCO(microsoft common objects in context)(Lin 等，2014)、Pascal VOC(pascal visual object classes)(Everingham 等，2015)等通用基准数据集上已表现出了超越人类的能力。然而，当这些方法被直接应用于复杂的野外大熊猫场景时却面临着巨大挑战。挑战主要来源于两个方面：数据资源限制和野外图像复杂性。野外大熊猫图像天然具有稀缺性，且高质量标注常需依赖动物学专家完成，导致数据获取成本高、标注难度大。另一方面，野外图像常伴随光照变化剧烈、前景遮挡严重等问题(孟继森 等，2025)，与现有通用目标检测模型基于大规模自然图像预训练学习到的视觉先验在成像条件、纹理分布和目标尺度等方面存在显著域差异，导致在野外大熊猫检测任务中性能受限。

针对目标检测任务中训练数据不足的问题，研究者提出了多种数据增强方法。传统方法多依赖图像旋转、裁剪等基本数据增强技术，或采用图像擦除(DeVries 和 Taylor，2017；Zhong 等，2020)和图像混合(Yun 等，2019；Zhang 等，2017)等更高级的增强方式。然而，这些增强方式难以生成兼具真实感、多样性和语义一致性的图像样本，可能对下游检测任务的实际效果影响不均(郭继昌 等，2022)。生成模型的发展为此提供了新的解决思路，越来越多的研究开始探索通过生成模型生成合成图像以辅助目标检测训练的路径(张珂 等，2025a；张永飞 等，2025b；郑天鹏 等，2025)。生成目标检测数据的关键在于，如何在生成包含目标图像时同步获得准确的边界框标注。早期方法独立生成前景与背景图像，通过将前景粘贴至背景中的指定位置构建目标检测数据。近期研究则尝试联合建模目标与背景，通过输入边界框来控制模型在指定位置合成目标。值得注意的是，上述方法均聚焦于通用检测领域，尚未应用于大熊猫检测任务。

对于野外图像环境的高度复杂性，现有大熊猫检测研究主要从模型结构入手，通过引入定制化模块以提升检测性能。方法通常围绕多尺度特征融合、注意力机制、网络结构重设计以及上下文信息建模等策略展开，旨在增强模型对复杂背景、剧烈光照变化和严重遮挡等干扰因素的鲁棒性，并在检测精度、推理速度与模型规模之间取得更优平衡，从而更好地适配野外应用场景的实际需求。

为解决大熊猫检测中的挑战，我们提出了一个集成图像生成与目标检测的统一生成-检测方法——PandaGenDet。该框架构建了“数据-图像-特征”三重协同优化：在数据维度，通过类别引导的生成模型提升合成数据的语义一致性与真实性；在图像维度，将图像增强器引入检测模型，通过可学习的像素映射使野外图像主动适应检测器的预训练权重；在特征维度，为检测模型设计生成特征注入器，将生成模型在建模野外分布时积淀的语义表征跨任务迁移至检测网络。通过数据级合成、图像级增强与特征级注入的三重协同优化，显著提升了检测模

型在复杂野外环境下的检测性能。据我们所知,本文首次将生成式数据增强和开放集目标检测方法应用于大熊猫检测任务,实现了对新物种的零样本泛化,为跨域检测与生态监测提供了新的思路与参考范式。

本文主要贡献可总结如下:1)以大熊猫为研究对象,实现了一个集成图像生成模型与目标检测模型的统一生成-检测方法——PandaGenDet。2)为生成模型设计了类别引导机制,使用文本特征对易产生语义偏移的野外参考目标图像进行约束与校正,以合成高质量的检测数据。3)对于检测模型,构建了一个即插即用的图像增强器,能够将复杂的野外图像转化为适应检测器预训练特征空间的增强图像;加入了生成特征注入器,从生成模型中抽取多层多尺度的语义表征,实现生成先验与检测特征的融合。

1 相关工作

1.1 面向目标检测的生成式数据增强

近年来,图像生成技术发展迅猛(龚帅 等,2025;叶国升 等,2023),基于大规模数据集训练的扩散模型在图像质量与细节还原方面取得了显著进展(Borji,2022;刘安安 等,2024a)。诸如Stable Diffusion(Rombach 等,2022)、DALL·E(Ramesh 等,2021)和Imagen(Saharia 等,2022)等先进模型,已能生成细腻且逼真的图像。随着生成模型在图像质量和细节保真度方面的显著提升,利用高质量合成图像辅助训练深度学习模型逐渐成为现实且具有吸引力的研究方向。这一设想最初被应用于图像分类任务(Azizi 等,2023;He 等,2022;Li 等,2023c;Sarıyıldız 等,2023;Shipard 等,2023;Zhou 等,2023),并取得了积极成果。然而,若希望将其推广至目标检测领域,则必须确定前景目标在背景图像中的空间位置。根据前景目标与背景图像的生成策略,现有研究大致可分为两类:(1)前景与背景分别生成。Ge等人(2022)基于类别文本生成前景目标图像并提取其掩码,再单独生成背景图像并将两者合成。Lin等人(2023)将文本生成前景目标、显著性剪切-粘贴、以及基于CLIP(contrastive language-image pre-training)的样本筛选相结合。(2)前景与背景联合生成。DetectorGAN(Liu 等,2019)将目标检测器嵌入GAN(generative adversarial network)训练流程中,引导模型生成最有利于检测器训练的图像。ODGEN(Zhu 等,2024)提出以目标边界框为条件的扩散式图像生成方法,提升了合成图像的结构合理性。Fang等人(2024)构建了一个结合可控扩散模型与CLIP的合成数据增强管道,用于提升检测性能。InstaGen(Feng 等,2024)则基于扩散模型,直接生成包含实例级目标及其对应边界框的图像,用于训练目标检测模型。

1.2 大熊猫目标检测

目标检测旨在通过对图像内容的深度理解,实现对目标位置和类别的精确判定(许德刚 等,2021;赵永强 等,2020)。大熊猫属于濒危野生动物,其目标检测研究是在野生动物目标检测领域内的的一个重要分支。在野生动物目标检测领域,现有研究大致可分为以下两类:1)通过数据增强与自适应学习提升数据稀缺下的检测性能。Feng和Li(2022)提出了一种结合地理空间约束的自适应嵌入网络,利用数据增强策略和上下文感知模块,有效增强了图像样本生成并优化了模型训练效果。2)在主流检测框架上引入定制化模块,以应对特定领域挑战。Roy等人(2023)在YOLOv4中集成残差块和密集连接模块,以提升特征提取与传递能力,并结合空间金字塔池化与改进路径聚合网络强化特征融合效果。Su等人(2024)重新设计了YOLOv5的C3模块,以降低野外干扰特征的权重,另一方面引入WIoU边界回归损失函数以缓解低质量图像对检测精度的影响。

而具体到大熊猫目标检测,其研究思路与野生动物目标检测在总体上趋于一致,主要围绕引入领域适应与任务定制化的改进策略展开。Fang等人(2020)在SSD框架中引入小波变换以同时提取空间域的轮廓特征与频率域的高频纹理特征,使得模型能够更好地利用熊猫的纹理高频信息,从而实现野外大熊猫的快速且高精度检测。Wang等人(2022)将基于注意力的深度目标检测与时空/位置信息构建的上下文记忆库以及独立的物种分布模型融合,通过自注意力与坐标注意力提取多尺度语境并计算联合概率,从而提升了低可见性红外相机图像中野生大熊猫的检测性能。吕皓天和贾小林(2024)在YOLOv5n的颈部引入融合注意力的深度可分离结构并在边界框回归中采用Alpha-IoU损失,提升了复杂环境下的大熊猫定位与识别精度。整体来看,现

有研究以YOLO系列模型为主，持续在特征增强、结构优化和轻量化等方面展开探索，面向实际部署场景演进相关技术。

2 方　法

为应对野外大熊猫检测中标注数据稀缺及预训练数据与野外图像之间的域差异问题，我们提出了一种融合图像生成模型与目标检测模型的统一生成-检测方法——PandaGenDet，在合成高质量目标检测数据的同时，增强检测模型在野外图像场景中的适应能力。本节将依次介绍该方法的整体设计思路，并详细阐述生成模型与检测模型的结构与改进策略。

2.1 概述

方法的总体设计目标是通过可控且逼真的合成图像有效扩展数据分布，并从模型层面增强检测器对野外复杂场景的适应能力，从而显著提升数据稀缺条件下的复杂野外环境中大熊猫检测性能与泛化能力。图1展示了我们提出的统一生成-检测方法PandaGenDet的整体框架。

框架由生成与检测两个阶段构成。1）数据生成阶段：图像生成模型将多样化的大熊猫目标合成到真实野外背景中，从而在保持背景分布真实性的同时丰富目标样本空间。其中，类别引导机制用于缓解参考野外参考目标图像语义偏移，增强合成数据的真实性与稳定性。2）目标检测阶段：目标检测模型对合成图像进行伪标签标注，并将合成图像数据集与真实图像数据集联合用于模型的继续训练。其中，图像增强器对野外输入进行分布对齐，使其更适配检测器的预训练表示空间；生成特征注入器将生成模型中捕捉的结构与语义先验注入检测网络，以增强模型鲁棒性。通过将图像生成与目标检测相统一，本方法实现了从高质量数据合成到检测模型领域适应性优化的联合设计。

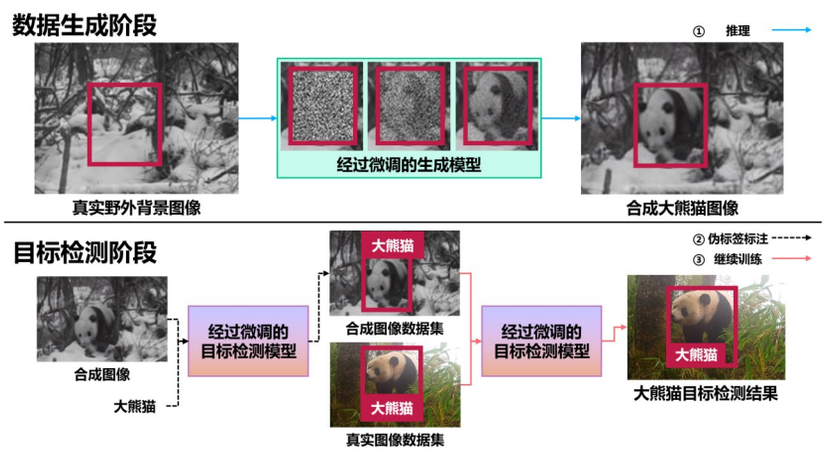


图1　大熊猫生成-检测方法——PandaGenDet

Fig. 1　PandaGenDet: a generation-detection unified method for giant panda object detection

2.2 基于类别引导的图像生成模型

通过在图像生成过程中对生成区域及目标外观多样性的精准控制，可以为后续检测训练提供具有稳定语义、准确边界和多样外观的高质量补充样本。由于文本条件难以充分描述大熊猫等特定物种的细粒度结构与外观特征，本文采用基于图像条件的生成范式，以引入类别相关且具有判别性的视觉先验，提升合成图像的真实性与多样性。在众多基于图像条件的生成模型中，Paint by Example通过引入参考图像并结合掩码约束，在指定区域内对齐并迁移示例图像的语义结构与外观模式，在保持整体一致性的同时实现精细化、可控化的内容生成（Yang等，2023）。在野外场景中，参考目标图像常受到光照变化剧烈、分辨率较低以及遮挡严重等因素的影响，导致其视觉特征存在偏差。为此，我们设计了类别引导机制：将目标类别的文本先验与参考图像的视觉特征融合作为条件向量，以修正参考图像特征偏差并增强生成结果的语义一致性与可控性。

2.2.1 Paint by Example图像生成模型

Paint by Example模型的训练流程如图2所示。

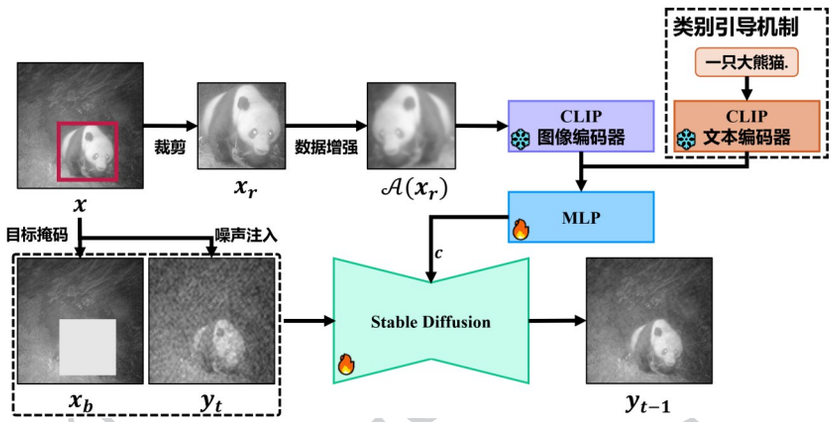


图2　大熊猫图像生成训练流程图

Fig. 2　Giant panda image generation training flow

使用目标检测数据集进行训练，对于给定的原始完整图像$\boldsymbol{x}$和目标边界框，使用边界框作为掩码图像$\boldsymbol{m}$。模型的输入包括去除目标的背景图像$\boldsymbol{x}_b$以及注入噪声的完整图像$\boldsymbol{y}_t$。参考图像$\boldsymbol{x}_r$经过数据增强处理$\mathcal{A}$后，首先通过冻结的CLIP图像编码器提取语义特征，并接入一个可训练的MLP（multi-layer perceptron）以生成条件向量$\boldsymbol{c}$。Stable Diffusion在条件向量$\boldsymbol{c}$的引导下，根据输入的$\boldsymbol{x}_b$和$\boldsymbol{y}_t$，预测$\boldsymbol{y}_t$中含有的噪声$\boldsymbol{\epsilon}_{t-1}$，并将图像还原为$\boldsymbol{y}_{t-1}$。

2.2.2　类别引导机制

野外图像常伴随强烈光照变化、分辨率降低等特点，冻结的CLIP图像编码器在此类图像上的语义提取能力受限，可能导致引导方向偏移和生成图像误差。为克服上述问题，我们提出了类别引导机制，以提升生成模型在野外环境下的语义准确性和目标清晰度。我们将类别信息用于辅助生成过程，利用目标检测数据集现有的类别标签，在无需额外标注的情况下，对条件向量进行进一步引导。使用文本特征为模型提供目标类别的强先验，以缓解参考图像特征偏差，进一步提升生成图像的语义一致性、合理性与可控性。

具体而言，对于输入图像$\boldsymbol{x}$和目标边界框，在获取边界框作为掩码$\boldsymbol{m}$的同时，提取边界框的类别，并格式化的构建标签l（如：一只大熊猫）。使用冻结的预训练CLIP文本编码器对类别标签l进行嵌入，得到类别特征向量（维度与CLIP图像编码器的输出相同）。随后，通过逐元素加法将参考图像特征与类别特征进行融合，共同作为MLP的输入，以生成最终的条件向量：

$$\boldsymbol{c} = MLP\left(CLIP\left(\mathcal{A}\left(\boldsymbol{x}_r\right)\right) + \lambda_{label} CLIP\left(l\right)\right) \tag{1}$$

式中，λ_{label}为可学习的权重系数，用于平衡图像特征与文本特征对条件向量的相对贡献。在去噪过程中，类别特征持续引导噪声预测方向，使生成结果更符合目标类别语义，对参考图像的语义偏差进行校正，从而提升生成图像的语义一致性与稳定性。

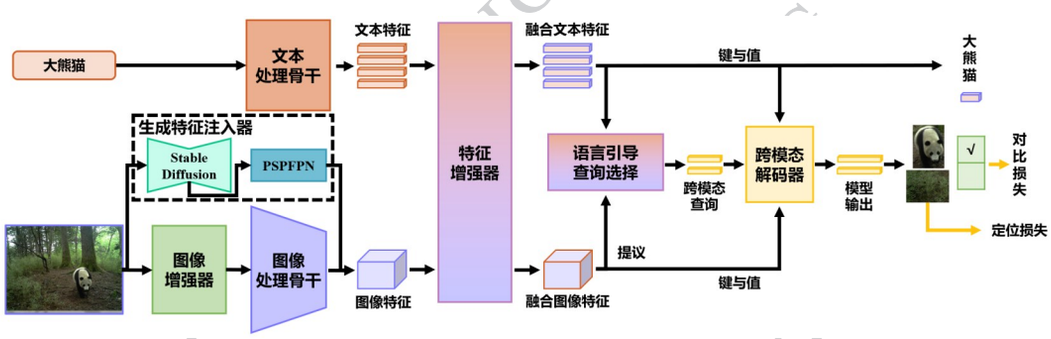


图3　大熊猫目标检测模型结构图

Fig. 3　Giant panda object detection model structure

2.3 结合图像增强与生成特征的目标检测模型

现有大熊猫目标检测方法多依赖单一视觉特征，在复杂自然场景下易受到光照变化、背景干扰及个体姿态差异的影响，导致检测鲁棒性不足。为提升模型在真实野外环境中的检测性能，本文选用Grounding DINO（Liu 等，2024b）作为基础检测框架。该模型通过将目标检测任务与语言语义信息进行深度融合，在目标定位精度和复杂背景建模等方面表现优异，已在多个基准数据集上取得领先结果。

然而，野外图像与检测器预训练数据之间存在显著域差异，生成模型蕴含的有效表征也尚未被充分利用。为此，本文在Grounding DINO框架基础上构建了图像增强器与生成特征注入器。图像增强器通过多尺度上下文建模与细节特征强化的协同机制，自适应地将复杂野外场景图像映射到与检测器联合优化的特征空间中；生成特征注入器则充分挖掘扩散生成模型中蕴含的多尺度语义信息，并将其以特征级方式注入至检测器骨干网络中，从而有效提升模型在复杂野外环境下的表征能力与检测鲁棒性。

2.3.1 Grounding DINO目标检测模型

Grounding DINO建立在DETR（detection transformer）（Carion 等，2020）的改进版本DINO（detr with improved denoising anchor boxes）（Zhang 等，2022）之上，通过引入紧密模态融合和大规模预训练，能够深度理解广泛的语义概念，具备开放集检测能力。如图3所示，Grounding DINO采用双编码器-单解码器架构。图像端使用Swin Transformer（shifted window transformer）（Liu 等，2021）提取多尺度视觉特征，而文本端通过BERT（bidirectional encoder representations from transformers）（Devlin 等，2019）生成文本嵌入。图像和文本特征通过紧密模态融合实现交互，最终预测目标的边界框与类别信息。

2.3.2 图像增强器

为有效利用可微调预训练模型的特征提取能力，我们设计了基于自适应特征增强（Ali 等，2024）的即插即用图像增强器。该增强器通过多尺度上下文建模与细节增强的双路径机制，完成图像到图像的映射过程，能够将复杂的野外图像转化为适配检测器预训练特征空间的增强图像，从而改善模型在背景复杂、目标遮挡等场景下的检测性能。在实际训练过程中，图像增强器与目标检测模型采用端到端的联合优化方式，对检测模型参数进行全量微调，以充分释放增强特征对检测性能的促进作用。

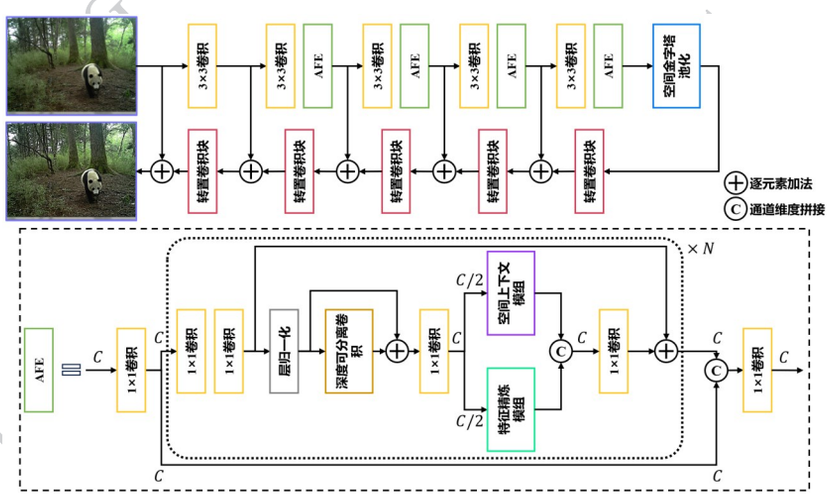


图4 图像增强器结构图

Fig. 4 Image enhancer structure

具体而言，图像增强器的结构如图4所示：网络整体采用U型编解码架构与特征精炼机制协同的设计范式。对于输入的图像 $\boldsymbol{x}_{img} \in \mathbb{R}^{H_0 \times W_0 \times 3}$，编码阶段首先使用3×3卷积进行下采样，生成初级特征图 $\boldsymbol{I}_1 \in \mathbb{R}^{H_1 \times W_1 \times C_1}$。随后，特征图依次经过4级特征提取阶段，每个阶段均通过3×3卷积进行下采样，并利用自适应特征增强模块（Adaptive Feature Enhancement，AFE）提取图像信息，该模块由空间上下文模组和特征精炼模组组成，分别用于捕获多尺度空间信息与强化目标边缘细节，从而提升对关键目标的

表征能力。最终得到最高级特征 $\boldsymbol{I}_5 \in \mathbb{R}^{H_5 \times W_5 \times C_5}$，式中 H_5，W_5 = H_1/16，W_1/16，特征维度 $C_5 = 16C_1$。在编码器的末端，加入空间金字塔池化模块，采用多尺度最大池化方法，进一步捕获全局上下文信息。解码阶段通过转置卷积模块实现跨尺度特征对齐，并采用跳跃连接以融合各层级语义信息。转置卷积模块包含四个基本操作：转置卷积上采样、1×1卷积通道压缩、批量归一化以及ReLU非线性激活。转置卷积模块的输出与相同尺寸的特征图逐元素相加，从而实现不同抽象层级特征的有效融合。在最终与输入图像逐元素相加之前，特征需要使用可学习的权重系数 λ_{AFE} 进行缩放，以在训练初期减弱其对图像的影响，防止对模型产生误导。

2.3.3　生成特征注入器

扩散生成模型在图像生成领域的卓越表现引发了学术界对其内部特征表征能力的广泛关注。研究表明，扩散模型的中间特征可以显著提升图像分类（Li 等，2023a；Xiang 等，2023；Yang 等，2023）、语义分割（Baranchuk 等，2022；Xu 等，2023）等任务的性能。其中间层特征通常具备较强的语义一致性，例如背景结构连续性与主体边缘完整性，这些特征能够帮助检测模型更有效地区分目标与背景，在遮挡、低照度及伪装场景下具有显著优势。

基于此，本文提出生成特征注入器，将微调后的生成模型内作为特征提取器，通过轻量化适配网络将多尺度语义表征注入至检测器骨干网络中。该设计不仅实现了生成与检测阶段的特征级协同融合，充分复用生成模型中的知识表征，同时在几乎不增加额外训练成本的情况下，提升了检测模型在复杂野外环境中的鲁棒性与跨域泛化能力。

具体而言，生成特征注入器的结构如图5所示，其整体设计参考了DreamTeacher（Li 等，2023b）的方案，包含两个阶段：特征提取阶段与特征处理阶段。在特征提取阶段，输入图像 $\boldsymbol{x}_{img} \in \mathbb{R}^{H_0 \times W_0 \times 3}$

首先经过标准的扩散过程。使用编码器 $\mathcal{E}$ 将图像嵌入至潜空间，得到低维表示 $\mathbf{z} \in \mathbb{R}^{32 \times 32 \times 4}$，经过 T 步噪声注入，生成含噪声的低维表示 $\mathbf{z}_T \in \mathbb{R}^{32 \times 32 \times 4}$。由于待检测图像的目标信息未知，因此使用全零向量 $\boldsymbol{c}_{zero}$ 作为条件输入，并从输出侧提取多层特征 $f_i \in \mathbb{R}^{(32/2^{i-1}) \times (32/2^{i-1}) \times 320i}$，$i \in \{1, 2, 3\}$。在特征处理阶段，最高层特征 f_3 被输入至一个来自PSPNet（pyramid scene parsing network）的金字塔池化模组，以在最小空间分辨率上进行上下文汇聚，从而增强特征的全局感受野与语义一致性。随后，该特征与 f_2 和 f_3 共同输入至特征金字塔网络（feature pyramid network，FPN），通过自顶向下路径与横向连接实现多尺度特征融合，以弥补低层特征语义不足的问题。融合后的特征表示为 f'_i，使用可学习的权重系数 λ_{PSPFPN} 进行缩放，以在训练初期抑制其对预训练检测模型权重的扰动。最终，所有输出特征经过双线性插值上采样至与目标检测模型骨干网络输出相同的空间尺度，并通过逐元素加法与检测模型特征进行融合，从而实现生成特征与检测特征的集成。

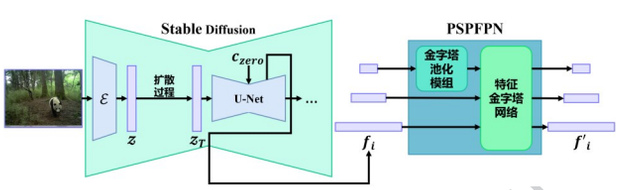


图5　生成特征注入器结构图

Fig. 5　Generative feature injector structure

3　实　验

3.1　实验设置

3.1.1　数据集

图6　LoTE-Animal数据集图像示例

Fig. 6　LoTE-Animal Dataset Image Examples

本研究采用LoTE-Animal数据集（long time-span dataset for endangered animal）（Liu 等，2023）作为主要的训练与测试数据来源。该数据集主要采集自卧龙国家级自然保护区，数据涵盖2009年至2021年间不同生态系统、季节、天气、时间、视角与栖息地场景，是一个面向濒危野生动物行为理解的大规模、长时序多模态数据集，典型图像如图6所示。基于

研究目标，我们从中筛选并构建了一个野外大熊猫图像子集，包含1,518张训练图像和380张测试图像。在图像生成实验中，训练集用于对生成模型进行微调，而测试集则用于评估生成结果与真实图像之间的分布差异。在目标检测实验中，采用相同的训练集对检测模型进行微调训练，并在测试集上对模型的检测性能进行系统评估。

图像生成使用的背景图像来自Wildlife Insights邛崃山脉项目数据集（McShea，2014）。该数据的采集区域与目标检测数据集高度重叠，图像分布特征相近。共获取了7,610张含复杂光照、遮挡的野外背景图像，经严格人工筛选，剔除了天空占比过大、植被覆盖率过低等不符合大熊猫栖息地样貌的图像，最终保留1,619张背景用于大熊猫图像合成。

3.1.2　评估指标

图像生成模型采用无明确目标参考图像质量评价指标进行评估（Rombach 等，2022；Yang 等，2023）。FID和KID为图像级别指标，用于量化生成图像与真实图像在分布上的相似程度，从而综合衡量生成图像的质量与多样性。而CLIP Score为目标级别指标，在本研究中用于评估可编辑区域内生成的合成目标图像与参考图像之间的语义一致性。

目标检测模型采用国际通用的综合评估指标mAP和mAR对模型性能进行分析，相关计算遵循MS COCO数据集（Lin 等，2014）所定义的标准化评估协议。mAP和mAR的计算采用IoU（intersection over union）阈值从0.5至0.95、步长为0.05的多阈值策略，先后对各阈值、各类别下的结果取平均得到最终结果。

3.1.3　基线模型

在图像生成实验中，我们选择了InstaGen（Feng 等，2024）作为对比基线。该方法是当前较为成熟且具有代表性的生成式数据增强方法，能够通过生成目标实例来扩充训练样本，具有较强的参考价值。

而鉴于现有野生动物与大熊猫检测相关研究未公开代码，本工作选取了具有代表性的实时检测器与基于Transformer的端到端检测器作为基线。我们选取了在该领域应用广泛的YOLOv5（Jocher，2020）作为对比。同时，为了与当前最先进的检测器进行全面对比，引入了YOLOv12（Tian 等，2025）与Co-DETR（Zong 等，2023）两种补充基线。

3.1.4　实现细节

图像生成模型的训练配置与默认设置保持一致，加载在OpenImages数据集（Kuznetsova 等，2020）上预训练的权重进行微调。训练过程中，基础学习率设为1.0e-05，并采用10,000步的预热策略。输入图像尺寸为256×256，扩散时间步数为1000，批处理大小为4，最大训练轮数为400。参考图像经过水平翻转、旋转、模糊与弹性变换等方式进行数据增强。在图像生成阶段，将图像裁剪并缩放为512×512后输入模型，掩码图像超出正方形区域的图像直接舍弃。最终，经随机筛选，获得1,518张合成的大熊猫图像。该生成图像数量参考文献（Fang 等，2024）的推荐，与微调生成模型所使用的图像数量一致。

目标检测模型采用Swin-Tiny（Liu 等，2021）作为图像骨干网络，加载在Objects365（Shao 等，2019）、GoldG（Kebe 等，2021）、GRIT（Gupta 等，2022）和V3Det（Wang 等，2023）数据集上预训练的权重。优化器选用AdamW（Loshchilov 和 Hutter，2017），基础学习率设为1.0e-5，文本与图像骨干的初始学习率为1.0e-6，图像增强器为5.0e-5，PSPFPN模块为1.0e-4。加入Multistep学习率衰减策略，衰减系数为0.1。批处理大小为4，最大训练轮数为6。训练过程中使用水平翻转、缩放与裁剪等数据增强方式。自适应特征增强模块的核心组件配置为[3,6]。考虑到大熊猫检测任务对细粒度特征的需求，将生成特征注入的噪声步数T设置为150（低于DreamTeacher（Li 等，2023b）推荐的250）；并从U-Net输出侧的12个特征层中抽取第5、8和11层进行注入。模型基于MMDetection工具箱（Chen 等，2019b）构建，并以MM Grounding DINO框架（Zhao 等，2024）为基础实现。

在生成与检测模型中，可学习权重系数（λ_{label}、λ_{AFE}和λ_{PSPFPN}）初始化为0.01，全部训练与测试实验均在NVIDIA GeForce RTX 4090 GPU上完成。

3.2　大熊猫图像生成实验结果

为了验证类别引导机制在图像生成任务中的实际效果，我们对引入该机制前后模型生成结果进行了对比。需要说明的是，类别引导机制仅作用于训练阶段，在推理阶段的输入依然只包含背景图像、参考图像与掩码图像三元组，不额外输入类别信息。

可视化结果如图7所示。加入类别引导机制

(a)背景图像 (b)Paint by Example (c)加入类别引导机制
((a)background image;(b)paint by example;(c)implement category-guidance mechanism)

图7 类别引导机制模型生成结果图

Fig. 7 Category-guidance mechanism model's generation results

后,生成图像质量得到显著提升,合成大熊猫目标在光照强度与色温方面与背景高度一致,且毛发纹理与周围植被之间过渡自然,展现出更强的融合能力。

定量评价结果如表1所示,InstaGen在图像级评价指标上均明显弱于本文方法,其在图像整体质量和分布一致性方面与真实数据存在更大差距。引入类别引导机制后,Paint by Example生成结果的各项指标均得到提升。在图像级指标方面,FID从147.00降至123.13,KID从0.059降至0.038,表明类别引导下合成的大熊猫图像在整体特征分布上更接近真实图像;同时,CLIP Score由0.625提升至0.636,说明类别引导机制有效增强了模型的语义理解与生成控制能力,提高了生成图像的语义一致性。

表1 图像生成结果定量评价表

Table 1 Quantitative evaluation form for image generation results

模型设置	图像级			目标级
	FID(↓)	KID(↓)	KID std(↓)	CLIP Score(↑)
InstaGen	164.05	0.098	**0.0021**	-
Paint by Example	147.00	0.059	0.0023	0.625
加入类别引导机制	**123.13**	**0.038**	**0.0021**	**0.636**

注:↑和↓分别表示数值越高和越低越好,加粗字体为每列最优值。

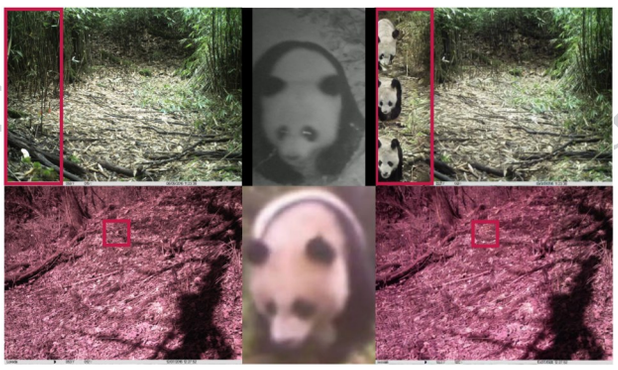
类别引导机制虽然没有在推理阶段注入类别语义信息,但通过在训练阶段强化模型的语义约束,不仅保证了生成图像的感知质量,还提升了其在高层次特征分布和语义一致性方面与真实图像的匹配程度。

类别引导机制提升了生成图像的分布一致性与视觉质量,但受限于模型的边界框选取策略,个别结果仍具有较强的视觉违和感或并未成功生成,如图8所示。具体而言,第一行中掩码区域比例设置不合理,影响了生成目标与背景的融合效果;第二行中目标尺度与背景比例失衡,从而导致目标生成失败。未来可进一步探索如何自适应地确定合成目标在图像中的位置与尺度。

3.3 大熊猫目标检测实验结果

为评估图像增强器与生成特征注入器对目标检测性能的提升效果,本文对引入不同组件前后的模型进行了系统性对比分析,并探讨了图像增强器的潜在增强原理。

表2展示了目标检测定量结果。目标边界框的像素面积在32×32至96×96之间为中型目标,中型目标仅占全部数据的0.68%。基线结果表明,YOLO系列模型(Jocher,2020;Tian等,2025)虽然参数更少,但在整体性能上明显低于Co-DETR(Zong等,2023)和PandaGenDet。Co-DETR和PandaGenDet模型的计算量接近,前者的全部目标mAP略高于后者的原版模型,但低于改进后的模型。

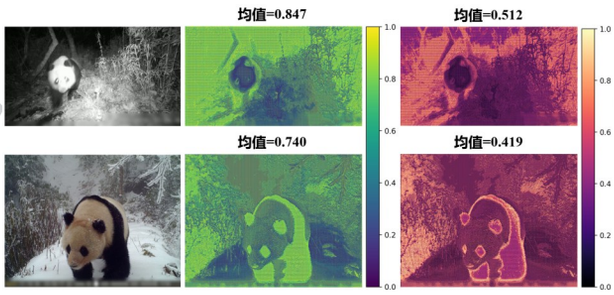


(a)背景图像 (b)参考图像 (c)加入类别引导机制
((a)background image;(b)reference image;(c)implement category-guidance mechanism)

图8　类别引导机制模型生成失败示例

Fig. 8　Example of category-guidance mechanism model's generation failure

对比 PandaGenDet 的不同版本，加入图像增强器后，模型参数量仅增加了 3.4M，而检测性能得到明显改善(mAP：88.8→89.7，mAR：94.9→95.5)，尤其在中型目标上提升最为突出(mAP：66.2→72.6，mAR：80.0→82.0)。表明轻量化的图像增强器在强化细节特征和缓解背景干扰方面发挥了积极作用，使模型在复杂背景下的多尺度目标检测性能显著提升。虽然单独引入生成特征注入器对整体性能的提升并不显著，但生成特征注入器确实增强了模型对中型目标的判别能力和定位稳定性。这说明，生成模型提取的多尺度特征为检测器提供了更强的上下文建模能力和边界信息。最终，在图像增强器与生成特征注入器并行使用时，虽然 mAR 略有下降，但模型整体性能达到最优，mAP 达到了 89.8。这表明两者的结合能够实现特征层面的互补：图像增强器优化了图像输入特征的可分辨性，而生成特征注入器提供了来自生成模型的语义先验，从而共同提升了检测模型的整体鲁棒性与泛化能力。实验结果充分验证了该设计在野外场景中的有效性，也表明其在其他视觉任务中可具有良好的迁移潜力。



(a)原图像 (b)绝对差值图 (c)感知色差图
((a)original image;(b)absolute difference map;(c)perceptual difference map)

图9　图像增强器差异图可视化结果图

Fig. 9　Visualization of Image Enhancer Difference Map Results

为深入分析图像增强器在特征优化过程中的具体作用，采用绝对差值图和感知色差图对像素级编辑效果进行量化与解释，结果如图9所示。绝对差值图结果表明，图像增强器对图像的修改集中出现在原图亮度较低和细节模糊的区域，而亮度较高或纹理平滑区域则响应较弱。感知色差图进一步表明，增强器在许多高响应区域不仅改变了像素幅度，而且引入了可被人眼感知的色彩/亮度偏移。这表明增强操作并非单纯的噪声级微调，而是在色彩与亮度层面对目标局部特征进行了有意义的增强，甚

表2　目标检测结果定量评价表

Table 2　Quantitative evaluation form for object detection results

模型	图像增强器	生成特征注入器	参数量	GFLOPs	全部目标		中型目标	
					mAP	mAR	mAP	mAR
YOLOv5l	×	×	12.4M	8.2	86.7	90.7	**78.0**	78.0
YOLOv12l	×	×	26.3M	88.6	87.9	90.1	60.3	60.0
Co-DETR	×	×	64.5M	37.2	89.3	92.4	70.1	72.0
Panda GenDet	×	×	172.9M	38.6	88.8	94.9	66.2	80.0
	×	√	206.4M	38.6	88.8	93.7	68.1	78.0
	√	×	176.3M	38.6	89.7	**95.5**	72.6	**82.0**
	√	√	209.8M	38.6	**89.8**	95.1	73.2	76.0

注：加粗字体为每列最优值。

至部分图像出现了如右下角所示的边缘增强效果。综上所述，图像增强器呈现出基于图像内容的自适应编辑特性：既能针对关键边缘、纹理结构进行增强，又能在不影响大面积背景的情况下减少冗余修改。

3.4 联合生成图像数据集继续训练实验结果

本节通过加入生成模型合成的大熊猫图像，构建合成图像数据集，并与真实图像数据集联合用于目标检测模型的继续训练，以评估合成数据对检测性能的提升效果。在实验中，生成模型与检测模型均采用前述的最优版本，以未进行继续训练的模型作为对比基线，在其基础上额外训练2个轮次。“无”表示未继续训练的模型；“LoTE”表示仅使用LoTE-Animal数据集继续训练；“LoTE+Syn”表示联合使用真实图像数据集与合成图像数据集进行继续训练。

表3 生成式数据增强定量评价表

Table 3 Quantitative evaluation form for generative data augmentation results

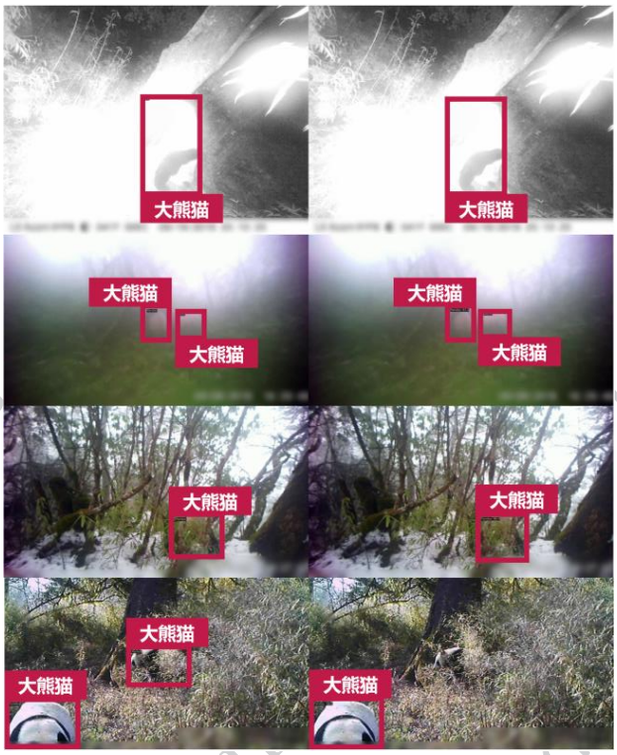
继续训练数据集	全部目标		中型目标	
	mAP	mAR	mAP	mAR
无	89.8	**95.1**	73.2	76.0
LoTE	89.7	94.9	**76.3**	**84.0**
LoTE+Syn	**90.1**	94.2	69.8	78.0

注：加粗字体为每列最优值。

测试结果如表3所示，仅使用LoTE-Animal数据集继续训练会降低模型性能，mAP从89.8下降至89.7；而在继续训练中加入合成图像数据后，模型整体性能进一步提升，mAP达到90.1，充分证明了合成数据在缓解数据稀缺与提升检测性能方面的有效性。综合来看，联合真实与合成图像继续训练的模型取得了本研究中最优的检测效果。

最佳模型的更多可视化结果如图10所示，模型在多种复杂场景下均能准确识别大熊猫目标，即使是人类难以辨认的图像（第1行至第3行），模型依然能够给出正确的检测结果，充分体现了其良好的鲁棒性与泛化能力。仅在部分极少数困难样本中（如第四行仅露出头部的大熊猫），模型出现了漏检现象。总体而言，该模型在真实野外场景中展现出了稳定而强大的目标识别能力。

为进一步验证生成式数据增强对检测性能的提升效果，我们在不同初始数据规模下开展了对比实验。首先，在数据生成阶段，使用不同数量的真实图像训练生成模型，生成并筛选出1,518张大熊猫合成图像。随后，在目标检测阶段，以该生成模型作为特征提取器，在相同的真实数据上训练检测模型，同时使用合成图像与真实图像联合继续训练。



(a)真值 (b)最佳检测结果

((a)ground truth;(b)best detection result)

图10 本研究最佳目标检测模型检测结果

Fig. 10 Detection results of the best object detection model

实验结果如图11所示，在生成相同数量的合成图像时，真实数据量越小，合成数据带来的提升越显

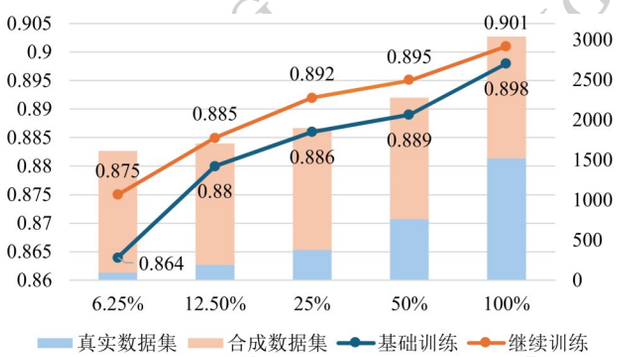


图11 不同数据量条件下检测模型性能图

Fig. 11 Performance of detection models under different dataset volume

著。当仅使用6.25%的真实数据(即94张图像)进行训练时，加入合成图后mAP从86.4提升至87.5，提升幅度最大。而随着真实数据量的增加，合成图对模型性能的边际贡献逐步减小。比较相邻数据量的结果可知，使用1,518张合成图像与真实图联合继续训练所得的模型，其性能仍略低于只使用2倍真实图像训练得到的模型。这可能是因为合成图像在多样性或真实感上仍存在不足，以及合成图像的伪标签质量不及人工标注等原因。因此，尽管合成图像的视觉质量随着训练数据规模增长而有所改善，但其仍难以完全替代等量的真实图像。

3.5 开放集检测实验结果

为评估模型在新类别上的泛化能力，我们对模型的开放集检测性能进行了分析。值得注意的是，本文以大熊猫检测为核心，开放集检测旨在评估模型在遇到训练集中未出现物种时的稳健性与语义泛化能力，检验其在真实生态监测场景中的适用性。

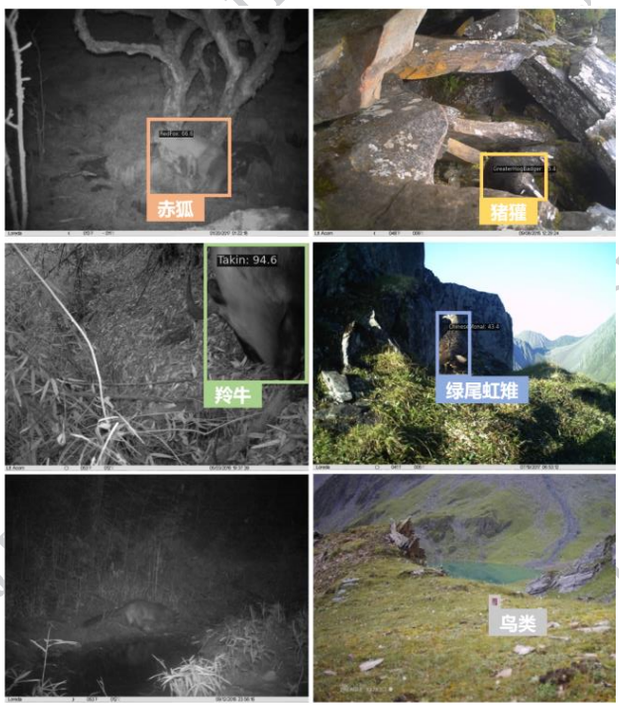


图12 本研究最佳目标检测模型开放集检测结果

Fig. 12 Open-set detection results of the best object detection model

在包含全部物种的LoTE-Animal数据集上进行定量测试，虽然模型仅使用了大熊猫类别数据进行训练，但是模型在藏酋猴和岩羊类别上的mAP依然达到了25.8和33.7，说明该方法具备跨类别检测能力，并能够对未见类别产生有效响应。使用Wildlife Insight邛崃山脉项目数据集(McShea，2014)的其他野生动物图像进行定性评估。结果如图12所示，除果子狸未被正确识别、羚牛检测框未完全覆盖角部外，其余检测结果均准确无误。该结果表明，所提出的模型能够有效识别训练阶段未见的物种，包括部分鸟类与哺乳动物个体，展现出良好的开放集检测能力与语义泛化能力。然而，目前尚缺乏统一的野生动物开放集检测评估基准，未来可进一步推动开放集野生动物检测基准的标准化与系统化建设，以实现对开放集性能的全面、客观与可复现评估。

4 结 论

为了缓解大熊猫检测任务中普遍存在的数据稀缺问题以及预训练数据与野外图像之间的域差异，本文提出了一个生成-检测一体化方法——PandaGenDet。在图像生成部分，我们设计了一种类别引导机制，使用文本特征提供额外的强先验信息引导生成过程，从而合成高质量的训练数据。在目标检测部分，我们提出了一个即插即用的图像增强器模块，利用多尺度上下文建模与细节增强的双路径策略增强图像，使野外图像更加适应检测器预训练权重。进一步地，我们在检测模型中设计了生成特征注入器，将生成模型的多层语义表征注入检测器骨干网络，实现了生成知识向检测任务的深层转化。最终实现了数据级合成、图像级增强与特征级注入的三重协同优化。实验验证表明，PandaGenDet在合成图像质量和检测鲁棒性方面均取得了明显改进，具有良好的开放集检测能力。然而，当前生成模型在部分极端条件下仍存在合成失败的问题，合成样本对检测器的正向贡献依然受限。未来工作可以聚焦于降低合成失败率、优化可编辑区域设置策略，增强合成样本对检测模型的有效增益，以及构建更加系统化的野外开放集检测评估体系方向。

参考文献(References)

Ali M, Javaid M, Noman M, Fiaz M and Khan S. 2024. Fanet: feature amplification network for semantic segmentation in cluttered background//Proceedings of the IEEE International Conference on Image Processing. Abu Dhabi: IEEE: 2592-2598 [DOI: 10.1109/

ICIP51287.2024.10647349]

Azizi S, Kornblith S, Saharia C, Norouzi M and Fleet D J. 2023. Synthetic data from diffusion models improves imagenet classification [EB/OL]. [2023-04-17]. https://arxiv.org/pdf/2304.08466.pdf

Baranchuk D, Rubachev I, Voynov A, Khrulkov V and Babenko A. 2021. Label-efficient semantic segmentation with diffusion models [EB/OL]. [2021-12-06]. https://arxiv.org/pdf/2112.03126.pdf

Borji A. 2022. Generated faces in the wild: quantitative comparison of stable diffusion, midjourney and dall-e 2 [EB/OL]. [2022-10-02]. https://arxiv.org/pdf/2210.00586.pdf

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with transformers//Proceedings of the 16th European Conference on Computer Vision (ECCV 2020). Glasgow: Springer: 213-229 [DOI: 10.1007/978-3-030-58452-8_13]

Chen K, Wang J, Pang J, Cao Y, Xiong Y, Li X, Sun S, Feng W, Liu Z, Xu J, Zhang Z, Cheng D, Zhu C, Cheng T, Zhao Q, Li B, Lu X, Zhu R, Wu Y, Dai J, Wang J, Shi J, Ouyang W, Loy C C and Lin D. 2019b. Mmdetection: open mmlab detection toolbox and benchmark [EB/OL]. [2019-06-17]. https://arxiv.org/pdf/1906.07155.pdf

Devlin J, Chang M W, Lee K and Toutanova K. 2019. Bert: pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis: Association for Computational Linguistics: 4171-4186 [DOI: 10.18653/v1/n19-1423]

DeVries T and Taylor G W. 2017. Improved regularization of convolutional neural networks with cutout [EB/OL]. [2017-08-15]. https://arxiv.org/pdf/1708.04552.pdf

Everingham M, Eslami S M A, Van Gool L, Williams C K I, Winn J and Zisserman A. 2015. The pascal visual object classes challenge: a retrospective. International Journal of Computer Vision, 111(1): 98-136 [DOI: 10.1007/s11263-014-0733-5]

Fang H Y, Han B, Zhang S, Zhou S, Hu C X and Ye W M. 2024. Data augmentation for object detection via controllable diffusion models//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2024). Waikoloa: IEEE: 1246-1255 [DOI: 10.1109/WACV57701.2024.00129]

Fang J Z, Yang H Y, Chen P, Wang C Y and Hu S X. 2020. A detection algorithm of giant panda in wild video image based on wavelet-SSD network//Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics. Toronto: IEEE: 3655-3660 [DOI: 10.1109/SMC42975.2020.9283247]

Feng C J, Zhong Y J, Jie Z Q, Xie W D and Ma L. 2024. Instagen: enhancing object detection by training on synthetic dataset//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024). Seattle: IEEE: 14121-14130 [DOI: 10.1109/CVPR52733.2024.01339]

Feng J F and Li J C. 2022. An adaptive embedding network with spatial constraints for the use of few-shot learning in endangered-animal detection. ISPRS International Journal of Geo-Information, 11(4): 256 [DOI: 10.3390/ijgi11040256]

Ganguly R, Bah M D and Dahmane M. 2024. Diffusion models as a representation learner for deepfake image detection//Proceedings of the 27th International Conference on Pattern Recognition (ICPR 2024). Kolkata: Springer: 228-241 [DOI: 10.1007/978-3-031-78305-0_15]

Ge Y, Xu J, Zhao B N, Joshi N, Itti L and Vineet V. 2022. Dall-e for detection: language-driven compositional image synthesis for object detection [EB/OL]. [2022-06-20]. https://arxiv.org/pdf/2206.09592.pdf

Girshick R B, Donahue J, Darrell T and Malik J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014). Columbus: IEEE Computer Society: 580-587 [DOI: 10.1109/CVPR.2014.81]

Girshick R B, Donahue J, Darrell T and Malik J. 2015. Region-based convolutional networks for accurate object detection and segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(1): 142-158 [DOI: 10.1109/TPAMI.2015.2437384]

Gong S, Deng Y and Xiang J H. 2025. A review of image generation methods based on diffusion models. Journal of Wuhan University (Engineering Edition), 58(2): 292-305 (龚帅, 邓勇, 向金海. 2025. 基于扩散模型的图像生成方法研究综述. 武汉大学学报(工学版), 58(2): 292-305) [DOI: 10.14188/j.1671-8844.2024.0148]

Guo J C, Yue H H, Zhang Y, Liu D, Liu X W and Zheng S D. 2022. The analysis of image enhancement on salient object detection. Journal of Image and Graphics, 27(7): 2129-2147 (郭继昌, 岳惠惠, 张怡, 刘迪, 刘晓雯, 郑司达. 2022. 图像增强对显著性目标检测的影响研究. 中国图象图形学报, 27(07): 2129-2147) [DOI: 10.11834/jig.200735]

Gupta T, Marten R, Kembhavi A and Hoiem D. 2022. Grit: general robust image task benchmark [EB/OL]. [2022-04-28]. https://arxiv.org/pdf/2204.13653.pdf

He R F, Sun S Y, Yu X, Xue C H, Zhang W Q, Torr P H S, Bai S and Qi X J. 2022. Is synthetic data from generative models ready for image recognition? [EB/OL]. [2022-10-14]. https://arxiv.org/pdf//arXiv.2210.07574.pdf

Jiao L C, Gao X B, Han J W, Li Y S, Bai X, Yang S Y, Meng D Y, Ren W Q, Shi Z H and Chen X Y. 2023. Introduction to the special issue on intelligent detection of target in complex scene images. Journal of Image and Graphics, 28(9): 2561-2562 (焦李成, 高新波, 韩军伟, 李云松, 白翔, 杨淑媛, 孟德宇, 任文琦, 石争浩, 陈秀妍. 2023.《中国图象图形学报》复杂场景图像目标智能

检测专栏简介. 中国图象图形学报，28(09)：2561-2562)[DOI：10.11834/jig.2300009]

Jocher G. 2020. Ultralytics yolov5 [CP/OL]. [2020-06-25]. https://github.com/ultralytics/yolov5

Kebe G Y, Higgins P, Jenkins P, Darvish K, Sachdeva R, Barron R, Winder J, Engel D, Raff E, Ferraro F and Matuszek C. 2021. A spoken language dataset of descriptions for speech-based grounded language learning//Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. Virtual：NeurIPS：

Kuznetsova A, Rom H, Alldrin N, Uijlings J R R, Krasin I, Pont-Tuset J, Kamali S, Popov S, Malloci M, Duerig T and Ferrari V. 2018. The open images dataset v4：unified image classification, object detection, and visual relationship detection at scale [EB/OL]. [2018-11-02]. http://arxiv.org/abs/1811.00982.pdf

Li A C, Prabhudesai M, Duggal S, Brown E and Pathak D. 2023a. Your diffusion model is secretly a zero-shot classifier//Proceedings of the IEEE/CVF International Conference on Computer and Communications. Paris：IEEE：2206-2217 [DOI：10.1109/ICCV51070.2023.00210]

Li D, Ling H, Kar A, Acuna D, Kim S W, Kreis K, Torralba A and Fidler S. 2023b. Dreamteacher：pretraining image backbones with deep generative models//Proceedings of the IEEE/CVF International Conference on Computer and Vision. Paris：IEEE：16652-16662 [DOI：10.1109/ICCV51070.2023.01531]

Li Z, Li Y X, Zhao P H, Song R J, Li X and Yang J. 2023c. Is synthetic data from diffusion models ready for knowledge distillation? [EB/OL]. [2023-05-22]. https://arxiv.org/pdf/2305.12954.pdf

Lin S B, Wang K, Zeng X Y and Zhao R. 2023. Explore the power of synthetic data on few-shot object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR 2023). Vancouver：IEEE：638-647 [DOI：10.1109/CVPRW59228.2023.00071]

Lin T Y, Maire M, Belongie S J, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft coco：common objects in context//Proceedings of the 13th European Conference on Computer Vision (ECCV 2014). Zurich：Springer：740-755 [DOI：10.1007/978-3-319-10602-1_48]

Liu A A, Su Y T, Wang L J, Li B, Qian Z X, Zhang W M, Zhou L N, Zhang X P, Zhang Y D and Huang J W. 2024a. Aigc visual content generation and provenance research progress. Journal of Image and Graphics, 29(6)：1535-1554 (刘安安，苏育挺，王岚君，李斌，钱振兴，张卫明，周琳娜，张新鹏，张勇东，黄继武，俞能海. 2024a. AIGC视觉内容生成与溯源研究进展. 中国图象图形学报，29(6)：1535-1554)[DOI：10.11834/jig.240003]

Liu D, Hou J, Huang S, Liu J, He Y, Zheng B, Ning J and Zhang J. 2023. Lote-animal：a long time-span dataset for endangered animal behavior understanding//Proceedings of the IEEE/CVF International Conference on Computer and Vision. Paris：IEEE：20007-20018 [DOI：10.1109/ICCV51070.2023.01836]

Liu L L, Muelly M, Deng J, Pfister T and Li L J. 2019. Generative modeling for small-data object detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul：IEEE：6073-6081 [DOI：10.1109/ICCV.2019.00617]

Liu S L, Zeng Z Y, Ren T H, Li F, Zhang H, Yang J, Jiang Q, Li C Y, Yang J W, Su H, Zhu J and Zhang L. 2024b. Grounding dino：marrying dino with grounded pre-training for open-set object detection//Proceedings of the 18th European Conference on Computer Vision (ECCV 2024). Milan：Springer：38-55 [DOI：10.1007/978-3-031-72970-6_3]

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S E, Fu C Y and Berg A C. 2016. Ssd：single shot multibox detector//Proceedings of the 14th European Conference on Computer Vision (ECCV 2016). Amsterdam：Springer：21-37 [DOI：10.1007/978-3-319-46448-0_2]

Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S and Guo B. 2021. Swin transformer：hierarchical vision transformer using shifted windows//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal：IEEE：9992-10002 [DOI：10.1109/ICCV48922.2021.00986]

Loshchilov I and Hutter F. 2019. Decoupled weight decay regularization [EB/OL]. [2019-11-14]. https://arxiv.org/pdf/1711.05101.pdf

Luo H L and Chen H K. 2020. Survey of object detection based on deep learning. Acta Electronica Sinica, 48(6)：1230-1239 (罗会兰，陈鸿坤. 2020. 基于深度学习的目标检测研究综述. 电子学报，48(6)：1230-1239)[DOI：10.3969/j.issn.0372-2112.2020.06.026]

Lyu H T and Jia X L. 2024. A lightweight giant panda object detection model with attention mechanism. Laser Journal, 45(8)：61-68 (吕皓天，贾小林. 2024. 融合注意力机制的轻量化大熊猫目标检测模型. 激光杂志，45(08)：61-68)[DOI：10.14016/j.cnki.jgzz.2024.08.061]

McShea W. 2014. Qionglai mountains project [EB/OL]. http://n2t.net/ark:/63614/w12004579

Meng J S, Ma Y M, Yang Z, Sun Q, Ju Y J, Xie J J and Zhang J G. 2025. Wild animal detection method based on pseudo-labels and YOLOv4. Chinese Journal of Wildlife, 46(03)：523-532 (孟继森，马玉明，杨紫合，孙茜，巨友娟，谢将剑，张军国. 2025. 基于伪标签和YOLOv4的野生动物检测方法. 野生动物学报，46(03)：523-532)[DOI：10.12375/ysdwxb.20250306]

Nguyen H, Maclagan S J, Nguyen T D, Nguyen T, Flemons P, Andrews K, Ritchie E G and Phung D Q. 2017. Animal recognition and identification with deep convolutional neural networks for automated wildlife monitoring//Proceedings of the 2017 IEEE International Conference on Data Science and Advanced Analytics. Tokyo：IEEE：40-49 [DOI：10.1109/DSAA.2017.31]

Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M

and Sutskever I. 2021. Zero-shot text-to-image generation [EB/OL]. [2021-02-24]. https://arxiv.org/pdf/2102.12092.pdf

Redmon J, Divvala S K, Girshick R B and Farhadi A. 2016. You only look once: unified, real-time object detection//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). Las Vegas: IEEE Computer Society: 779-788 [DOI: 10.1109/CVPR.2016.91]

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022). New Orleans: IEEE: 10674-10685 [DOI: 10.1109/CVPR52688.2022.01042]

Roy A M, Bhaduri J, Kumar T and Raj K. 2023. Wildect-yolo: an efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection. Ecological Informatics, 75: 101919 [DOI: 10.1016/j.ecoinf.2022.101919]

Saharia C, Chan W, Saxena S, Li L, Whang J, Denton E, Seyed Ghasemipour S K, Karagol Ayan B, Mahdavi S S, Gontijo Lopes R, Salimans T, Ho J, Fleet D J and Norouzi M. 2022. Photorealistic text-to-image diffusion models with deep language understanding [EB/OL]. [2022-05-23]. https://arxiv.org/pdf/2205.11487.pdf

Sariyildiz M B, Alahari K, Larlus D and Kalantidis Y. 2023. Fake it till you make it: learning transferable representations from synthetic imagenet clones//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE: 8011-8021 [DOI: 10.1109/CVPR52729.2023.00774]

Shipard J, Wiliem A, Nguyen Thanh K, Xiang W and Fookes C. 2023. Diversity is definitely needed: improving model-agnostic zero-shot classification via stable diffusion//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2023 Workshops. Vancouver: IEEE: 769-778 [DOI: 10.1109/CVPRW59228.2023.00084]

Su X H, Zhang J W, Ma Z B, Dong Y Q, Zi J L, Xu N, Zhang H Y, Xu F and Chen F X. 2024. Identification of rare wildlife in the field environment based on the improved YOLOv5 model. Remote Sensing, 16(9): 1535 [DOI: 10.3390/rs16091535]

Tian Y, Ye Q and Doermann D S. 2025. Yolov12: attention-centric real-time object detectors [EB/OL]. [2025-02-18]. https://doi.org/10.48550/arXiv.2502.12524.pdf

Wang H L, Zhong J S, Xu Y F, Luo G, Jiang B Y, Hu Q, Lin Y C and Ran J H. 2022. Automatically detecting the wild giant panda using deep learning with context and species distribution model. Ecological Informatics, 72: 101868 [DOI: 10.1016/j. ecoinf. 2022. 101868]

Wang J, Zhang P, Chu T, Cao Y, Zhou Y, Wu T, Wang B, He C and Lin D. 2023. V3det: vast vocabulary visual detection dataset//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris: IEEE: 19787-19797 [DOI: 10.1109/ICCV51070.2023.01817]

Xiang W, Yang H, Huang D and Wang Y. 2023. Denoising diffusion autoencoders are unified self-supervised learners//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris: IEEE: 15756-15766 [DOI: 10.1109/ICCV51070.2023.01448]

Xu D G, Wang L and Li F. 2021. A review of typical object detection algorithms based on deep learning. Computer Engineering and Applications, 57(8): 10-25 (许德刚, 王露, 李凡. 2021. 深度学习的典型目标检测算法研究综述. 计算机工程与应用, 57(8): 10-25) [DOI: 10.3778/j.issn.1002-8331.2012-0449]

Xu J, Liu S, Vahdat A, Byeon W, Wang X and De Mello S. 2023. Open-vocabulary panoptic segmentation with text-to-image diffusion models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023). Vancouver: IEEE: 2955-2966 [DOI: 10.1109/CVPR52729.2023.00289]

Yang B X, Gu S Y, Zhang B, Zhang T, Chen X J, Sun X Y, Chen D and Wen F. 2023. Paint by example: exemplar-based image editing with diffusion models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023). Vancouver: IEEE: 18381-18391 [DOI: 10.1109/CVPR52729.2023.01763]

Ye G S, Wang J M, Yang Z Z, Zhang Y H, Cui R K and Xuan S. 2023. Survey of image composition based on deep learning. Journal of Image and Graphics, 28(12): 3670-3698 (叶国升, 王建明, 杨自忠, 张宇航, 崔荣凯, 宣帅. 2023. 深度学习图像合成研究综述. 中国图象图形学报, 28(12): 3670-3698) [DOI: 10.11834/jig.220713]

Yun S D, Han D Y, Chun S H, Oh S J, Yoo Y J and Choe J S. 2019. Cutmix: regularization strategy to train strong classifiers with localizable features//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE: 6022-6031 [DOI: 10.1109/ICCV.2019.00612]

Zhang H, Cisse M, Dauphin Y N and Lopez-Paz D. 2017. Mixup: beyond empirical risk minimization [EB/OL]. [2017-10-25]. https://arxiv.org/ pdf/1710.09412.pdf

Zhang H, Li F, Liu S L, Zhang L, Su H, Zhu J, Ni L M and Shum H Y. 2022. Dino: detr with improved denoising anchor boxes for end-to-end object detection [EB/OL]. [2022-03-07]. https://arxiv.org/pdf/2203.03605.pdf

Zhang K, Sheng X, Xiao Y J, Yang J Y, Chen M J and Ren Z H. 2025a. Stable diffusion model for few-shot generation of meter defect images. Journal of Image and Graphics. Journal of Image and Graphics, 30(11): 3451-3464 (张珂, 盛鑫, 肖扬杰, 杨资远, 陈美娟, 任泽华. 2025a. 面向少样本表计缺陷图像生成的稳定扩散模型. 中国图象图形学报, 30(11): 3451-3464) [DOI: 10.11834/jig.240777]

Zhang Y F, Liu J Y, Ma H M, Liu S X, Jia W, Liu W and Han X D.

2025b. Introduction to the data generation and application special column for computer vision tasks. Journal of Image and Graphics, 30(11): 3411-3412 (张永飞, 刘家瑛, 马惠敏, 刘世霞, 贾伟, 刘武, 韩向娣. 2025b.《中国图象图形学报》面向计算机视觉任务的数据生成与应用专栏简介. 中国图象图形学报, 30(11): 3411-3412) [DOI: 10.11834/jig.2500011]

Zhao X, Chen Y, Xu S, Li X, Wang X, Li Y and Huang H. 2024. An open and comprehensive pipeline for unified object grounding and detection [EB/OL]. [2024-01-04]. https://arxiv.org/pdf/2401.02361.pdf

Zhao Y Q, Rao Y, Dong S P and Zhang J Y. 2020. Survey on deep learning object detection. Journal of Image and Graphics, 25(4): 629-654 (赵永强, 饶元, 董世鹏, 张君毅. 2020. 深度学习目标检测方法综述. 中国图象图形学报, 25(4): 629-654) [DOI: 10.11834/jig.190307]

Zhao Z X, Liu Y, Sun X D, Liu J T, Yang X T and Zhou C. 2021. Composited fishnet: fish detection and species recognition from low-quality underwater videos. IEEE Transactions on Image Processing, 30: 4719-4734 [DOI: 10.1109/TIP.2021.3074738]

Zheng T P, Chen Y X, Wen X Z, Li Y C and Wang Z Y. 2025. Diffusion model-generated video dataset and detection benchmarks. Journal of Image and Graphics, 30(04): 1059-1071 (郑天鹏, 陈雁翔, 温心哲, 李严成 and 王志远. 2025. 扩散模型生成视频数据集及其检测基准研究. 中国图象图形学报, 30(04): 1059-1071) [DOI: 10.11834/jig.240259]

Zhong Z, Zheng L, Kang G, Li S and Yang Y. 2020. Random erasing data augmentation//Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020). New York: AAAI Press: 13001-13008 [DOI: 10.1609/AAAI.V34I07.7000]

Zhou Y C, Sahak H and Ba J. 2023. Training on thin air: improve image classification with generated data [EB/OL]. [2023-05-24]. https://doi.org/10.48550/arXiv.2305.15316

Zhu J Y, Li S Y, Liu Y X A, Yuan J, Huang P and Shan J. 2024. Odgen: domain-specific object detection data generation with diffusion models. Advances in Neural Information Processing Systems, 37: 63599-63633 [DOI: 10.52202/079017-2031].

Zong Z, Song G L and Liu Y. 2023. Detrs with collaborative hybrid assignments training//Proceedings of the IEEE/CVF International Conference on Computer and Vision. Paris: IEEE: 6725-6735 [DOI: 10.1109/ICCV51070.2023.00621]

作者简介

刘祯，男，硕士研究生，主要研究方向为计算机视觉和目标检测。E-mail: liu_zhen@buaa.edu.cn

史振威，通讯作者，男，教授，主要研究方向为图像处理、模式识别和机器学习。E-mail: shizhenwei@buaa.edu.cn

杨沁哲，男，博士研究生，主要研究方向为遥感图像解译模型。E-mail: yangqinzhe@buaa.edu.cn

刘丽芹，女，博士后，主要研究方向为高光谱遥感数据处理和深度学习。Email: liuliqin@buaa.edu.cn

刘辰阳，男，博士研究生，主要研究方向为遥感图像多模态解译。Email: liuchenyang@buaa.edu.cn

邹征夏，男，教授，主要研究方向为遥感图像处理、人工智能、计算机视觉和深度学习。E-mail: zhengxiazou@buaa.edu.cn