重庆理工大学学报 自然科学
Journal of Chongqing University of Technology(Natural Science)

面向3D目标检测的多模态生成式图像数据增强的研究

张光钱, 周广利, 黄飞, 刘文兵, 向阳开

引用本文:
张光钱, 周广利, 黄飞, 刘文兵, 向阳开. 面向3D目标检测的多模态生成式图像数据增强的研究[J]. 重庆理工大学学报（自然科学）, 2024, 38(10): 13–20.

重庆理工大学学报(自然科学)
Journal of Chongqing University of Technology(Natural Science)
2024 年 第 38 卷 第 10 期
Vol. 38 No. 10 2024
doi: 10.3969/j. issn. 1674-8425(z). 2024. 10. 002

面向 3D 目标检测的多模态生成式图像数据增强的研究

张光钱[1],周广利[2],黄 飞[2],刘文兵[2],向阳开[1]

(1. 重庆交通大学 机电与车辆工程学院, 重庆 400074;
2. 中国路桥工程有限责任公司, 北京 100010)

摘 要:针对传统生成式图像数据增强算法丢失 3D 属性信息,无法应用于自动驾驶领域 3D 目标检测任务的问题,提出了一种基于稳定扩散模型的多模态图像生成算法,并基于该算法设计了一种面向 3D 目标检测的数据增强方法。算法通过增加多模态输入进一步约束图像的生成过程。算法设计了一种多模态特征在线生成模块,在线提取场景描述、语义分布和深度特征等信息;同时针对多模态特征融合网络设计了一种增强型门控自注意力模块,精准地捕捉潜在特征空间中的深度信息,从而保留图像的 3D 属性信息,实现对图像纹理、颜色以及光照等 2D 特征的针对性修改。基于算法出色的深度保持特性,将新图像与 3D 伪标签结合,构成新的图像样本,实现对图像样本的数据增强。在 nuScenes 公开数据集上 3D 检测结果表明,算法针对公交车、卡车等体积较大类别的 3D 属性保留效果较好,AP 值分别提高了 17.2% 和 14.1%,同时 mAP 提高了 6.8%,NDS 提高了 3.4%。

关键词:数据增强;稳定扩撒;图像生成;目标检测;特征融合
中图分类号:TP391 **文献标识码**:A **文章编号**:1674 - 8425(2024)10 - 0013 - 08

0 引言

3D 目标检测是自动驾驶领域车辆环境感知的关键内容,而数据增强能够丰富样本特征,有助于提升 3D 目标检测算法的泛化性能,进而为自动驾驶车辆在复杂交通场景中实现智能决策和安全导航提供重要支持。

传统数据增强方法主要通过几何变换、颜色扰动、随机反转和裁剪等简单图像变换来增强数据 2D 特征[1-2],在控制图像中场景或目标的 3D 属性特征方面存在一定局限性。近几年,对抗性数据增强[3-4]、跨模态数据增强[5]等方法通过图像生成的方式较好地丰富了图像的 2D 属性特征。随着潜在扩散模型(latent diffusion models, LDM)[6]的提出,采用步扩散、迭代的图像生成方式能够更好地控制图像的特征和形态,缓解图像细节形变和失真等,为图像生成和数据增强提供了全新的解决方案。如 Burg 等[7]利用扩散模型增强图像 2D 特征成功实现了小样本分类任务,但该方法基于对实验数据的检索和选择,模型泛化性能较差。

目前利用生成技术的图像扩充研究多应用在 2D 任务[8-10],针对 3D 数据增强的研究相对较少。3D 数据生成问题相关研究如 Rombach 等[6]提出的稳定扩散模型(stable diffusion, SD),利用深度输入控制图像生成过程,但单一的控制元素未能较好地控制场景目标的数量和位姿,导致生成图像与原始场景的 3D 属性一致性较差。Li 等[10]提出了一种开放场景下的文本到图像的生成方法(grounded language to image generation, GLIGEN),采用定位边界框或关键点等条件信息作为对齐输入,精确地控制图像生成过程。然而双模态输入的 GLIGEN 方法可控特征相对单一,生成的图像易出现失真、扭曲和目标位置偏移等问题。可以认为,目前图像生成技术多数只针对深度或语义分布等单维度信息进行控制,不

收稿日期:2024 - 01 - 15
基金项目:重庆市科技创新重大研发项目(CSTB2022TIAD-STX0003)
作者简介:张光钱,男,硕士研究生,主要从事智能网联汽车感知方向的研究,E-mail:gqzhang77@163.com。

本文引用格式:张光钱,周广利,黄飞,等. 面向 3D 目标检测的多模态生成式图像数据增强的研究[J]. 重庆理工大学学报(自然科学),2024,38(10):13 - 20.
Citation format:ZHANG Guangqian, ZHOU Guangli, HUANG Fei, et al. A multimodal generative image data enhancement for 3D object detection[J]. Journal of Chongqing University of Technology(Natural Science),2024,38(10):13 - 20.

能同时保证生成场景中目标深度、目标类别以及画面分布等属性不变,因此难以应用于自动驾驶 3D 目标检测任务的数据增强。

针对上述问题,设计了一种多模态图像生成算法,并提出了一种面向 3D 目标检测的数据增强方法。本文中贡献主要包含以下几个方面:

1) 设计了一种基于扩散模型的多模态图像生成算法。其中,设计的多模态特征在线生成模块能高效地提取场景多维特征;提出的增强型门控制自注意力模块(enhanced ggated self-attention,EGSA)有效融合了多模态特征,较好地保留了场景的 3D 属性。

2) 基于所设计的多模态图像生成算法,提出了一种面向 3D 目标检测任务的数据增强方法。该方法保留目标 3D 位置属性的新样本图像与伪标签数据对齐,有效地扩充了数据样本。

1 算法框架

多模态生成式数据增强算法网络框架如图 1 所示。算法充分考虑多模态输入的特点,设计了一种模态特征在线生成模块,自动提取出场景的各模态特征,极大地降低了算法的输入条件。为了继承 SD 算法模型强大的开放场景图像生成能力,算法通过冻结该模型大部分权重的方式继承其大量先验知识。算法在保留原始模型的去噪扩散模型(DDPM)、采样器、文本编码器、变分编码器(VAE)等结构[6]的基础上,针对特征融合网络进行改进设计。特征预处理部分,算法利用卷积网络以及文本编码器对图像输入和文本描述输入进行处理,分别得到嵌入位置编码的模态特征和文本特征序列。

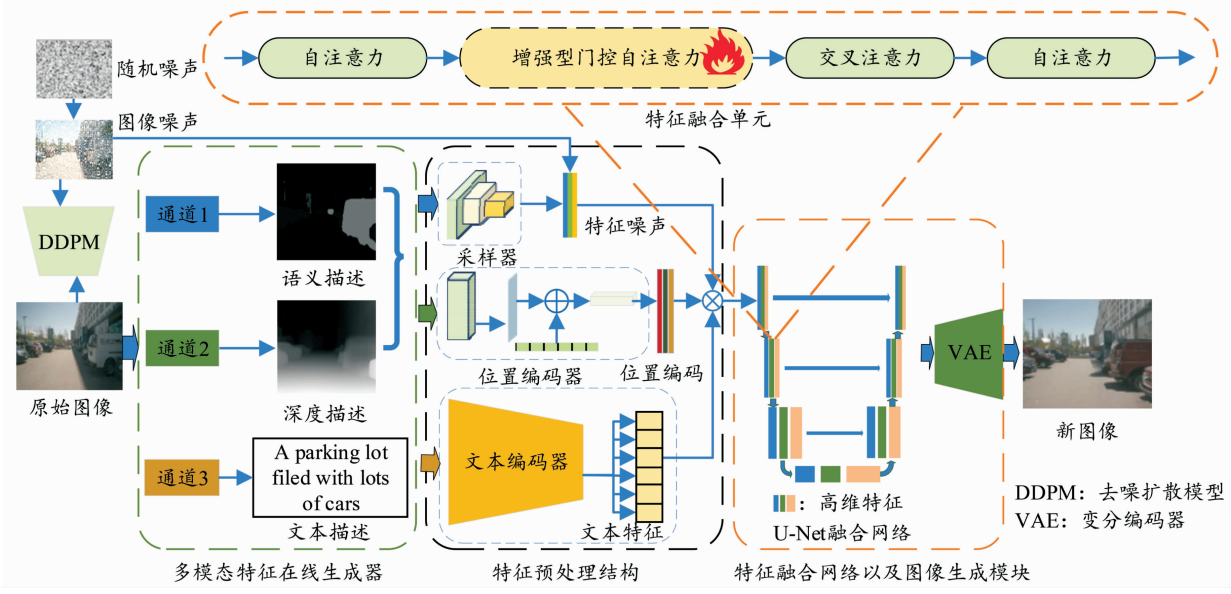

图 1 多模态生成式数据增强算法网络框架结构示意图

在 U 型网络(U-Network,U-Net)[11]中,各模态特征将得到充分融合,多层注意力结构的特征融合单元提高了算法对关键特征的捕捉能力。所提出的 EGSA 模块能够深度融合图像场景的文本、语义以及深度等多模态特征,进一步提升融合网络对于高维度特征的细粒度的理解。此外,变分编码器将得到的高维特征采样生成新图像,使其与原始场景深度保持一致。

本文中多模态图像生成算法生成的新图像具备与原图相近 3D 属性,将其与伪标签数据结合,完成对数据样本的大幅度扩充,具体流程如图 2 所示。新图像的伪标可通过直接映射原始样本标签的方式获得,其与新生成的图像构成了新数据样本。由于生成图像与原始图像具有相同的对象以及其相似的 3D 位置关系,伪标签能够与新图像中的目标一一对应。为保证样本间的差异性,通过设置提示词以及随机参数的方式使每次生成的图像特征在目标颜色、背景天气等方面具有一定随机性,进而丰富样本特征。

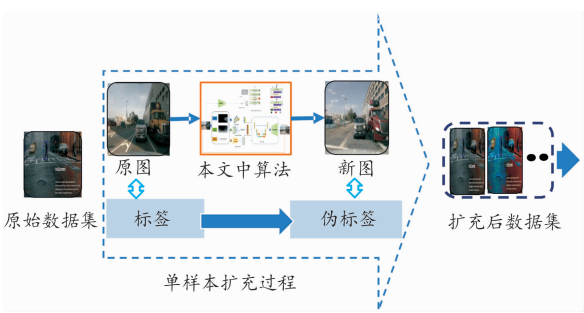

图 2 数据增强流程示意图

1.1 多模态特征在线生成模块

为了降低算法多模态输入特征的准备条件,多模态图像生成算法设计了一个特征在线生成模块,通过 3 个特征生成

通道分离出场景的语义描述、深度描述以及文本描述等各模态特征,具体如图1所示。受到Ade20K数据集[12]的启发,模块在平面维度上设计的通道1使用语义分割模块InterImage[13]分割图像各类别实体,生成场景内各个目标像素级别的150个类别标签,为场景中的目标提供准确的类别描述。单一模态输入的生成过程无法精确控制生成场景中多个目标的类别和具体位置。因此在描述词控制的基础上,需要从不同维度对场景中的目标类别以及目标位置进行不断约束。在空间维度上设计的通道2利用单目深度估计方法Midas[14]提取图像场景像素深度,每个像素深度扩散到各个图片通道,为生成过程提供场景的潜在深度特征。通道3利用BLIP(bootstrapping language-image pre-training)[15]算法生成关于场景信息的文字描述作为生成算法的文本输入,为新图像生成提供了基本场景信息。每个原始场景对应的文本输入基本结构为:

$$W_{\mathrm{n}} + W_{\mathrm{a}} + W_{\mathrm{ca}} + W_{\mathrm{v}} \tag{1}$$

式中:W_{n}、W_{a}、W_{ca}和W_{v}分别为对场景目标数量、目标颜色、目标类别和动作的描述词。

上述设计的多模态特征在线生成模块自动提取原始场景中的多模态特征,为生成算法提供了多维度特征约束。

1.2 特征预处理

为了提高各模态之间的交互效率,需要对各原始模态信息进行处理。场景文本序列记为$\boldsymbol{c}$,潜在深度特征记为$\boldsymbol{d}$,语义特征序列定义为$\boldsymbol{s}$,则算法对齐输入可记为$\boldsymbol{g}=[(s_1,d_1),(s_2,d_2),\cdots,(s_n,d_n)]$,其中$n$为场景中目标数量,网络输入即为$\boldsymbol{I}=[\boldsymbol{c},\boldsymbol{g}]$。

对于文本输入,算法继承了SD中OpenCLIP(open contrastive language-image pre-training)模块以获得描述词的特征序列,记为$\boldsymbol{h}^{\mathrm{c}}=[h_1^{\mathrm{c}},h_2^{\mathrm{c}},\cdots,h_n^{\mathrm{c}}]$。对于语义输入和深度输入,算法采用多层4×4的卷积模块构建空间采样器,通过多次下采样后得到不同模态下与原图采样后相同尺寸的特征噪声$\boldsymbol{h}^{\mathrm{g}}$,具体流程如图3所示。通过该采样器得到在潜在空间中不同模态特征表示如下:

$$\boldsymbol{h}^{\mathrm{g}} = P_{\mathrm{cat}}(f_{\mathrm{s}}(s),f_{\mathrm{d}}(\boldsymbol{d})) \tag{2}$$

$$\boldsymbol{h}^{i} = S(F_{\mathrm{AE}}(i)+n_{\mathrm{r}}) \tag{3}$$

式中:$f_{\mathrm{s}}(\cdot)$和$f_{\mathrm{d}}(\cdot)$分别表示对语义输入和深度输入进行下采样;$P_{\mathrm{cat}}(\cdot,\cdot)$表示在通道维度对输入进行拼接;$i$表示原始图像输入;$n_{\mathrm{r}}$表示随机噪声;$F_{\mathrm{AE}}(\cdot)$表示利用自编码器对其编码;$S(\cdot)$表示扩散模型的采样方法;$\boldsymbol{h}^{i}$表示通过上述方法得到的图像噪声。

为了进一步提取到模态输入的关键特征,式(4)利用轻量ConvNeXt-T[15]网络模块提取模态特征,再通过线性模块嵌入位置编码信息。

$$P(x) = N_{\mathrm{L}}(N_{\mathrm{CN}}(\boldsymbol{x})+\boldsymbol{p}) \tag{4}$$

$$\boldsymbol{h}^{\mathrm{e}} = [P(\boldsymbol{g}_{\mathrm{s}}),P(\boldsymbol{g}_{\mathrm{e}})] \tag{5}$$

式中:$\boldsymbol{x}$表示对齐输入;$N_{\mathrm{CN}}(\cdot)$和N_{L}分别表示ConvNeXt-T网络模块和线性网络模块;$\boldsymbol{p}$表示预设的位置编码;$\boldsymbol{h}^{\mathrm{e}}$表示经过尺寸调整等预处理后得到的高维度特征序列。

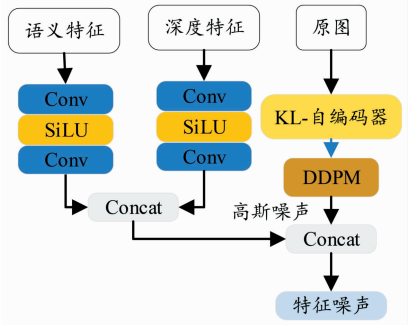

图3 多模态输入下采样流程框图

1.3 特征融合

EGSA模块在U-Net中通过多层注意力结构分级融合潜在空间中复杂的约束信息,保证新图像中场景的3D属性稳定。如图4所示,该模块中不仅设置初始为0的可学习参数λ和α来调节潜在特征和条件信息[16],同时还利用残差结构和多次交叉注意力分层、分次融合语义和深度特征,提高算法对语义分布以及深度位置等多维度特征的控制能力。第一层门控制自注意力模块对应式(6),通过注意力机制提取潜在空间中各模态特征,使模型容易理解场景与目标间的联系,得到一个富含语义信息的潜在特征,即融合特征1。进一步,第二层门控制自注意力模块将深度信息与视觉特征再次融入到空间特征,得到融合特征2。式(7)利用带可学习参数的门限制结构网络调节各种模态特征之间的平衡,促进跨模态的特征融合。

$$\boldsymbol{v} = N_{\mathrm{CA}}(\boldsymbol{v}+T_{\mathrm{s}}(N_{\mathrm{SA}}([\boldsymbol{v},\boldsymbol{h}_{\mathrm{s}}^{\mathrm{e}}])),\boldsymbol{c}) \tag{6}$$

$$\boldsymbol{v} = \boldsymbol{v}+\alpha\cdot\tanh(\lambda)\cdot N_{\mathrm{CA}}(T_{\mathrm{s}}(N_{\mathrm{SA}}([\boldsymbol{v},\boldsymbol{h}_{\mathrm{d}}^{\mathrm{e}}])),\boldsymbol{c}) \tag{7}$$

式中:$\boldsymbol{v}=[v_1,v_2,\cdots,v_N]$表示场景的视觉特征表征,$N_{\mathrm{SA}}$和$N_{\mathrm{CA}}$分别表示自注意力网络和交叉注意力网络,$T_{\mathrm{s}}(\cdot)$表示视觉特征选择器,由其选择图像特征。参数$\alpha$只在训练时为1,并且根据扩散模块采样表参数变化[10]。

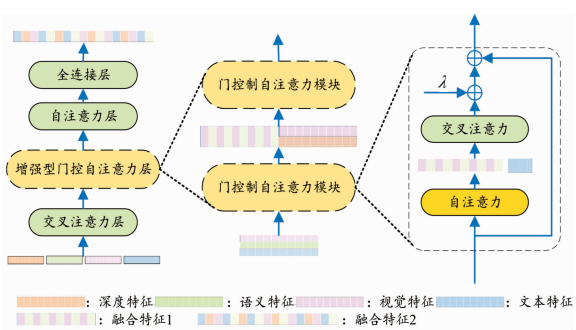

图4 增强型门控制注意力模块结构框图

相较于LDM网络,算法在U-Net中新模块增加参数记为θ',对应的目标函数可以改写成式(8)。

$$\min_{\theta'}\mathcal{L} = \mathbb{E}(Z,\varepsilon)\sim N(0,\boldsymbol{I}),t[\|\varepsilon-f_{(\theta,\theta')}(\boldsymbol{z},\boldsymbol{t},\boldsymbol{I})\|_2^2] \tag{8}$$

式中:t 表示在采样时间集合中均匀采样获得的参数;$f_{(\theta,\theta')}$ 表示变分自编码器。

2 实验验证

2.1 实验设计

为了更好地反映本文中数据增强方法在自动驾驶领域小样本学习任务中的作用,选用 nuScenes 公开数据集 v1.0-mini 进行测试[17],其原始样本较少但场景丰富。实验分为 2 个阶段:在数据增强阶段对原始图像数据分别进行2 倍、3 倍和 5 倍扩充;在验证阶段利用多种单目检测算法验证增强后数据的有效性。

数据增强阶段,实验预设场景描述的结构如式(9)所示,通过随机增改场景描述中特征属性关键词的方式来丰富样本特征,如颜色、天气等。

$$\{W_s\}\cdots\{W_{co}\}+W_{ca}\cdots \tag{9}$$

式中:W_s、W_{co}和 W_{ca}分别表示天气属性词、目标颜色属性词和目标类别描述词,…表示原始图片描述词中其他未改动的内容,$\{\cdot\}$表示在预设属性集合中随机选择一个属性词,其中目标颜色属性词集合为红色、黄色、橙色、白色和蓝色,天气属性词集合为晴天、雨天、多云和夜晚。

算法验证阶段保留了数据集标签中汽车、卡车、公交车、行人、锥桶等关键类别,利用 PGD(probabilistic and geometric depth)[18]、FCOS3D(fully convolutional one-stage monocular 3D object detection)[19]和 Epro-PnP(end-to-end probabilistic perspective-n-points)[20]等 3D 检测算法在扩充前后的数据集中进行训练和测试验证。

实验所用平台配置为:GeForceRTX-4090(24GB)显卡、i9-12700 处理器。为了能够客观反应算法数据增强的能力,验证阶段参照 OpenMMLab[21]开源训练平台参数设置实验参数,统一使用 AdamW 优化器,初始学习率为 0.004,权重衰减为 0.01,数据遍历次数为 21 次。

2.2 实验结果

2.2.1 定性分析

对于针对性地修改图像 2D 特征方面,算法通过随机配置属性关键词的方式有效地控制图像生成,实现效果如图 5 所示。针对天气和光线,新图像各个场景中的建筑物、地面和天空都出现了不同程度的光线变化;针对目标颜色,新图像中汽车颜色变化丰富、图案配色自然;针对场景建筑,新图像中教学楼类型风格以及颜色都发生了对应的显著改变。

针对保留图像场景中的 3D 属性特征,算法实现效果如图 6 所示。

图 5 图像特征生成效果

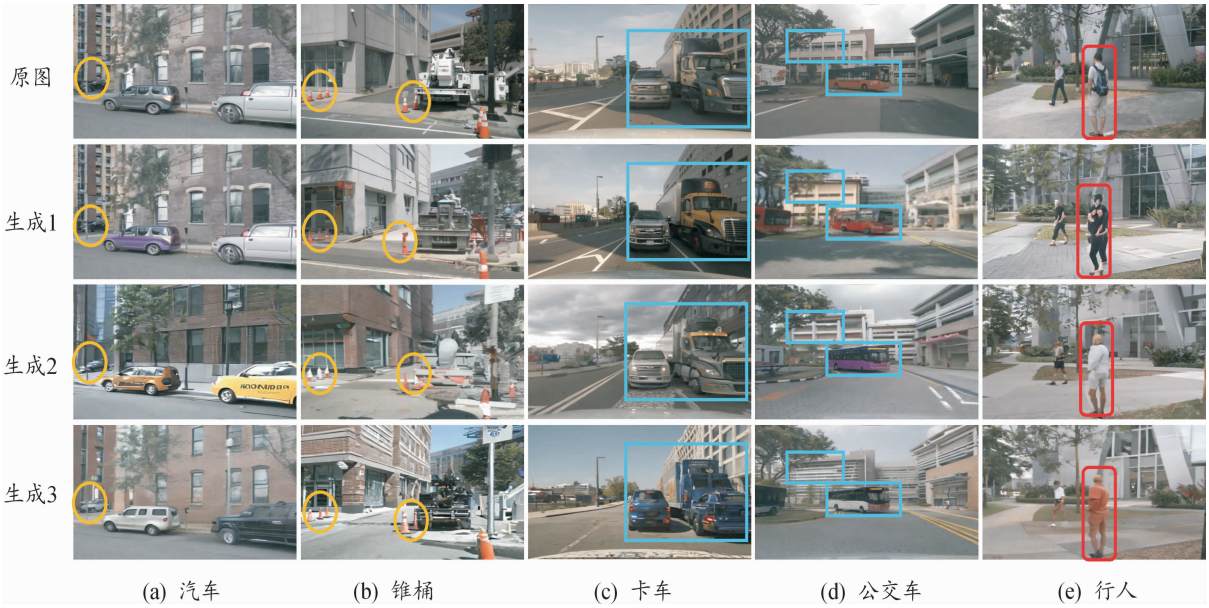

图6 各类目标生成效果

图6中各类目标3D位置基本不变,景深还原度较高,在一定距离内车辆等大尺寸目标生成效果较好,证明本文中所提出的图像生成算法能够较好地融合上述各模态信息。如图6椭圆框位置所示,多次生成结果中,原图像中远景角落处的车辆和车辆附近的锥桶皆被较好地还原,证明算法对于场景边缘的小目标物体也具有较好的生成能力和深度还原能力。图6蓝色直角方框位置所示,新图像场景光照、车辆或建筑物风格、外观图案等变化明显,表现了算法较好的图像生成能力和强大的特征改造能力。只是对于行人等细节较多的目标出现轻微变形,人物局部细节还原度不够,具体如图6行人类别生成结果中红色圆角方框所示。

实验结果表明,新图像各种场景中目标位置、姿态和建筑布局等3D特征都基本得以还原,具备与伪标签构成3D数据样本的条件。该算法不仅能够在2D特征上实现有针对性的修改,为图像样本提供了丰富可控的特征;还能保留图像场景的3D属性特征,为图像生成和数据增强提供可行而有效的解决思路。

图7为本文算法与SD、GLIGEN等类似基于扩散模型的图像生成效果。图7(b)为SD算法以图7(a)作为输入的推理结果。与原始图像相比,图中椭圆框位置生成了多余的车辆并且方框位置车辆位姿变化较大。图7(c)为利用GLIGEN方法将原图深度信息作为输入的推理结果。相较于原图,椭圆框与方框所示位置的车辆语义信息已经丢失,出现生成类别错误、位置偏移、目标丢失的问题。图7(d)为利用本文中算法推理得到的新图像,新图像中车辆颜色、车身图案都发生了变化,且车辆位置以及姿态基本不变,场景深度高度还原,3D位置特征的还原效果优于其他算法。

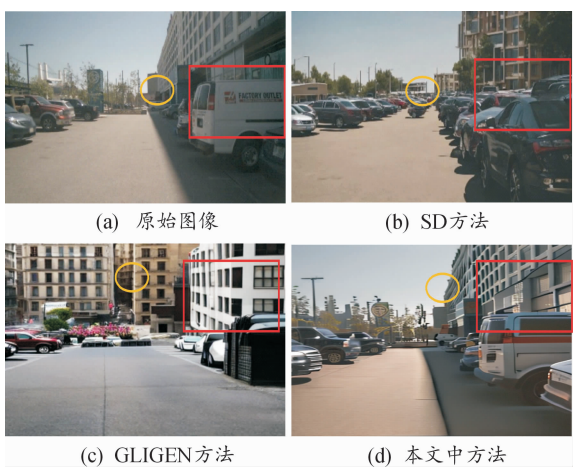

图7 多模态生成效果

利用本文中算法扩充的数据集进行测试验证,单目检测算法对新图像的检测结果如图8所示。在目标密集的停车场环境中,图像场景还原度较高,检测结果较好;在复杂的路口场景中,较远的锥桶和行人等小目标亦能够被还原和检测;在景深较大的街道场景中,道路两边的建筑存在轻微变形,但场景深度基本还原,且较远的微小目标仍能被算法较好地检测。检测算法在新数据中的停车场、路口、街道等多场景中的良好检测表现证明了新图像样本特征的有效性。

图 8　采用本文中数据增强算法后的 3D 单目检测结果

2.2.2　定量分析

图 9 为 PGD 算法在不同倍数扩充数据集进行训练后得到的检测结果。

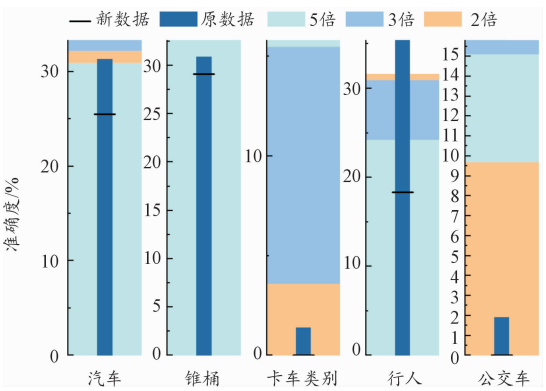

图 9　不同扩充规模下 PGD 算法对不同类别的检测结果

使用等量新数据样本训练检测算法后，汽车、公交车、行人、锥桶、卡车等类别测试 AP 值为 25.5%、0%、18.3%、29.1% 和 0%。由于生成的新图像中存在目标局部轻微变形、未能较好反应目标速度和朝向的问题，相较于原始数据，各类别 AP 值略有下降，分别下降 5.8%、1.9%、17%、1.8% 和 1.4%。如图 6 所示，行人细节较多，生成难度较大，存在明显变形，细节还原度较低，导致图 6 中行人类别在 3 倍扩充数据后 AP 值下降 4.5%。锥桶因其本身的特征较为简单，产生的局部畸变并不影响算法检测精度，3 倍扩充数据后 AP 值提高 4.8%。另外，汽车、货车、公交车等较大目标特征明显，3 倍扩充数据后 AP 值分别提高 3%、17.2% 和 14.1%。各类别目标的检测结果表明，本文中算法生成的新数据与原始数据具有相似目标特征和相近特征质量，生成结果较好地保留了目标原始 3D 位置信息，使样本数据与伪标签高度匹配，有效地扩充了数据样本。由于生成图像中一些中小目标类别存在细节特征质量较差问题，并且算法对数据集中黑夜场景的还原度不高，过度扩充数据样本后造成检测模型泛化性能轻微下降。

表 1 为不同 3D 单目检测算法对增强数据集的测试结果，以原始数据集为基准，其中粗体数字表示该指标明显提升，箭头示意指标的升降情况。

表 1　3D 单目检测算法对增强数据集测试结果

算法	扩充倍数	*mAP* ↑	*mATE* ↓	*mASE* ↓	*mAOE* ↓	*mAVE* ↑	*mAAE* ↑	*NDS* ↑
PGD	基准	0.201	0.972	0.367	1.306	1.594	0.400	0.227
	2	0.229	0.949	0.382	1.434	1.592	0.414	0.240
	3	**0.269**	**0.925**	**0.347**	1.501	1.766	0.470	**0.261**
	5	**0.238**	**0.902**	**0.363**	1.162	1.542	0.480	**0.245**
FCOS3D	基准	0.254	0.977	0.389	1.149	1.541	0.345	0.256
	2	0.262	0.875	0.337	1.583	1.593	0.407	0.269
	3	**0.292**	**0.921**	**0.364**	1.387	1.665	0.435	**0.274**
	5	**0.274**	**0.949**	**0.360**	1.343	1.590	0.483	**0.258**

续表(表1)

算法	扩充倍数	*mAP* ↑	*mATE* ↓	*mASE* ↓	*mAOE* ↓	*mAVE* ↑	*mAAE* ↑	*NDS* ↑
Epro-PnP	基准	0.191	1.015	0.601	1.568	1.257	0.215	0.214
	2	0.201	1.021	0.308	1.509	1.848	0.280	0.242
	3	**0.216**	**0.917**	**0.261**	1.324	2.295	0.394	**0.251**
	5	**0.219**	**0.941**	**0.285**	1.487	1.671	0.336	**0.254**

由表1可知,数据增强后,各检测算法性能指标均有显著提升,尽管平均方向误差(mAOE)、平均速度误差(mAVE)和平均属性误差(mAAE)略微上升,但平均平移误差(mATE)、平均尺度误差(mASE)均有明显下降,说明该数据增强方法对于数据的深度特征以及尺寸特征具有良好的增强效果,该方法有利于提高3D目标检测算法准确预测目标深度、尺寸等3D属性的能力。以数据集3倍扩充为例,各类算法综合性能均有提升,其测试指标中mAP值分别提高了6.8%、3.8%和2.5%,nuScenes检测分数(NDS)分别提高3.4%、1.8%和3.7%。

3 结论

1)提出的多模态图像生成算法,通过引入3个维度的模态特征输入,结合增强型门控自注意力融合机制,提升了算法对与图像3D信息的控制能力。相较于传统单模态或双模态图像生成算法,本文中多模态图像生成算法能够更好地保留场景的3D属性,在多种场景和目标类别的生成结果中表现优异。

2)结合所设计的多模态图像生成算法,通过改变图像背景、场景天气、目标颜色等方式增强了图像特征,并利用新图像场景3D位置还原度较高的特性,结合伪标签实现了面向3D目标检测任务的数据增强功能。在多组测试结果中,多个检测算法的mAP分别提高6.8%、2.3%和2.5%。

面向3D检测的数据增强方法可提高训练数据的多样性和改善模型的泛化性能,为3D数据集增强和小样本学习研究提供借鉴。

参考文献:

[1] ZHANG H Y, CISSé M, DAUPHIN Y N, et al. Mixup: Beyond empirical risk minimization[C]//Proceedings of the 6th International Conference on Learning Representations. Vancouver, 2018.

[2] YUN S, HAN D, OH S J, et al. Cutmix: Regularization strategy to train strong classifiers with localizable features[C]//Proceedings of the IEEE/CVF International Conference on Computer vision. Waikoloa, HI, USA, 2019: 6023 - 6032.

[3] GOODFELLOW I, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//International Conference on Learning Representations. San Diego, CA, USA, 2015.

[4] 陈辉,王硕,许家昌,等. 基于多尺度特征融合生成对抗网络的水下图像增强[J]. 计算机工程与应用,2023,59(21):231 - 241.

[5] HAO X, ZHU Y, APPALARAJU S, et al. Mixgen: a new multi-modal data augmentation[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 379 - 389.

[6] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10684 - 10695.

[7] Burg M F, Wenzel F, Zietlow D, et al. Image retrieval outperforms diffusion models on data augmentation[EB/OL]. (2023 - 04 - 20)[2024 - 01 - 28]. https://arxiv.org/abs/2304.10253.

[8] 师红宇,王嘉鑫,李怡. 基于改进ACGAN算法的带钢小样本数据增强方法[J/OL]. 计算机集成制造系统,https://kns.cnki.net/kcms/detail/11.5946.TP.20230104.1047.004.html.

[9] XIAO C, XU S X, ZHANG K. Multimodal data augmentation for image captioning using diffusion models[C]//Proceedings of the 1st Workshop on Large Generative Models Meet Multimodal Applications. New York, NY, USA: Association for Computing Machinery, 2023: 23 - 33..

[10] LI Y, LIU H, WU Q, et al. Gligen: open-set grounded text-to-image generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 22511 - 22521.

[11] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[J/OL]. (2015 - 03 - 18)[2024 - 01 - 28]. https://arxiv.org/abs/1505.04597.

[12] ZHOU B, ZHAO H, Puig X, et al. Semantic understanding of scenes through the ade20k dataset[J]. International Journal of Computer Vision, 2019, 127: 302 - 321.

[13] WANG W, DAI J, CHEN Z, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions[C]//Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition. 2023: 14408 – 14419.

[14] RANFTL R, LASINGER K, HAFNER D, et al. Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 44(3): 1623 – 1637.

[15] LI J, LI D, XIONG C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//International Conference on Machine Learning. PMLR, 2022: 12888 – 12900.

[16] LIU Z, MAO H, WU C Y, et al. A convnet for the 2020s [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 11976 – 11986.

[17] CAESAR H, BANKITI V, LANG A H, et al. Nuscenes: A multimodal dataset for autonomous driving[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11621 – 11631.

[18] WANG T, Xinge Z H U, PANG J, et al. Probabilistic and geometric depth: Detecting objects in perspective[C]// Conference on Robot Learning. PMLR, 2022: 1475 – 1485.

[19] WANG T, ZHU X, PANG J, et al. Fcos3d: Fully convolutional one-stage monocular 3d object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 913 – 922.

[20] CHEN H, WANG P, WANG F, et al. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 2781 – 2790.

[21] CHEN K, WANG J, PANG J, et al. MMDetection: Open MMLab Detection Toolbox and Benchmark[J/OL]. (2019 – 06 – 17)[2024 – 01 – 28]. https://arxiv.org/abs/1906.07155.

A multimodal generative image data enhancement for 3D object detection

ZHANG Guangqian[1], ZHOU Guangli[2], HUANG Fei[2], LIU Wenbing[2], XIANG Yangkai[1]

(1. School of Mechatronics and Vehicle Engineering, Chongqing Jiaotong University, Chongqing 400074, China; 2. China Road & Bridge Corporation, Beijing 100010, China)

Abstract: The traditional generative image data augmentation algorithms usually lose 3D attribute information, rendering them unsuitable for 3D object detection in autonomous driving. To address the problem, we propose a multimodal image enhancement algorithm based on stable diffusion model. A data augmentation method specifically designed for 3D object detection is developed employing our proposed algorithm. It further constrains the image generation process by introducing more modal inputs. In addition, it has devised a multimodal feature online generation module to extract real-time information such as scene descriptions, semantic distributions, and depth features. Meanwhile, for the multimodal feature fusion network, an enhanced gating self-attention module is designed to accurately capture depth information in the latent feature space. This effectively preserves the 3D attribute information of the image, facilitating targeted modifications to 2D features like texture, color, and illumination. Leveraging the algorithm's exceptional depth-preserving characteristics, the new images are combined with 3D pseudo-labels to create novel image samples, thereby achieving data augmentation for image samples. The 3D detection results on the nuScenes public dataset demonstrate the effectiveness of our algorithm in preserving 3D attributes, particularly for larger categories such as buses and trucks. The AP values exhibit noticeable improvement of 17.2% and 14.1% respectively. Additionally, the indicator of mAP and DNS is increased by 6.8% and 3.4% respectively.

Key words: data enhancement; stable diffusion; image generation; object detection; feature fusion

(责任编辑 林 芳)