Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.087314

ARTICLE

3DGe-Aug: 3D Generative Model-Based Data Augmentation for Few-Shot Object Detection

Jichao Xie[1,*], Xinlei Liu[1], Tong Duan[1], Baolin Li[1], Zhen Zhang[1] and Peng Yi[2]

[1]Information Engineering University, Zhengzhou, China
[2]Security Research Center, Songshan Laboratory, Zhengzhou, China
*Corresponding Author: Jichao Xie. Email: jichaoxie@foxmail.com

Received: 16 June 2026; Accepted: 14 August 2026; Published: 22 October 2026

ABSTRACT: In few-shot object detection scenarios involving emerging, rare, or specialized objects, the lack of diverse training samples severely constrains model performance. Recently, data augmentation methods leveraging 2D text-to-image generative models have gained traction; however, they exhibit limited capabilities in generating fine-grained categories and under-represented objects. To address these limitations, we propose 3DGe-Aug, a 3D generative model-based data augmentation framework for few-shot object detection. Our framework constructs 3D assets from limited 2D images via a 3D generative pipeline, thereby overcoming the rendering limitations of 2D generative models for fine-grained objects. Building upon these 3D assets, multi-perspective object images with rich viewpoints and high spatial-geometric consistency are rendered through precise camera pose control. Furthermore, we design an instance pasting algorithm that incorporates multi-scale transformations and dynamic occlusion elimination mechanisms to synthesize training data. Comprehensive experiments conducted across two representative object detection scenarios—unmanned aerial vehicles (UAVs) and autonomous driving—validate the effectiveness of the proposed method across various detection architectures, such as RT-DETR and the YOLO series. To facilitate future research, we also contribute a few-shot object detection benchmark dataset comprising three base classes and six novel classes, along with a corresponding 3D asset set. Our code and datasets are publicly available at https://github.com/xiejichao/DataAug.

KEYWORDS: Few-shot object detection; data augmentation; 3D generative models; multi-perspective rendering; instance pasting

1 Introduction

Object detection is a fundamental task in computer vision, aiming to localize and classify objects in images or videos. It has diverse applications spanning smart wearables, intelligent transportation, industrial inspection, medical image analysis, and military reconnaissance [1,2]. However, training object detection models relies heavily on large-scale annotated datasets, which are labor-intensive and expensive to construct. When dealing with emerging or rare object categories, obtaining real training samples with diverse viewpoints and rich backgrounds becomes prohibitively challenging, degrading model performance under such few-shot settings [3,4]. Therefore, efficiently and scalably synthesizing training data holds significant research value [5].

To mitigate these challenges, numerous data augmentation methods have been proposed in both academia and industry. Among these, Copy-Paste [6] is widely recognized as a simple yet effective strategy.

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

However, this method relies on the availability of object instances within the training set, rendering it ineffective when target categories are entirely absent. To synthesize absent or rare object instances, researchers have proposed several methods [5,7–9] based on text-to-image generative models. These methods demonstrate strong generative capabilities for common object categories, such as motorcycles, sofas, cats, and dogs. However, due to the inherent limitations of 2D generative models, they struggle to generate fine-grained vehicle categories, emerging objects, or specialized equipment [8]. For instance, Latent Diffusion Models (LDMs) [10], a representative text-to-image method, often fail to generate accurate fine-grained structures, as exemplified in the middle row of Fig. 1. Additionally, X-Paste [5] complements text-to-image generation with a retrieval-based image collection strategy. Although this expands the object image pool by retrieving existing resources from the internet, it suffers from inaccurate category verification and a lack of perspective diversity.



Figure 1: Comparison of object instances generated by text-to-image and 3D generative models.

The rapid progress of open-source 3D generative models [11–13] has enabled the reconstruction of realistic 3D object assets from a single or a few 2D images. Such generative approaches substantially reduce the cost of 3D asset creation and promote their widespread deployment in downstream domains, including film and gaming. Motivated by this progress, we propose a data augmentation method for few-shot object detection based on 3D generative models. Our method leverages multi-perspective images rendered from 3D assets to enhance the viewpoint diversity and geometric richness of few-shot object data. Representative 3D-generated objects are shown at the bottom of Fig. 1.

To effectively integrate diverse object images into the copy-paste framework, we propose 3DGe-Aug, a unified data generation and processing framework. As illustrated in Fig. 2, the pipeline is concise and efficient, consisting of the following three stages: **(1) 3D Asset Construction.** For few-shot objects, the framework employs a 3D generative model to reconstruct 3D assets with detailed geometric structures and surface textures from either a single-view image or a limited set of multi-view images, laying a solid foundation for subsequent multi-perspective rendering. **(2) Multi-Perspective Image and Mask Generation.** Leveraging computer graphics rendering, the spatial pose of a virtual camera is precisely controlled within a spherical coordinate system to automatically generate large-scale, viewpoint-rich 2D object images. Subsequently, a salient object segmentation model is employed to extract accurate object masks, effectively eliminating irrelevant backgrounds. **(3) Instance Pasting.** We design an instance pasting algorithm that incorporates multi-scale transformation and occlusion elimination mechanisms. This module not only simulates the physical phenomenon of multi-scale object distribution but also dynamically filters out heavily occluded object instances by computing inter-object overlap, thereby augmenting few-shot object samples while maintaining high reliability in label annotations.

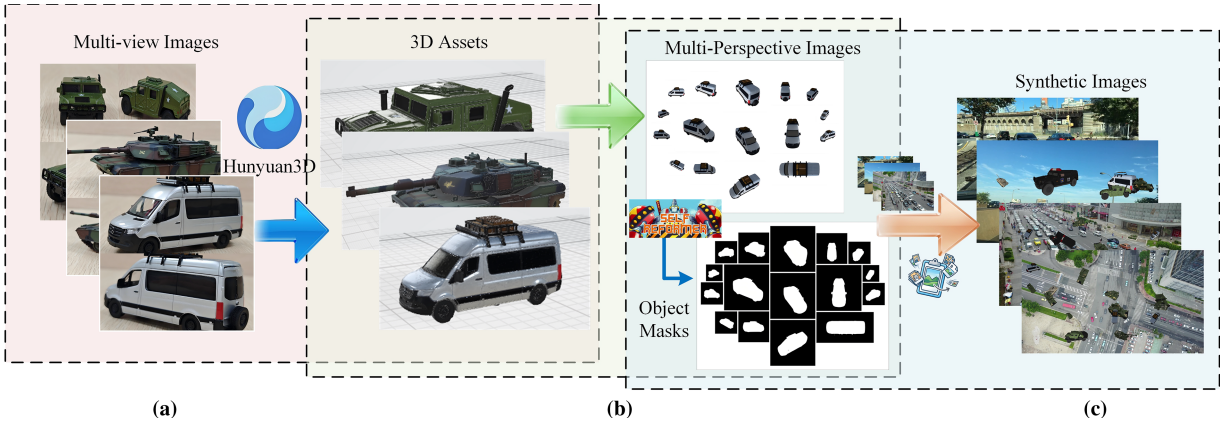


Figure 2: Overview of the proposed 3DGe-Aug framework for few-shot object detection data augmentation. (**a**) 3D asset construction; (**b**) Multi-perspective image and mask generation; (**c**) Instance pasting.

Additionally, to validate the effectiveness of the proposed method, we construct few-shot object detection datasets for two representative scenarios: UAV-based object detection and autonomous driving. These datasets comprise three common base classes (Car, Truck, Bus) and six novel classes (Humvee, ICV, Mengshi, PLZ-05A, Sprinter, ZTQ-15).

The main contributions of this paper are summarized as follows:

- We propose 3DGe-Aug, a data augmentation framework for few-shot object detection based on 3D generative models. This framework effectively addresses the generation limitations of traditional text-to-image models when dealing with fine-grained, scarce, and emerging objects. By constructing high-quality 3D assets and rendering multi-perspective images, it significantly enhances the diversity and spatial-geometric consistency of the few-shot object feature space.
- We design a training data synthesis algorithm that integrates multi-perspective image sampling, multi-scale transformation, and dynamic occlusion elimination. This algorithm not only reproduces complex visual representations, including multi-perspective views, multi-scale variations, and natural object overlaps in real scenes, but also effectively filters out heavily occluded instances whose discriminative features are degraded.
- We conduct extensive experiments to validate the effectiveness of the proposed framework. On the VisDrone dataset for UAV object detection, we achieve mean average precision (mAP) scores of 0.828 for all classes and 0.883 for novel classes, representing improvements of 6.0% and 8.3% over the baseline X-Paste method. On the BDD100K dataset for autonomous driving, our method achieves an average detection accuracy of 0.636 across all classes, corresponding to a 21.9% improvement over the X-Paste method. Furthermore, we open-source the project code, release a 3D asset set constructed from multi-perspective images, and contribute a few-shot test dataset comprising nine object classes.

2 Related Work

Few-Shot Object Detection. Object detection has achieved remarkable success, driven by the availability of massive annotated datasets. However, in many practical scenarios, collecting and labeling large-scale real-world data is often expensive or even infeasible. Few-shot object detection (FSOD) aims to identify novel objects using only a small number of samples, thereby alleviating model reliance on abundant annotated data [3]. Currently, mainstream FSOD approaches are categorized into three paradigms: meta-learning, transfer learning, and data augmentation [14,15].

Among them, meta-learning-based methods [16] adhere to the core principle of "learning to learn." These methods typically adopt an episodic training strategy, constructing subtasks composed of support and query sets to acquire task-agnostic prior knowledge. This class of methods exhibits rapid adaptation capabilities and can extract key features from a small number of support samples. However, the training process is complex and prone to convergence difficulties; moreover, computational efficiency degrades significantly as the number of classes or samples increases [15].

Transfer learning-based methods are grounded in the principle of knowledge transfer. These methods [4] first pre-train a feature extractor on sample-abundant base classes, freeze the backbone network, and fine-tune only the classification and regression heads on a balanced dataset containing both base and novel classes. This approach is simple in structure and robust in performance, effectively preserving general discriminative features learned from base classes. However, such methods are prone to overfitting when novel class samples are scarce, failing to capture the full distribution of visual variations and leading to degraded performance under significant viewpoint changes.

Although meta-learning and transfer learning algorithms have achieved significant progress in model architecture optimization and learning strategy refinement, the lack of feature richness and perspective diversity in training data remains a fundamental bottleneck under extreme few-shot conditions. This has prompted a shift toward data-centric approaches, aiming to directly enrich the feature space of novel objects via generative data augmentation.

Data Augmentation Methods for Few-Shot Object Detection. Among data augmentation methods, Copy-Paste [6] is a simple yet powerful strategy that boosts model performance by copying object instances from one image and pasting them into the complex background of another. However, traditional copy-paste heavily relies on existing object instances in the dataset, encountering a "nothing-to-paste" bottleneck when dealing with completely novel or unseen categories.

To overcome data scarcity, recent studies have increasingly employed text-to-image generative models to synthesize training data. For instance, X-Paste [5] combines CLIP and Stable Diffusion with an image retrieval mechanism to scale up object instance data. This method either synthesizes objects via generative models or filters web-crawled images using zero-shot recognition models, significantly improving the scalability of the copy-paste strategy. DataAugmentation [7] explores data augmentation for object detection using controllable diffusion models. This method constrains the diffusion inpainting process by introducing visual priors (e.g., bounding box specifications) and leverages class-calibrated CLIP scores for post-hoc data filtering, effectively ensuring the semantic correctness of synthesized objects within specified regions. LLM-DiffAug [8] leverages Large Language Models (LLMs) to guide diffusion models, thereby enhancing data diversity for few-shot object detection. MPAD [9] tackles the narrow variance of synthetic samples by introducing a multi-perspective augmentation framework that generates typical samples via fine-grained attribute prompts, creates hard samples by mixing base and novel classes, and filters hard backgrounds based on visual similarity, thereby mitigating overfitting and enhancing generalization.

Although these text-to-image-based methods have achieved remarkable success on common objects such as motorcycles, sofas, cats, and dogs, they struggle to generate accurate object instances for fine-grained categories, under-represented emerging objects, or specialized military equipment. This limitation restricts their applicability in challenging few-shot object detection tasks. To more explicitly highlight the differences and our unique contributions, we summarize the core ideas, addressed problems, and limitations of existing few-shot data augmentation methods compared to our proposed 3DGe-Aug in Table 1.

Table 1: Comparison of data augmentation methods for few-shot object detection.

Method	Core Idea	Primary Problem Addressed	Limitation
Copy-Paste [6]	Copies and pastes existing object instances within the training set.	Insufficient number of object instances in the training set.	Fails when target objects are entirely absent from the dataset (nothing to paste).
X-Paste [5]	Generates object instances via text-to-image models and retrieves images.	Insufficient number of object instances in the training set.	Limited generation capability for fine-grained and rare objects; potential image retrieval failures.
DataAugmentation [7]	Augments existing instances in training images using controllable diffusion models.	Lack of rich feature representations for training instances.	Inapplicable when target objects are completely absent from the training set.
LLM-DiffAug [8]	Generates object instances via LLM-guided text-to-image models.	Lack of rich feature representations for training instances.	Limited generation capability for fine-grained and rare categories.
MPAD [9]	Generates object instances via text-to-image models to construct hard samples.	Lack of hard samples among training instances.	Limited generation capability for fine-grained and rare categories.
3DGe-Aug (Ours)	Constructs 3D assets of objects and renders multi-perspective images.	Absence of objects for copy-pasting and limited generation capability of 2D text-to-image models.	Potential domain shift between synthetic and real instances on large-scale datasets.

3D Generative Models. Unlike 2D image generative models, 3D generative models aim to construct 3D assets with consistent geometric topology and surface materials. In recent years, driven by the deep integration of deep learning and computer graphics, 3D generation technology has achieved remarkable breakthroughs. For example, the open-source Hunyuan3D model [11] can rapidly reconstruct production-grade 3D assets with physically based rendering (PBR) materials from a single or a few 2D prompt images. 3D generation technology has not only revolutionized game and film asset creation pipelines, but has also demonstrated significant potential as a controllable synthetic data generator for computer vision perception tasks such as autonomous driving [17].

Compared to 2D images produced by text-to-image models, 3D assets exhibit superior spatial-geometric consistency, enabling continuous and precise camera pose control within a spherical coordinate system and rendering 2D object projections under arbitrary lighting conditions and viewing distances. In this

paper, we leverage 3D generative models to reconstruct 3D assets of scarce objects and render geometrically consistent multi-perspective views, thereby overcoming the structural synthesis limitations of 2D generative approaches for fine-grained and specialized categories.

3 Methodology

While prior data augmentation methods leveraging text-to-image generative models demonstrate strong capabilities in synthesizing common object categories, they struggle when dealing with fine-grained, rare, or emerging objects that are under-represented in generative pre-training data, failing to produce accurate object images.

To address this limitation, we investigate how to automatically and efficiently synthesize multi-perspective views of rare objects to augment the training sets. We achieve this by constructing 3D assets of rare objects and rendering diverse perspective images. To seamlessly integrate these multi-perspective images into a copy-paste framework, we propose 3DGe-Aug, a unified data synthesis and processing framework. As illustrated in Fig. 2, the proposed framework comprises three core modules: 3D asset construction, multi-perspective image and mask generation, and instance pasting. The following subsections elaborate on each module.

3.1 3D Asset Construction

To address the scarcity of real samples for rare objects in few-shot object detection, our method first utilizes a small set of 2D images to construct 3D assets. By building 3D meshes with accurate geometric structures and detailed surface textures, we lay a solid foundation for rendering downstream spatially and geometrically consistent multi-perspective views.

Formally, let the input set of 2D object images be denoted as $\mathcal{I} = \{I_{v_i}\}_{i=1}^{N}$, where I_{v_i} denotes an image captured from a specific viewpoint v_i. In our framework, v_1 corresponds to the front view of the object, which serves as the primary visual prior for driving the 3D reconstruction process.

- **Multi-view joint reconstruction**: When $N > 1$, the input set includes the front view alongside additional viewpoints (e.g., left, right, top). The 3D generative model infers the complete object geometry through cross-view feature matching and fusion, effectively minimizing structural uncertainty in unobserved or occluded regions.
- **Single-view reconstruction**: When $N = 1$, the input set simplifies to $\mathcal{I} = \{I_{v_1}\}$. In this single-view mode, the framework relies on strong geometric priors pre-trained within the 3D generative model to complete unobserved surface structures. Single-view reconstruction can thus be viewed as a special case of multi-view reconstruction.

The proposed framework leverages an advanced 3D generative model $\mathcal{F}_{3D}$ (e.g., Hunyuan3D [11]). Given a multi-view object image set $\mathcal{I}$, the 3D generation process can be formalized as:

$$\mathcal{M} = \mathcal{F}_{3D}(\mathcal{I}; \Theta) \tag{1}$$

where Θ denotes the pre-trained parameters of the 3D generative network. The generated high-dimensional 3D object $\mathcal{M}$ is a mesh embedded with rich material and structural information, represented as a triplet:

$$\mathcal{M} = (\mathcal{V}, \mathcal{F}, \mathcal{T}) \tag{2}$$

In this triplet, $\mathcal{V} \in \mathbb{R}^{V\times 3}$ denotes the set of 3D spatial coordinates of V vertices, defining the geometric shape of the object; $\mathcal{F} \in \mathbb{Z}^{F}$ denotes the F polygonal faces connecting these vertices, forming the topological surface; $\mathcal{T}$ denotes the Physically Based Rendering (PBR) texture map.

Implementation Details. To balance fine geometric detail preservation with computational efficiency during subsequent batch rendering, we cap the face count of the 3D model at 500k. Furthermore, to maximize visual fidelity, we activate PBR material generation mode. The PBR texture $\mathcal{T}$ explicitly defines the albedo, roughness, and metallic properties of the surface, fostering realistic light-shadow interactions during multi-perspective rendering [11]. The final 3D object is uniformly exported as a GLB file to interface seamlessly with the downstream multi-perspective rendering and mask generation pipeline.

3.2 Multi-Perspective Image and Mask Generation

After constructing the 3D asset, we employ computer graphics rendering techniques to automatically generate large-scale, viewpoint-rich 2D object projections and leverage salient object segmentation to produce precise masks. The general workflow is illustrated in Fig. 3, laying the foundation for downstream instance pasting onto real background images.

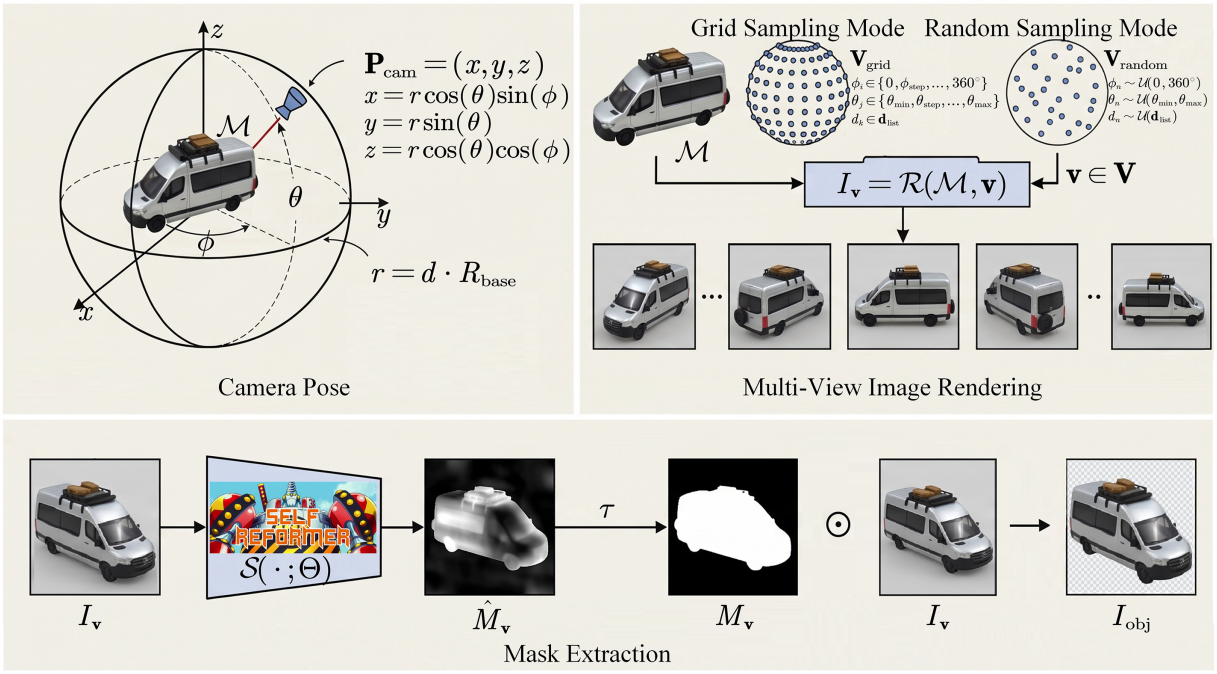


Figure 3: Multi-perspective image and mask generation.

To render 2D images from diverse viewpoints using the 3D asset $\mathcal{M}$, we control the virtual camera pose within a spherical coordinate system. To ensure that the camera remains focused on the object center while avoiding collisions, we compute the base radius R_{base} from the bounding box of the model. The spatial position of the virtual camera is determined by three spherical coordinate parameters: the azimuth angle ϕ, elevation angle θ, and distance multiplier d. The distance r from the camera to the object center is defined as:

$$r = d \cdot R_{\text{base}} \tag{3}$$

To prevent gimbal lock when the camera is directly overhead ($\theta = 90°$), we dynamically compute the camera's up vector $\mathbf{u}$. The 3D Cartesian coordinates of the camera $\mathbf{P}_{\text{cam}} = (x, y, z)$ are computed as:

$$\begin{cases} x = r\cos(\theta)\sin(\phi) \\ y = r\sin(\theta) \\ z = r\cos(\theta)\cos(\phi) \end{cases} \tag{4}$$

Given the object center coordinates and the dynamic up vector $\mathbf{u}$, the camera pose is determined via the standard LookAt matrix.

To accommodate varying demands for viewpoint diversity across different tasks, we design two multi-perspective image sampling modes:

- **Grid Sampling Mode.** Uniform discrete sampling is performed within preset angle and distance ranges. Given the azimuth range $[0°, 360°]$ with step size ϕ_{step}, the elevation range $[\theta_{\min}, \theta_{\max}]$ with step size θ_{step}, and a distance list $\mathbf{d}_{\text{list}}$, the grid sampling parameter space $\mathbf{V}_{\text{grid}}$ is defined as:

$$\mathbf{V}_{\text{grid}} = \left\{ (\phi_i, \theta_j, d_k) \;\middle|\; \begin{array}{l} \phi_i \in \{0, \phi_{\text{step}}, \ldots, 360° - \phi_{\text{step}}\}, \\ \theta_j \in \{\theta_{\min}, \theta_{\min} + \theta_{\text{step}}, \ldots, \theta_{\max}\}, \\ d_k \in \mathbf{d}_{\text{list}} \end{array} \right\} \tag{5}$$

  where $\mathbf{d}_{\text{list}}$ is the predefined distance list controlling the relative distance between the camera and the object center. This mode systematically traverses the object's visual appearance across structured spatial orientations.
- **Random Sampling Mode.** To enhance data randomness and mitigate the discrete artifacts of fixed grids, given a sampling budget N, the algorithm randomly samples camera parameters from uniform distributions:

$$\mathbf{V}_{\text{random}} = \{(\phi_n, \theta_n, d_n)\}_{n=1}^{N} \tag{6}$$

  where $\phi_n \sim \mathcal{U}(0, 360°)$, $\theta_n \sim \mathcal{U}(\theta_{\min}, \theta_{\max})$, and $d_n \sim \mathcal{U}(\mathbf{d}_{\text{list}})$. This mode yields unstructured observations.

Given a specific camera parameter combination $\mathbf{v} \in \mathbf{V}$, the corresponding multi-view image $I_{\mathbf{v}} = \mathcal{R}(\mathcal{M}, \mathbf{v})$ is rendered via the rasterization function $\mathcal{R}$.

To eliminate redundant backgrounds and isolate target instances from the rendered images, we extract precise object masks using a salient object segmentation framework [18]. Specifically, we employ SelfReformer [18] to execute accurate mask extraction. Given the input image $I_{\mathbf{v}} \in \mathbb{R}^{H \times W \times 3}$, the object mask generation network is defined as a non-linear mapping $\mathcal{S}(\cdot; \Theta)$, where Θ denotes the pre-trained network parameters. The mask extraction process is formalized as:

$$\hat{M}_{\mathbf{v}} = \mathcal{S}(I_{\mathbf{v}}; \Theta), \tag{7}$$

where $\hat{M}_{\mathbf{v}} \in [0,1]^{H \times W}$ is the continuous probability map predicted by the network. To obtain sharp boundaries, a threshold τ (e.g., $\tau = 128/255$) is applied for binarization, yielding the final binary mask $M_{\mathbf{v}}$:

$$M_{\mathbf{v}}(i, j) = \begin{cases} 1, & \text{if } \hat{M}_{\mathbf{v}}(i, j) > \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Finally, the clean object instance image $I_{\text{obj}} = I_{\mathbf{v}} \odot M_{\mathbf{v}}$ for subsequent background fusion is obtained via the Hadamard product of the original image and its mask. This decoupled formulation ensures accurate instance extraction, enabling seamless integration of the generated object into complex real-world backgrounds without introducing rendering artifacts.

3.3 Instance Pasting

Having rendered multi-perspective object images and generated their precise masks, we synthesize augmented training samples by integrating these generated instances into complex real background images. To reflect the physical dynamics of scale variations and mutual occlusion in natural scenes, we propose an instance pasting algorithm that incorporates dynamic random scaling and occlusion-aware label elimination.

Formally, let the original dataset comprise a background image set $\mathcal{B}$ and a label set $\mathcal{L}_{\text{ori}}$. The goal of augmentation is to produce an augmented image set $\mathcal{B}'$ with updated labels $\mathcal{L}_{\text{aug}}$.

For each background image $B \in \mathcal{B}$, the corresponding ground-truth annotations are represented as $L_{\text{ori}} = \{b_i \mid b_i = (c_i, x_i, y_i, w_i, h_i)\}_{i=1}^{M}$, where c_i denotes the object class and (x_i, y_i, w_i, h_i) are normalized bounding box coordinates in YOLO format. During augmentation, the algorithm randomly samples K multi-view instances from the object database for pasting.

Let I_{obj}^{j} and M_{obj}^{j} denote the j-th object instance image and its corresponding binary mask, respectively. To mimic perspective-dependent scale variations resulting from varying object distances, we introduce an area ratio range $[S_{\min}, S_{\max}]$ to govern the scale of the pasted instance relative to the total background area.

For the j-th instance, we sample the area ratio S_{obj}^{j} from a uniform distribution:

$$S_{\text{obj}}^{j} \sim \mathcal{U}(S_{\min}, S_{\max}). \tag{9}$$

While preserving the original aspect ratio, the scaled object dimensions (w_j', h_j') satisfy:

$$w_j' \cdot h_j' = S_{\text{obj}}^{j} \cdot W_{\text{bg}} \cdot H_{\text{bg}}, \tag{10}$$

where W_{bg} and H_{bg} denote the width and height of the background image in pixels.

Once the pasting center coordinates are randomly determined, pixel-level compositing is executed using the mask. We allow natural overlaps among newly pasted instances as well as between pasted and pre-existing objects. The final augmented image B' after sequentially pasting K objects onto background B is derived via iterative mask-based blending:

$$B^{(j)} = B^{(j-1)} \odot (1 - M_{\text{obj}}'^{j}) + I_{\text{obj}}'^{j} \odot M_{\text{obj}}'^{j}, \quad j = 1, \ldots, K, \tag{11}$$

where $B^{(0)} = B$, $B^{(K)} = B'$, and $M_{\text{obj}}'^{j}$ and $I_{\text{obj}}'^{j}$ denote the scaled and translated binary mask and instance image for the j-th object, respectively.

Because inter-object overlap is permitted, pre-existing or previously pasted objects may become heavily occluded. To prevent such heavily occluded instances—whose discriminative visual features are severely degraded—from introducing noisy labels during training, we introduce a dynamic label elimination mechanism based on occlusion ratios.

Given an occlusion threshold γ, when the bounding box of a newly placed object b_{new} overlaps with an existing object b_{old}, the occlusion ratio of b_{old} is defined as:

$$\text{Occ}(b_{\text{old}}, b_{\text{new}}) = \frac{\text{Area}(b_{\text{old}} \cap b_{\text{new}})}{\text{Area}(b_{\text{old}})}. \tag{12}$$

Throughout the sequential pasting process, the validity of all active labels is dynamically re-evaluated. If $\mathrm{Occ}(b_{\mathrm{old}}, b_{\mathrm{new}}) > \gamma$, the underlying object is deemed severely occluded and purged from the annotation set.

Let $L^{(j-1)}$ represent the label set prior to inserting the j-th instance b^j_{new}. The label update rule following the j-th insertion is formalized as:

$$L^{(j)} = \left\{ b \in L^{(j-1)} \mid \mathrm{Occ}(b, b^j_{\mathrm{new}}) \leq \gamma \right\} \cup \{ b^j_{\mathrm{new}} \}. \tag{13}$$

After completing K iterations of scale sampling, compositing, and label filtering, the final augmented annotation set is obtained as $\mathcal{L}_{\mathrm{aug}} = L^{(K)}$. This mechanism expands object quantity and enriches feature representations while preserving annotation integrity via strict label filtering. Representative pasted examples are illustrated in Fig. 4, featuring VisDrone [19] on the left and BDD100K [20] on the right.



Figure 4: Examples of object instance pasting on VisDrone (**left**) and BDD100K (**right**).

It is worth noting that the current compositing strategy relies on a simple random pasting mechanism. As a result, pasted objects may occasionally exhibit visual inconsistencies with the target scene in terms of illumination, shadowing, contact geometry, perspective, road-plane alignment (in autonomous driving), and altitude-aware scaling (in UAV scenarios). We explicitly acknowledge these visual inconsistencies as a limitation of the current pipeline. However, as demonstrated by Ghiasi et al. [6], the simple mechanism of randomly pasting objects is often sufficiently powerful and can provide significant improvements over strong baselines. In our preliminary small-scale experiments, we also observed that context-aware or reference-based placement strategies did not yield better training outcomes than random pasting. Therefore, we adopt random pasting as the default strategy to maintain algorithmic simplicity and high efficiency. Addressing these visual discrepancies through advanced compositing techniques—such as shadow synthesis, dynamic lighting adaptation, and perspective-aware placement—remains an important direction for future research.

4 Experiments and Analysis

4.1 Experimental Setup

Datasets. We conduct few-shot object detection experiments on the VisDrone UAV dataset [19] and the BDD100K autonomous driving dataset [20]. VisDrone contains 6471 training and 548 validation images, while BDD100K contains 69,863 training and 10,000 validation images. We select 3 classes shared by both datasets as base classes (Car, Truck, Bus) and incorporate 6 novel classes (Humvee, ICV, Mengshi, PLZ-05A, Sprinter, ZTQ-15). Additionally, all background images used for instance pasting are strictly drawn from the official training splits of the VisDrone and BDD100K datasets.

To evaluate model performance, we construct the few-shot-object-aerial test set (251 images, containing 5459 total instances with 1434 novel class instances) for UAV detection and the few-shot-object-ground test set (103 images, containing 1186 total instances with 407 novel class instances) for autonomous driving detection. Both test sets contain the aforementioned 9 classes with viewpoints consistent with their respective scenarios. To prevent data leakage, we ensure no overlap between the images in these test sets and the 2D source images used for 3D asset construction. Furthermore, all novel dataset annotations were assisted by the Segment Anything Model (SAM) on the Ultralytics platform. To ensure high-quality annotations, we implemented a rigorous three-tier quality control protocol: initial annotation by one researcher, cross-checking by a second researcher, and final verification by the project lead. For evaluation, we merge the original validation sets with our constructed test sets to comprehensively evaluate recognition performance on both base and novel classes.

Object Detection Models. We build the training and validation environment based on Ultralytics [21] and evaluate our method on RT-DETR, YOLOv5, YOLOv8, YOLOv11, and YOLOv26. We use COCO pre-trained weights with default hyperparameters from Ultralytics. Models are trained at a resolution of 640×640 with a batch size of 16, for 50 epochs on VisDrone and 10 epochs on BDD100K.

3DGe-Aug Settings and Implementation Details. For 3D asset construction, we directly invoke the Tencent Cloud Hunyuan3D API (API version: ai3dv20250513, Model version: v3.1) [11]. The generation type is set to Normal mode (generating geometry with textures), EnablePBR is turned on for realistic lighting interactions, and the output format is background-free GLB with a face count at 500k. The 2D source images used for reconstructing novel objects were captured via smartphones from physical scale models across multiple angles. Specifically, the multi-view 3DGe-Aug pipeline utilizes 6 images (front, left, right, top, left-front, right-front), whereas the single-view 3DGe-Aug-s uses only a single front-view image. To strictly prevent data leakage, none of these source images were included in downstream object detection training or testing sets. In terms of computational overhead and deployment costs, reconstructing a 500k-face 3D asset takes an average of 239.0 s and costs 0.5 RMB for the multi-view setup, yielding an average storage footprint of 49.9 MB. The single-view reconstruction requires 218.8 s and costs 0.4 RMB. During data synthesis, rendering a single multi-perspective image takes approximately 0.0635 s, while extracting the corresponding background mask using SelfReformer [18] averages 0.0331 s per image. Camera sampling parameters are set as follows: elevation angle is $[30°, 90°]$ for UAV and $[0°, 30°]$ for autonomous driving; azimuth is $[0°, 360°]$ for both. Elevation and azimuth step sizes are selected from $\{5°, 10°, 15°, 20°, 25°, 30°, 35°\}$, and sampling distance is 1. For dataset synthesis, the number of pasted objects per image is selected from $\{1, 2, 3, 4, 5, 6, 8, 10, 15, 20\}$. The area ratio range is $[0.005, 0.05]$ and the occlusion threshold is 0.7. It should be noted that the choice of the mask extractor, the occlusion threshold, the area range, the lighting condition, and the PBR configuration were fixed as empirical settings rather than being experimentally optimized. Therefore, the optimality of these specific parameters has not been exhaustively verified.

Compared Methods. We use **X-Paste** [5] as our primary baseline, which combines image generative models with retrieval mechanisms. Using this method, we retrieved 279 real-world images for the six novel classes (an average of 46.5 per category) from the internet. Our method using multi-view joint reconstruction is denoted as **3DGe-Aug**, while the single-view variant is **3DGe-Aug-s**. The combination of X-Paste and 3DGe-Aug is denoted as **3DGe-Aug-XPa**. Training without any augmentation serves as the **Baseline** to evaluate the impact of different augmentation strategies on base and novel class detection.

Furthermore, evaluating recent state-of-the-art data augmentation methods requires addressing their inherent prerequisites. As qualitatively demonstrated by the empirical failure cases in the middle row of Fig. 1, current text-to-image models fundamentally struggle to synthesize accurate visual features for fine-grained and emerging novel classes. Consequently, recent methods strictly reliant on these text-to-image generative priors, such as LLM-DiffAug [8] and MPAD [9], inherit this limitation and cannot stably produce valid training instances for our task, making them unsuitable for direct quantitative comparison.

Similarly, classical Copy-Paste [6] and recent controllable diffusion methods like DataAugmentation [7] fundamentally require existing object instances within the training set to serve as source material. Since our training sets lack instances of the 6 novel classes, these methods are structurally inapplicable. X-Paste [5] bypasses the text-to-image generation bottleneck via a web image retrieval mechanism, making it our primary strong baseline, though it undeniably remains vulnerable when dealing with ultra-rare objects that cannot be retrieved online. Finally, it is worth noting that 3DGe-Aug is entirely orthogonal to classical image-level augmentations and domain randomization techniques (e.g., those built into the YOLO training pipeline). Our framework provides robust synthetic geometric foundations, which can be seamlessly combined with these classical techniques during training to further enhance model generalization.

4.2 Experimental Results

4.2.1 Comparison with Other Data Augmentation Methods

To validate the effectiveness of the proposed method, we first evaluate the detection performance of different augmentation strategies on the VisDrone dataset. The per-category detection accuracy (mAP_{50}) is presented in Table 2. To further analyze generalization across base and novel classes, we compare mAP_{50}^{base}, mAP_{50}^{novel}, and mAP_{50}^{all} in Table 3.

Table 2: Per-category detection accuracy (mAP_{50}) on VisDrone.

Method	Car	Truck	Bus	Humvee	ICV	Mengshi	PLZ-05A	Sprinter	ZTQ-15	All
Baseline	0.879	0.460	0.780	0.000	0.000	0.000	0.000	0.000	0.000	0.235
3DGe-Aug-s	0.857	0.492	0.747	0.657	0.482	0.521	0.386	0.760	0.414	0.591
3DGe-Aug	0.848	0.467	0.728	0.697	0.791	0.345	0.545	0.366	0.578	0.596
X-Paste	0.862	0.510	0.771	0.769	0.866	0.833	0.841	0.757	0.821	0.781
3DGe-Aug-XPa	0.861	0.522	0.774	0.866	0.959	0.853	0.889	0.859	0.869	0.828

Table 3: Detection accuracy (mAP_{50}) on base, novel, and all classes of VisDrone.

Method	Base	Novel	All	95% CI for All
Baseline	0.706	0.000	0.235	/
3DGe-Aug-s	0.698 ± 0.018	0.537 ± 0.017	0.591 ± 0.012	[0.579, 0.602]
3DGe-Aug	0.681 ± 0.007	0.553 ± 0.026	0.596 ± 0.017	[0.579, 0.613]

(Continued)

Table 3 (continued)

Method	Base	Novel	All	95% CI for All
X-Paste	0.714 ± 0.009	0.815 ± 0.011	0.781 ± 0.007	[0.774, 0.788]
3DGe-Aug-XPa	0.719 ± 0.008	0.883 ± 0.014	0.828 ± 0.009	[0.819, 0.837]

To ensure statistical reliability and assess training stability, all reported results in Table 3 are presented as the mean ± standard deviation along with 95% confidence intervals (CI). The results represent the average values across different augmentation-density configurations, derived from YOLOv26l models trained under different numbers of pasted objects ($\{1, 2, 3, 4, 5, 6, 8, 10, 15, 20\}$) per training image. Note that the Baseline method does not involve any object pasting and was evaluated using a fixed random seed. Therefore, it achieves deterministic results without variance in our experiments, and we acknowledge that its underlying training uncertainty was not measured. Furthermore, to validate the robustness and stability of our proposed method, we conducted statistical significance testing on the detection performance (mAP_{50}^{all}). A one-sample t-test against the Baseline method (constant 0.235) yielded $t = 145.51$, $p < 0.0001$. Additionally, since 3DGe-Aug-XPa and X-Paste were evaluated under corresponding object pasting configurations, we adopted a paired samples t-test for their comparison. The test demonstrated a significant performance improvement over X-Paste ($t = 13.96$, $p < 0.0001$). These statistical results indicate that the substantial performance gains achieved by 3DGe-Aug-XPa are stable and reliable across varying augmentation densities.

Data Efficiency Under Few-Shot Conditions. As shown in Table 3, X-Paste achieves $mAP_{50}^{novel} = 0.815$ using an average of 46.5 real images per novel class. In contrast, our single-view method (3DGe-Aug-s) achieves $mAP_{50}^{novel} = 0.537$ using only a single real image for 3D reconstruction—approximately 2.2% of X-Paste's data while attaining 65.9% of its performance. This demonstrates the substantial potential of 3D generative models for addressing few-shot feature scarcity. With multi-view joint reconstruction (3DGe-Aug), 3D asset fidelity is further enhanced, raising mAP_{50}^{novel} to 0.553 (a 3.1% relative improvement).

Synergy Between 3D and 2D Augmentation. Although 3D-based augmentation is data-efficient, a performance gap remains relative to X-Paste, which utilizes real web-retrieved images. However, we observe that the geometrically consistent images rendered from 3D assets complement the textural diversity of real images from X-Paste. As shown in Tables 2 and 3, the joint strategy (3DGe-Aug-XPa) achieves superior performance, increasing mAP_{50}^{base} by 0.7% to 0.719 and mAP_{50}^{novel} by 8.3% to 0.883 compared to X-Paste alone. This indicates that 3DGe-Aug-XPa successfully integrates multi-view geometric consistency with real 2D textures, lifting mAP_{50}^{all} to 0.828.

Qualitative detection results obtained by 3DGe-Aug-XPa are illustrated in Fig. 5, featuring VisDrone dataset samples on the left and few-shot-object-aerial dataset samples on the right.

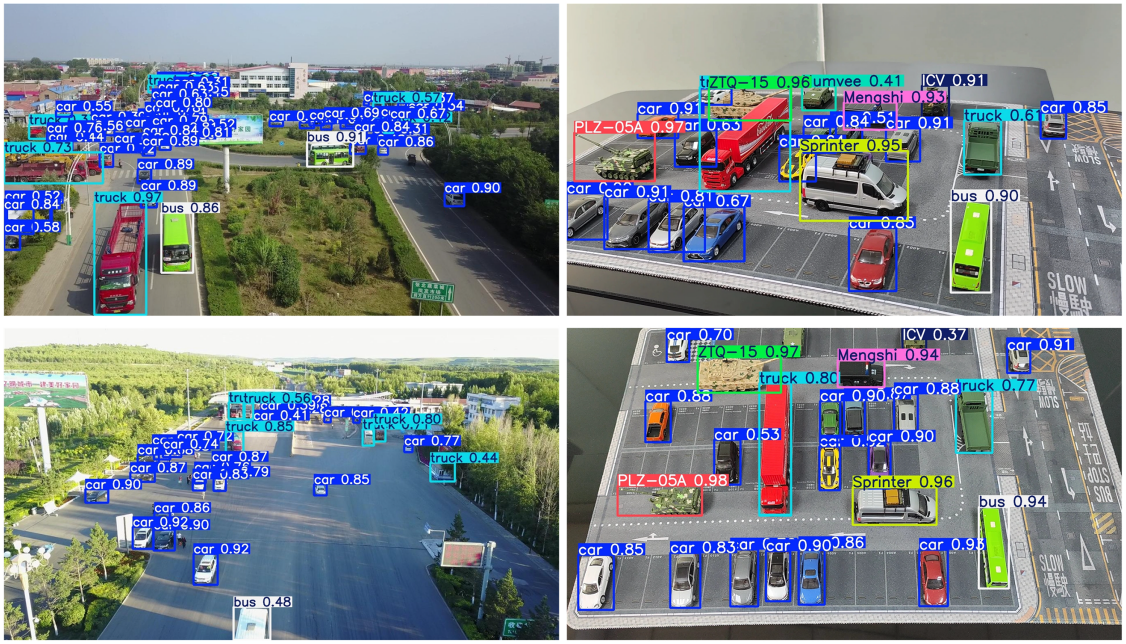


Figure 5: Detection results using 3DGe-Aug-XPa on VisDrone dataset (**left**) and few-shot-object-aerial dataset (**right**).

4.2.2 Generalization across Different Datasets

To validate generalization across different datasets and scenarios, we evaluate our approach on the BDD100K autonomous driving dataset, which features substantially different camera viewpoints. Per-category detection accuracy (mAP_{50}) is reported in Table 4, whereas base, novel, and overall results are summarized in Table 5. To maintain statistical rigor, all overall results in Table 5 are presented as the mean ± standard deviation along with 95% confidence intervals (CI). The results represent the average values across different augmentation-density configurations, derived from YOLOv26l models trained under different numbers of pasted objects($\{1, 2, 3, 4, 5\}$) per training image. Similar to the VisDrone setup, the Baseline model does not utilize object pasting and employs a fixed random seed, resulting in a deterministic performance without variance.

Table 4: Per-category detection accuracy (mAP_{50}) on BDD100K.

Method	Car	Truck	Bus	Humvee	ICV	Mengshi	PLZ-05A	Sprinter	ZTQ-15	All
Baseline	0.810	0.639	0.637	0.000	0.000	0.000	0.000	0.000	0.000	0.232
3DGe-Aug	0.794	0.620	0.606	0.207	0.444	0.106	0.251	0.224	0.356	0.401
X-Paste	0.790	0.613	0.591	0.403	0.557	0.305	0.465	0.546	0.425	0.522
3DGe-Aug-XPa	0.802	0.639	0.624	0.472	0.745	0.402	0.701	0.716	0.622	0.636

Table 5: Detection accuracy (mAP_{50}) on base, novel, and all classes of BDD100K.

Method	Base	Novel	All	95% CI for All
Baseline	0.695	0.000	0.232	/
3DGe-Aug	0.673 ± 0.017	0.265 ± 0.052	0.401 ± 0.030	[0.371, 0.431]
X-Paste	0.664 ± 0.011	0.450 ± 0.138	0.522 ± 0.090	[0.432, 0.611]
3DGe-Aug-XPa	0.688 ± 0.013	0.610 ± 0.073	0.636 ± 0.046	[0.590, 0.682]

Domain Shift on Large-Scale Datasets. Using 3DGe-Aug alone yields mAP_{50}^{novel} = 0.265 on BDD100K. This limitation fundamentally stems from the visual discrepancy between synthetic multi-view images and real-world scenes, as well as the dataset scale (nearly 70,000 training images), where frequent pasting of synthetic 3D objects amplifies the domain gap between synthetic training data and real test images. Consequently, the detector tends to overfit to the synthetic 3D features, thereby degrading its generalization accuracy on real-world test images. Nevertheless, even without any real novel objects, 3DGe-Aug achieves 58.8% of X-Paste's performance (0.450 vs. 0.265), demonstrating that it still provides substantial baseline detection capabilities for rare objects even in the complete absence of real-world novel samples. In contrast, X-Paste uses real images with minimal domain shift, making it less susceptible to this issue.

Complementarity and Base-Class Performance Retention. Despite the distribution shift limitations of pure generative methods, 3DGe-Aug and X-Paste exhibit strong feature complementarity. To compensate for the synthetic bias, X-Paste contributes essential real-world texture details, while 3D assets provide rich viewpoint and scale diversity. Their combination (3DGe-Aug-XPa) substantially improves detection performance, achieving mAP_{50}^{novel} = 0.610 and mAP_{50}^{all} = 0.636. Moreover, 3DGe-Aug-XPa effectively mitigates base-class accuracy degradation during joint training. As shown in Table 5, X-Paste alone causes base class accuracy to drop to 0.664, whereas 3DGe-Aug-XPa restores it to 0.688, approaching the Baseline (0.695) while simultaneously improving novel class accuracy. While this combined strategy alleviates the domain shift, it does not fundamentally resolve the rendering gap. Fundamentally bridging this gap by optimizing data synthesis strategies—such as incorporating physical environment priors (e.g., dynamic lighting conditions, diverse weather, and sensor noise) to enhance rendering realism—represents a critical direction for our future research.

Qualitative detection results produced by 3DGe-Aug-XPa are illustrated in Fig. 6, showcasing BDD100K dataset samples on the left and few-shot-object-ground dataset samples on the right.

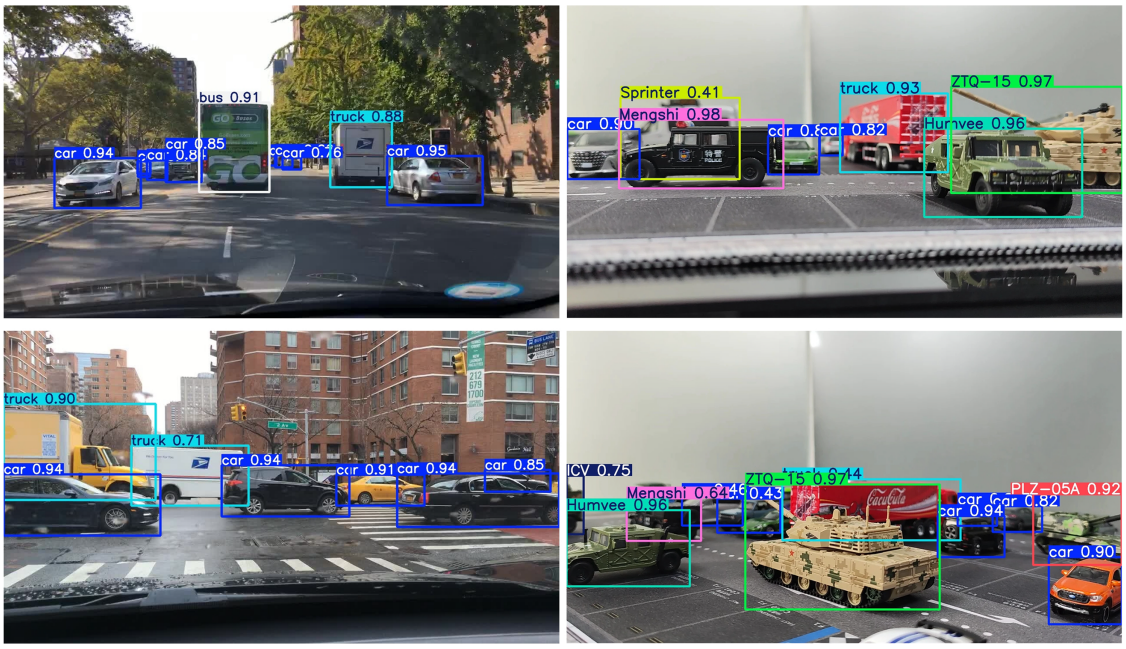


Figure 6: Detection results using 3DGe-Aug-XPa on BDD100K dataset (**left**) and few-shot-object-ground dataset (**right**).

4.2.3 Generalization across Detection Architectures

To validate the universality of 3DGe-Aug across different architectures, we evaluate various detectors on VisDrone, including CNN-based models (YOLOv5, YOLOv8, YOLOv11, YOLOv26) and the Transformer-based RT-DETR. In these experiments, the number of pasted objects per image is fixed to 3. To comprehensively evaluate the detection performance, we extend our evaluation to include mAP_{50-95}, Precision, Recall, and F1-score. Detailed quantitative results are summarized in Table 6.

Table 6: Performance across different detection architectures.

Model	mAP_{50}^{base}	mAP_{50}^{novel}	mAP_{50}^{all}	mAP_{50-95}^{all}	Precision	Recall	F1
RT-DETR-l	0.649	0.392	0.478	0.387	0.745	0.470	0.576
YOLOv5l	0.572	0.635	0.614	0.475	0.648	0.559	0.600
YOLOv8l	0.692	0.702	0.698	0.563	0.744	0.600	0.664
YOLOv11l	0.671	0.698	0.689	0.547	0.753	0.593	0.663
YOLOv26x	0.722	0.515	0.584	0.472	0.641	0.474	0.545
YOLOv26l	0.710	0.546	0.601	0.473	0.637	0.488	0.553
YOLOv26m	0.685	0.387	0.486	0.381	0.602	0.412	0.489
YOLOv26n	0.533	0.364	0.420	0.308	0.526	0.389	0.447

Universality across Architectures. Data generated by 3DGe-Aug is effectively leveraged across diverse detector architectures. Both CNN-based models (characterized by local receptive fields) and RT-DETR (leveraging global self-attention) learn discriminative features from augmented samples. Notably, YOLOv8l achieves the best overall performance with $mAP_{50}^{novel} = 0.702$ and $mAP_{50}^{all} = 0.698$, alongside a balanced Precision of 0.744 and F1-score of 0.664, demonstrating strong few-shot adaptability.

Confusion Matrix Analysis. To further investigate whether the performance gains come from improved class recognition or merely easier localization of pasted objects, Fig. 7 presents the normalized confusion matrices. Taking the YOLOv8l model as a representative example, the diagonal elements reveal high true positive classification rates for most novel classes (e.g., 0.83 for Humvee, 0.87 for ICV). Meanwhile, instances of certain specialized categories like PLZ-05A and Sprinter exhibit some confusion with visually similar vehicles (such as trucks) or the background. This confirms that the augmented data effectively enhances the semantic recognition capability of the model rather than just creating trivial localization shortcuts, although distinguishing highly similar fine-grained categories remains a challenge in extreme few-shot scenarios.

Impact of Model Scale on Few-Shot Learning. To investigate how model capacity affects the utilization of 3D-augmented data, we compare YOLOv26 variants (YOLOv26n, YOLOv26m, YOLOv26l, and YOLOv26x). **Base classes scale monotonically.** As model parameters increase from YOLOv26n to YOLOv26x, mAP_{50}^{base} rises steadily from 0.533 to 0.722, consistent with the conventional wisdom that larger models better fit abundant data. **Novel classes exhibit overfitting.** In contrast, mAP_{50}^{novel} improves from 0.364 (YOLOv26n) to 0.546 (YOLOv26l) before dropping to 0.515 (YOLOv26x). This reveals a typical overfitting phenomenon in few-shot learning: ultra-large models overfit underrepresented novel classes, whereas medium-to-large-sized models (YOLOv8l, YOLOv11l, YOLOv26l) achieve a better balance between learning and generalization. In summary, 3DGe-Aug consistently improves detection across architectures. However, moderate-capacity models best leverage 3D-augmented data in few-shot settings.

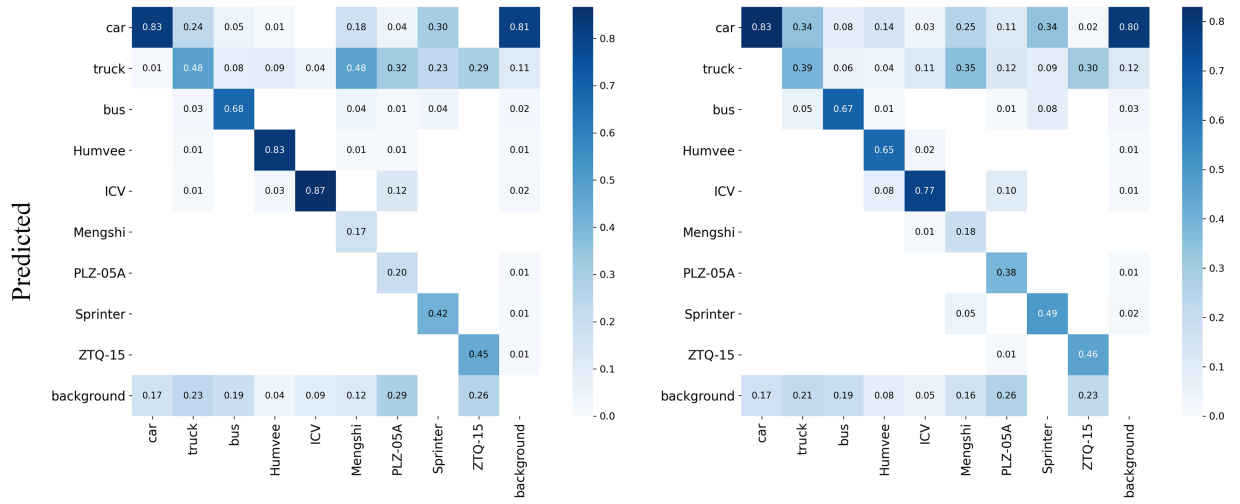


Figure 7: Normalized confusion matrixs of YOLOv8l (**left**) and YOLOv11l (**right**).

Performance on Separated Test Sets. To ensure a transparent evaluation, we separately report the detection performance on the original VisDrone validation set and the constructed few-shot-object-aerial test set in Table 7 (merged results are detailed in Table 6). As shown in Table 7, the models maintain robust base-class accuracy on the original validation images—which contain no novel objects—demonstrating that the augmentation strategy does not trigger severe false-positive degradation. Furthermore, on the constructed test set, the models exhibit strong recognition capabilities for both base and novel classes. For example, YOLOv8l achieves $mAP_{50}^{base} = 0.822$ and $mAP_{50}^{novel} = 0.716$. These disaggregated results confirm that the proposed 3DGe-Aug framework enhances the detector's capacity to recognize rare instances while preserving its stability on original instances.

Table 7: Performance on VisDrone validation set and few-shot-object-aerial test set.

Model	VisDrone Validation Set				Few-Shot-Object-Aerial Test Set			
	mAP_{50}^{base}	mAP_{50}^{novel}	mAP_{50}^{all}	mAP_{50-95}^{all}	mAP_{50}^{base}	mAP_{50}^{novel}	mAP_{50}^{all}	mAP_{50-95}^{all}
RT-DETR-l	0.616	/	0.616	0.438	0.729	0.412	0.518	0.440
YOLOv5l	0.532	/	0.532	0.368	0.662	0.646	0.651	0.525
YOLOv8l	0.592	/	0.592	0.418	0.822	0.716	0.751	0.630
YOLOv11l	0.589	/	0.589	0.413	0.786	0.707	0.733	0.608
YOLOv26x	0.664	/	0.664	0.481	0.804	0.529	0.620	0.527
YOLOv26l	0.627	/	0.627	0.448	0.816	0.564	0.648	0.539
YOLOv26m	0.605	/	0.605	0.428	0.806	0.396	0.533	0.443
YOLOv26n	0.440	/	0.440	0.284	0.671	0.489	0.550	0.432

4.2.4 Adaptability to Pasting Quantity

To evaluate how different augmentation methods respond to varying numbers of pasted objects, we evaluate pasting quantities ranging from 1 to 20 on the VisDrone dataset. The corresponding performance curves are illustrated in Figs. 8–10. The experimental setup adopts random placement, an area ratio within $[0.005, 0.05]$, an occlusion threshold of 0.7, and YOLOv26l as the primary detector.

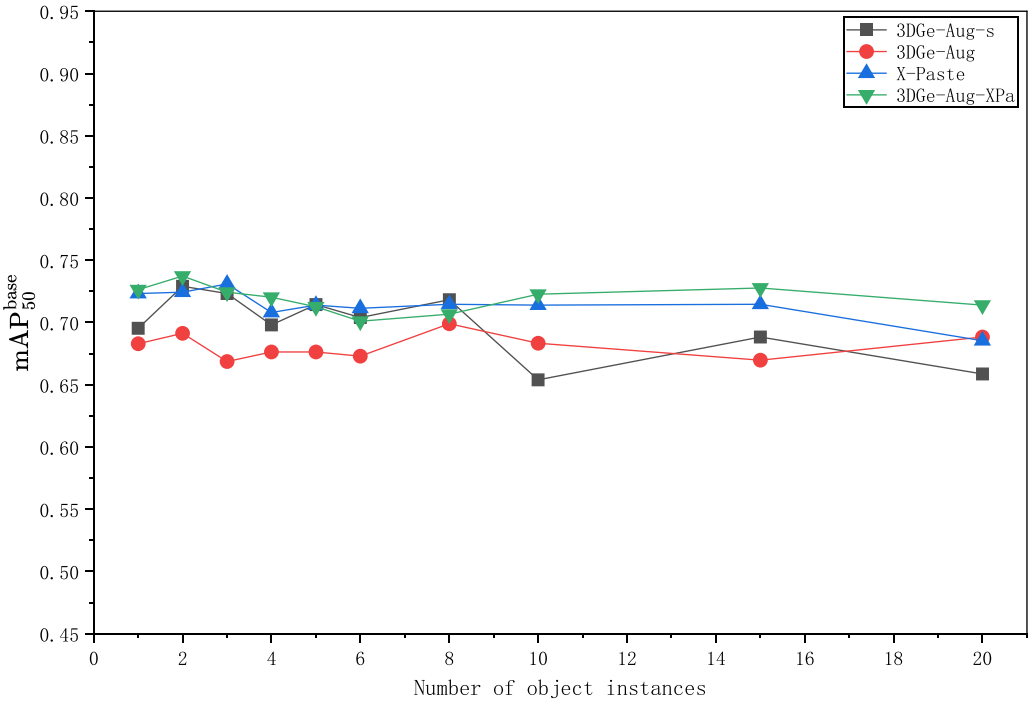


Figure 8: Base class accuracy vs. pasting quantity.

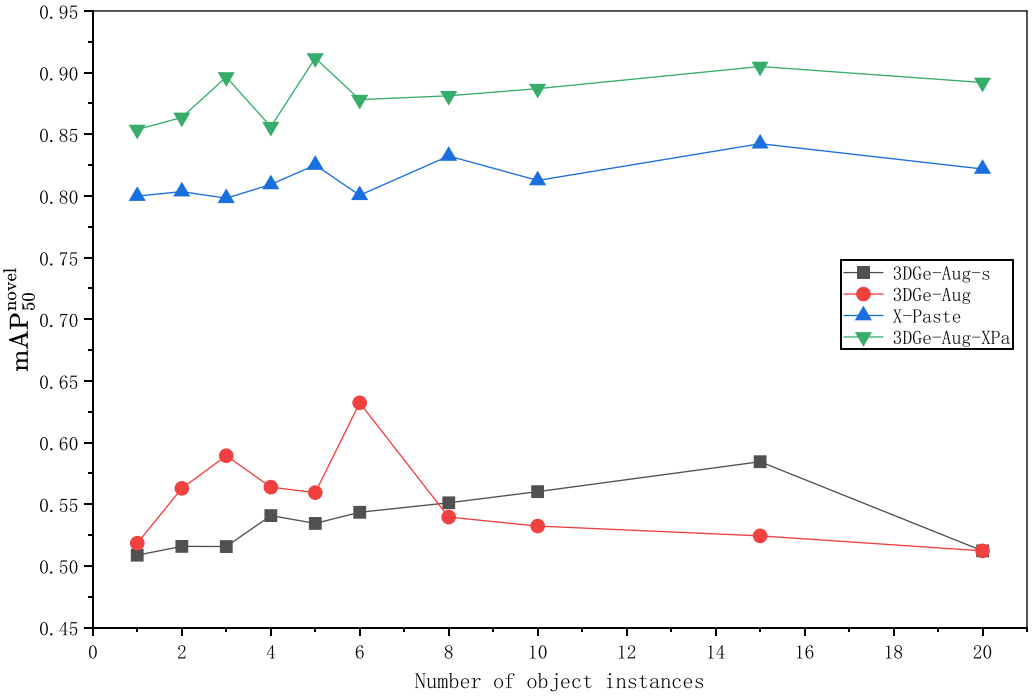


Figure 9: Novel class accuracy vs. pasting quantity.

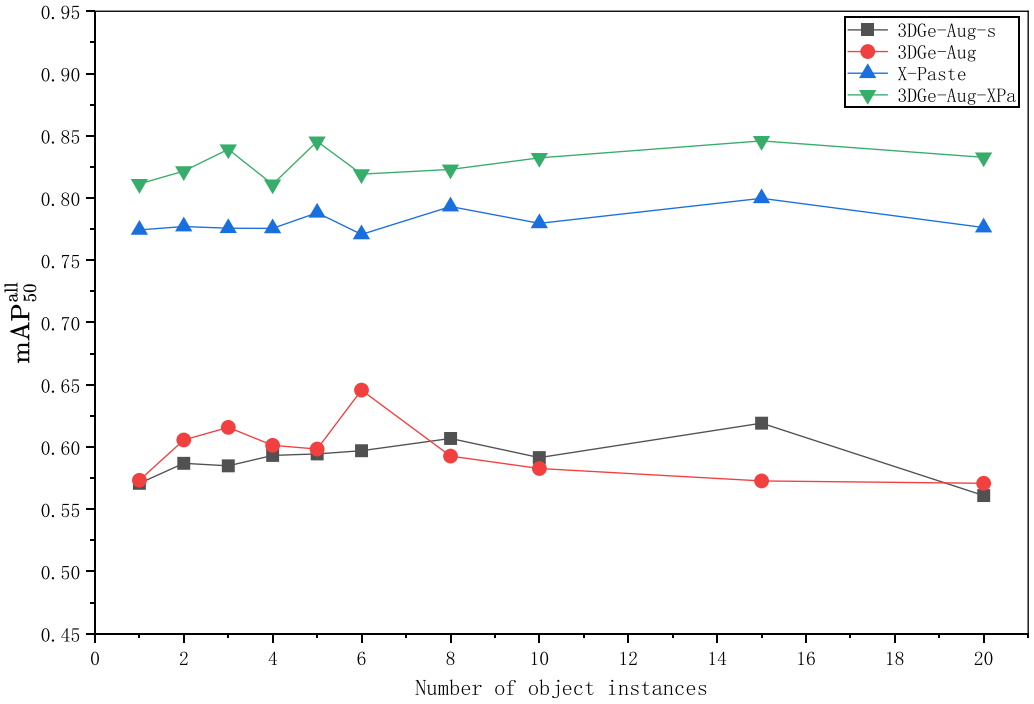


Figure 10: Overall accuracy vs. pasting quantity.

Quantity Adaptability and Training Stability. All evaluated methods exhibit robust adaptability across pasting quantities ranging from 1 to 20, achieving satisfactory detection accuracy. However, training stability varies: pure generative variants (3DGe-Aug and 3DGe-Aug-s) exhibit higher performance variance across different quantities, reflecting weaker stability. In contrast, methods incorporating real visual textures (X-Paste and 3DGe-Aug-XPa) maintain minimal fluctuations, demonstrating training stability alongside consistently high accuracy.

Base-Novel Trade-Off. Randomly pasting multiple novel objects into constrained image spaces inevitably causes partial occlusion of pre-existing base class instances, triggering dynamic label removal when the overlap threshold is exceeded. Consequently, base and novel class accuracies demonstrate opposing trade-off trends: mAP^{novel}_{50} increases with higher pasting density before saturating or slightly overfitting, whereas mAP^{base}_{50} gradually degrades owing to increased background clutter and occlusion. This phenomenon underscores the necessity of controlling pasting density to strike an optimal balance between novel class learning and base class retention.

4.2.5 Sampling Mode and Quantity

To evaluate the impact of multi-view sampling strategies, we compare Grid and Random sampling modes under different step sizes (ϕ_{step}, θ_{step}) on the VisDrone dataset. The sampling quantity denotes the total number of multi-view images generated under each parameter setting. We fix the number of pasted objects per image to 5 and employ YOLOv26l as the default detector. Quantitative results are presented in Table 8.

Table 8: Impact of sampling modes and quantities.

$\phi_{step}, \theta_{step}$	Number of Views	Grid Sampling			Random Sampling		
		Base	Novel	All	Base	Novel	All
5°, 5°	936	0.676	0.559	0.598	0.713	0.534	0.593
10°, 5°	468	0.670	0.607	0.628	0.705	0.528	0.587
10°, 10°	252	0.703	0.646	0.665	0.701	0.423	0.516
15°, 15°	120	0.683	0.638	0.653	0.680	0.530	0.580
20°, 20°	72	0.729	0.636	0.667	0.678	0.474	0.542
25°, 25°	45	0.711	0.598	0.636	0.634	0.575	0.595
30°, 30°	36	0.688	0.470	0.543	0.517	0.546	0.536
35°, 35°	22	0.715	0.554	0.608	0.642	0.503	0.549

Grid vs. Random Sampling. Both sampling modes demonstrate strong adaptability across a wide range of rendered view quantities (ranging from dozens to nearly one thousand). However, Grid Sampling consistently outperforms Random Sampling, achieving a higher average mAP_{50}^{novel} of 0.588 compared to 0.514. This performance disparity stems from the fact that grid sampling guarantees structured, uniform coverage of the virtual camera across the 3D spherical space, whereas random sampling tends to suffer from spatial clustering and unobserved blind spots, thereby constraining novel object feature representation.

Inverted U-Shaped Effect of Sampling Density. mAP_{50}^{novel} exhibits a clear inverted U-shaped relationship with sampling density, rather than a monotonic linear correlation. **Overly dense sampling causes feature redundancy.** With fine angular steps ($\phi_{step} = 5°, \theta_{step} = 5°$), the number of rendered views surges to 936, while mAP_{50}^{novel} drops to 0.559. Excessively fine camera increments introduce near-identical visual projections, causing feature redundancy and harming model generalization. **Overly sparse sampling causes feature loss.** Conversely, with coarse step sizes ($\phi_{step} = 30°, \theta_{step} = 30°$), only 36 views are generated, and mAP_{50}^{novel} falls to 0.470. An insufficient number of viewpoints fails to cover essential geometric characteristics, leading to degraded recognition. Moderate step sizes (10°–20°) strike an optimal balance between feature diversity and rendering efficiency. Within this regime, mAP_{50}^{base} remains robust (0.703–0.729), while mAP_{50}^{novel} and mAP_{50}^{all} reach their peaks. This observation provides practical guidelines for maximizing data augmentation effectiveness while minimizing computational and rendering overhead.

5 Conclusion

This paper proposes 3DGe-Aug, a 3D generative model-based data augmentation framework designed for few-shot object detection. To address the limitations of text-to-image models on fine-grained and emerging objects, 3DGe-Aug first reconstructs 3D assets with detailed geometry and Physically Based Rendering (PBR) textures from sparse 2D images. A virtual camera system operating in spherical coordinates then renders multi-view, geometrically consistent 2D images, while salient object segmentation extracts background-free masks. Finally, an instance pasting algorithm incorporating multi-scale transformations and dynamic occlusion elimination synthesizes training data by integrating these instances onto real backgrounds, expanding object quantity and feature diversity while maintaining label reliability. Comprehensive experiments conducted across two scenarios (VisDrone and BDD100K) and multiple detector architectures (RT-DETR and the YOLO series) demonstrate that: (1) 3DGe-Aug constructs robust feature spaces with high data efficiency, yielding substantial gains even from single-image reconstruction; (2) 3DGe-Aug exhibits

strong synergy with 2D augmentation (X-Paste), with their combination boosting novel class accuracy while mitigating base-class degradation and preserving overall detection performance.

Beyond the UAV and autonomous driving scenarios evaluated in this work, 3DGe-Aug exhibits broad applicability to other real-world domains constrained by severe data scarcity. Specifically, in **smart retail and unmanned commerce**, 3DGe-Aug can rapidly construct 3D visual assets for frequently updated or non-standard fresh produce, enabling automated checkout, smart weighing, and vending item recognition under rapid iterations. In **smart security and public safety**, it facilitates multi-perspective data synthesis for rare or prohibited objects in surveillance systems, where real-world training images are restricted or difficult to acquire.

Despite these advantages, practical insights reveal potential scalability challenges for large-scale deployment in real-world production settings: **Computational and Rendering Bottlenecks:** Constructing high-resolution 3D assets and batch-rendering massive multi-view image sets demand substantial GPU compute and memory resources. Scaling the framework to thousands of candidate object classes or integrating it into continuous pipelines may incur non-negligible processing latency. **Storage and I/O Overhead:** Generating multi-view images across high grid densities creates large volumes of synthetic visual data, placing heavy demands on dataset storage capacity and disk I/O bandwidth during large-scale model training. **Automated Domain Alignment at Scale:** As dataset scale increases, manual tuning of physical rendering parameters becomes impractical. Unaligned illumination, shadows, and sensor noise between synthetic instances and diverse background environments can amplify domain shift if not automatically harmonized. To address these scalability limitations, future work will focus on: (1) developing lightweight 3D asset representation and acceleration techniques to optimize generation and rendering throughput; (2) lightweight domain alignment and style transfer to reduce the synthetic-real gap; (3) incorporating physical environment priors (lighting, weather, shadows) during rendering to enhance realism and robustness in a fully automated, low-cost manner.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by the National Cyber Security-National Science and Technology Major Project of China, grant number 2025ZD1501300.

Author Contributions: The authors confirm contribution to the paper as follows: methodology, Jichao Xie, Xinlei Liu, Zhen Zhang and Peng Yi; software, Jichao Xie, Xinlei Liu and Tong Duan; validation, Tong Duan and Baolin Li; investigation, Jichao Xie and Xinlei Liu; data curation, Baolin Li; writing—original draft preparation, Jichao Xie, Xinlei Liu and Tong Duan; writing—review and editing, Zhen Zhang and Peng Yi; visualization, Baolin Li; funding acquisition, Zhen Zhang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The source code, 3D asset collection, and few-shot object detection benchmark datasets supporting the findings of this study are publicly available on GitHub at https://github.com/xiejichao/DataAug. The repository code is released under the MIT License, and the datasets/3D assets are distributed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0), strictly restricting their use to non-commercial academic research and educational purposes.

Ethics Approval: Not applicable for human or animal subjects. All novel vehicle categories (e.g., Humvee, ICV, Mengshi, PLZ-05A, ZTQ-15) were constructed from commercially available toy models and physical scale models collected from public domain resources. The dataset consists purely of exterior 2D visual representations and synthetic 3D surface meshes of toy models. No real military hardware, classified technical data, structural blueprints, or defense-restricted parameters were involved.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zou Z, Chen K, Shi Z, Guo Y, Ye J. Object detection in 20 years: a survey. Proc IEEE. 2023;111(3):257–76. doi:10.1109/jproc.2023.3238524.
2. Li Z, Dong Y, Shen L, Liu Y, Pei Y, Yang H, et al. Development and challenges of object detection: a survey. Neurocomputing. 2024;598(4):128102. doi:10.1016/j.neucom.2024.128102.
3. Köhler M, Eisenbach M, Gross HM. Few-shot object detection: a comprehensive survey. IEEE Trans Neural Netw Learn Syst. 2023;35(9):11958–78.
4. Wang X, Huang TE, Darrell T, Gonzalez JE, Yu F. Frustratingly simple few-shot object detection. In: Proceedings of the 37th International Conference on Machine Learning; 2020 Jul 13–18; Vienna, Austria. p. 9919–28.
5. Zhao H, Sheng D, Bao J, Chen D, Chen D, Wen F, et al. X-Paste: revisiting scalable copy-paste for instance segmentation using clip and stablediffusion. In: Proceedings of the 40th International Conference on Machine Learning; 2023 Jul 23–29; Honolulu, HI, USA. p. 42098–109.
6. Ghiasi G, Cui Y, Srinivas A, Qian R, Lin TY, Cubuk ED, et al. Simple copy-paste is a strong data augmentation method for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 2918–28.
7. Fang H, Han B, Zhang S, Zhou S, Hu C, Ye WM. Data augmentation for object detection via controllable diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2024 Jan 3–8; Waikoloa, HI, USA. p. 1257–66.
8. Jiang Y, Qiang S, Li W, Liang Y. LLM-DiffAug: enhancing few-shot object detection via LLM-guided diffusion augmentation. Knowl Based Syst. 2025;326:114066.
9. Nguyen Vu AK, Truong QT, Nguyen VT, Ngo T, Do TT, Nguyen T. Multi-perspective data augmentation for few-shot object detection. In: Proceedings of the International Conference on Learning Representations; 2025 Apr 24–28; Singapore. p. 65620–38.
10. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 10684–95.
11. Hunyuan3D T, Yang S, Yang M, Feng Y, Huang X, Zhang S, et al. Hunyuan3D 2.1: from images to high-fidelity 3D assets with production-ready PBR material. arXiv:2506.15442. 2025.
12. Li Y, Zou ZX, Liu Z, Wang D, Liang Y, Yu Z, et al. TripoSG: high-fidelity 3D shape synthesis using large-scale rectified flow models. IEEE Trans Pattern Anal Mach Intell. 2025;1–18.
13. Wu S, Lin Y, Zhang F, Zeng Y, Yang Y, Bao Y, et al. Direct3D-S2: gigascale 3D generation made easy with spatial sparse attention. In: Proceedings of the Advances in Neural Information Processing Systems; 2025 Dec 2–7; San Diego, CA, USA. p. 170778–04.
14. Song Y, Wang T, Cai P, Mondal SK, Sahoo JP. A comprehensive survey of few-shot learning: evolution, applications, challenges, and opportunities. ACM Comput Surv. 2023;55(13s):1–40.
15. Shi Y, Shi D, Qiao Z, Zhang Y, Liu Y, Yang S. A survey on recent advances in few-shot object detection. Chin J Comput. 2023;46(8):1753–80. doi:10.1016/j.neunet.2023.10.039.
16. Yan X, Chen Z, Xu A, Wang X, Liang X, Lin L. Meta R-CNN: towards general solver for instance-level low-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 19–23; Honolulu, HI, USA. p. 9577–86.
17. Zeng K, Wu Z, Xiong K, Wei X, Guo X, Zhu Z, et al. Rethinking driving world model as synthetic data generator for perception tasks. In: Proceedings of the Fourteenth International Conference on Learning Representations; 2026 Apr 23–27; Rio de Janeiro, Brazil. p. 1–26.
18. Yun YK, Lin W. Towards a complete and detail-preserved salient object detection. IEEE Trans Multimed. 2023;26:4667–80. doi:10.1109/tmm.2023.3325731.
19. Zhu P, Wen L, Du D, Bian X, Fan H, Hu Q, et al. Detection and tracking meet drones challenge. IEEE Trans Pattern Anal Mach Intell. 2021;44(11):7380–99. doi:10.1109/tpami.2021.3119563.

20. Seita D. BDD100K: a large-scale diverse driving video database. Berkeley Artif Intell Res Blog Version. 2018;511(41):11.
21. Jocher G, Qiu J, Liu M, Lyu S, Akyon FC, Kalfaoglu ME. Ultralytics YOLO26: unified real-time end-to-end vision models. arXiv:2606.03748. 2026.