Diagnosing Aerial-View Object Detectors with Foundational Image Generative Models

Stanislav Panev[1], Minhyek Jeon[1], Vaishnavi Khindkar[1], Ahish Deshpande[1], Celso M de Melo[2], Shuowen Hu[2], Shayok Chakraborty[1,3], and Fernando De la Torre[1]

[1] Carnegie Mellon University, USA
spanev@andrew.cmu.edu, minhyekj@alumni.cmu.edu, {vkhindka, ahishd}@andrew.cmu.edu, ftorre@cs.cmu.edu
[2] DEVCOM Army Research Laboratory, USA
{celso.m.demelo.civ, shuowen.hu.civ}@army.mil
[3] Florida State University, USA shayok@cs.fsu.edu

Abstract. Recent advances in large-scale image generative models enable photorealistic scene synthesis with controllable attributes. Beyond data augmentation, their potential as diagnostic tools for trained vision systems remains unexplored in the aerial and remote sensing domains. We introduce a synthetic diagnostic framework for aerial-view vehicle detection that combines text-guided generation, attribute-controlled editing, and automated attribute verification to construct a controllable synthetic testbed. This enables fine-grained evaluation of pretrained detectors under diverse scene types and environmental conditions that are difficult to isolate in real datasets. Across three detection architectures and three real aerial datasets, synthetic scene-wise performance trends closely match real-world weaknesses. Guided by these diagnostics, targeted supplementation with small real datasets from the identified weak categories yields improvements of up to 13% AP50 while requiring substantially fewer additional samples than non-targeted augmentation. Our results show that controlled synthetic probing can predict real-domain performance gaps and guide efficient data collection. The proposed diagnostic framework is modular and can incorporate alternative generative or vision-language models as capabilities evolve. Our code and datasets are available here: humansensinglab.github.io/AVODDiag/

Keywords: Aerial Object Detection · Remote Sensing Robustness · Synthetic Data Evaluation · Model Diagnosis · Applied Generative AI

1 Introduction

Recent advances in foundation image generative models—such as *Imagen 3* [2], *Gemini 2.5 Flash Image* [8], *GPT-Image-1* [29], *DALL·E 3* [28], and *Midjourney* [25]—enable the synthesis of highly realistic imagery with controllable semantic attributes. While these models are widely used for data augmentation,

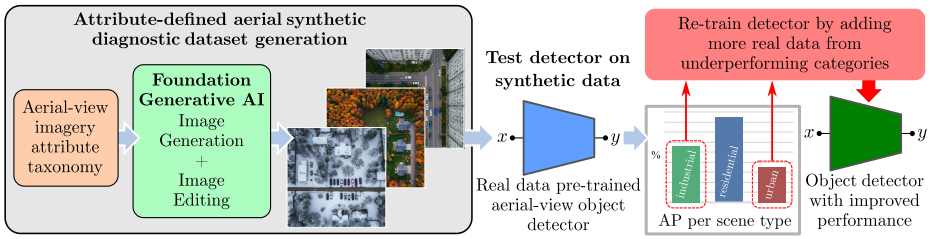

Fig. 1: We leverage foundation generative models to synthesize diverse, attribute-controlled aerial-view image datasets, thereby enabling systematic diagnosis of the weaknesses of aerial-view vehicle detectors.

their potential as diagnostic tools for analyzing trained vision systems remains largely unexplored in remote sensing and aerial imagery. *Aerial-view object detection* (AVOD) is a core component of *Earth observation* (EO) pipelines that support disaster response, infrastructure inspection, defense intelligence, urban planning, and environmental monitoring. In such high-stakes applications, systematic performance degradation under specific scene conditions can directly affect downstream decision-making, motivating reliable and interpretable diagnostic evaluation.

Modern aerial benchmarks such as *DOTAv2* [11], *LINZ* [12], and *UGRC* [12] are inherently imbalanced: certain scene types and environmental conditions are overrepresented, while others are scarce or entangled with confounding factors. As a result, detectors may exhibit systematic performance degradation in specific environments (*e.g.*, dense urban or industrial areas) that is difficult to isolate using standard aggregate metrics. Collecting geographically and environmentally balanced aerial datasets is costly and often infeasible, making controlled diagnostic evaluation a persistent challenge in remote sensing.

In this work, we investigate whether foundation image generative models can be leveraged to construct a controlled synthetic diagnostic testbed for aerial-view vehicle detection (Fig. 1). As an illustrative example, Fig. 2 shows synthetic images generated by *Imagen 3* under different attribute settings. Rather than proposing a new detector or generative model, we introduce a synthetic diagnostic framework that combines: (1) text-guided aerial image generation, (2) attribute-controlled image editing to reduce distributional imbalance, and (3) automated attribute extraction and validation using multimodal models. This process produces a fully synthetic dataset annotated with explicit environmental and object-level attributes, enabling scene-conditioned evaluation of pre-trained detectors under controlled variations of scene type, season, weather, and object properties. We evaluate three widely adopted detection architectures—*Faster R-CNN* [37], *YOLOv8* [9], and *ViTDet* [22]—trained on three real aerial datasets. Synthetic scene-wise performance trends consistently align with real-world weaknesses, with urban and industrial environments emerging as recurrent failure modes across models and training sources. Guided by these diagnostics, we perform targeted real-data supplementation by adding scene-specific images

Fig. 2: A showcase of synthetic aerial view imagery generated by *Imagen 3*. Each image represents a combination of a different environment and a season. We use these data to identify inherent weaknesses in object detectors.

corresponding to the identified weak categories. Despite being at least an order of magnitude smaller than the original training sets, this targeted supplementation yields improvements of up to 13% in AP50.

Our contribution is therefore a validated synthetic-to-real diagnostic framework for aerial-view detection systems. We demonstrate that controlled synthetic probing can predict real-domain performance gaps and provide actionable guidance for efficient data acquisition in remote sensing applications. While our empirical study focuses on RGB aerial imagery and vehicle detection, the proposed diagnostic framework is modular and generalizable, and can integrate alternative generative and multimodal models as capabilities evolve, and can be applied to non-aerial imagery systems as well. In summary, our contributions are:

1. A synthetic, attribute-controlled diagnostic framework for analyzing aerial-view object detectors using foundation generative models.
2. Empirical evidence that synthetic scene-wise evaluation reliably reflects real-world performance weaknesses across multiple detectors and datasets.
3. A targeted data supplementation strategy guided by synthetic diagnostics, achieving up to 13% AP50 improvement with limited additional real data.

2 Related Work

2.1 Model Diagnosis in Object Detection

Traditional model diagnostics rely on confusion matrices and performance metrics, which are insufficient for characterizing systemic failure modes. Recent work

moves beyond coarse summaries toward fine-grained failure analysis. Hoiem *et al.* [16] categorized detection errors into localization mistakes, confusions with similar classes, and background false positives. Building on this, Bolya *et al.* [4] proposed *TIDE*, decomposing detection errors into six types for fine-grained diagnosis, exposing previously hidden failure patterns.

However, these diagnostic approaches remain post-hoc, where multiple environmental factors are entangled and difficult to control. Explainability methods such as attribution maps [44] and local surrogate explanations [38] highlight decision-relevant regions but do not isolate the effect of scene-level attributes on detection performance. Synthetic datasets such as *InteriorNet* [21] provide large-scale rendered environments yet lack sufficient attribute-level controllability, while corruption benchmarks [15] reveal sensitivity patterns under distribution shift without enabling causal attribute manipulation of scene conditions. In contrast, our method leverages controllable synthetic generation to isolate individual scene attributes while keeping others fixed, enabling causal and interpretable failure analysis infeasible with real-world data.

2.2 Synthetic Data for Benchmarking

Using synthetic imagery to evaluate and improve vision models has a long history. Early work used game engines for semantic segmentation [39,41] and object tracking [14], but suffered domain gaps due to texture artifacts and unrealistic lighting. Müller *et al.* [26] improved photorealism with a physics-based *Unreal Engine* simulator, yet its fixed scene configurations remain impractical for large-scale, attribute-controlled diagnostic generation. In the *autonomous driving* domain, generative models have recently been used to synthesize realistic driving scenarios for evaluation and failure mode discovery [17, 27]. However, they are tailored to the temporal, multi-modal structure of driving data (video, LiDAR, RADAR) and do not transfer directly to aerial-view detection, which operates on single frames with unique viewpoint and scale characteristics.

Diffusion models have enabled scalable and photorealistic synthesis. For instance, *DatasetDM* [50] generates labeled datasets but targets augmentation rather than diagnosis, while zero-shot evaluation approaches [7, 20] using *Stable Diffusion* [40] lack fine-grained attribute control. Methods combining diffusion with *NeRF*s [34] or *3D Gaussian Splatting* [19] achieve high-fidelity, 3D-consistent synthesis; however, because they represent appearance as an entangled function of 3D geometry, lighting, and texture, they do not support the manipulation of individual 2D appearance attributes that is necessary for controlled diagnostic analysis. More recently, large multimodal models with integrated generation such as *ChatGPT* with *DALL·E 3* [3] and *Gemini*'s *Imagen 3* [2] enable high-fidelity text-to-image generation with strong prompt adherence. Despite these advances, no prior work in aerial-view detection has leveraged such foundation models to construct controlled diagnostic benchmarks isolating failure modes. We introduce a framework that systematically employs *Imagen 3*'s generative capabilities for interpretable diagnostic analysis and targeted data supplementation for aerial detectors.

2.3 Benchmarking Limitations in Aerial Vision

Object detection in aerial and satellite imagery remains challenging due to large-scale variation, arbitrary object orientations, and complex backgrounds [11]. Consequently, even modern detectors (*e.g.*, *YOLOv8* [18], *DINO* [51], and *ViTDet* [22]) still exhibit performance limitations on aerial benchmarks such as *DOTAv2* [11], and their accuracy degrades under geographic domain shift [6,12]. However, existing benchmarks cannot systematically diagnose these failures: *DOTAv2* suffers from geographic and temporal concentration biases with strong statistical confounds; *xView3* [31] provides SAR imagery but lacks correlated RGB data; and *RarePlanes* [45] includes synthetic imagery via AI.Reverie's platform but introduces a simulation-to-real domain gap.

Domain adaptation methods have attempted to mitigate cross-domain discrepancies – Fang *et al.* [12] addresses geographic shifts via weak supervision, *DA Faster* [6] leverages domain discriminators, *SWDA* [43] employs strong–weak alignment, and *UMT* [10] adopts a mean-teacher strategy—but these treat domain shift as a monolithic phenomenon without decomposing it into underlying causal factors. This indicates that adaptation alone is insufficient; advancing robust aerial detection requires a diagnostic perspective that disentangles and quantifies the impact of individual causal factors.

3 Method

This section presents our methodology for diagnosing the weaknesses of pretrained object detectors using synthetic aerial-view datasets generated by foundation generative models (Sec. 3.1). The process begins with the definition of an attribute taxonomy that guides data generation and ensures semantic coverage and visual diversity (Sec. 3.1). Synthetic data creation proceeds in two stages: (1) initial image generation (Sec. 3.1) and (2) dataset enrichment through image editing (Sec. 3.1). The resulting dataset is then annotated and used for model diagnosis based on attribute-conditioned evaluation (Sec. 3.2).

3.1 Diagnostic Dataset Generation

The diagnostic dataset generation is a two-stage process. First, an initial set of synthetic images $\mathcal{I}_{\mathrm{G}}$ is produced using a *text-to-image* approach. To enhance diversity and achieve a more balanced distribution of visual attributes, a second stage performs *image-and-text-to-image* generation, in which the initial images are edited based on textual prompts. This yields a complementary set of edited images $\mathcal{I}_{\mathrm{E}}$. The final diagnostic dataset, $\mathcal{I} = \mathcal{I}_{\mathrm{G}} \cup \mathcal{I}_{\mathrm{E}}$, is fully synthetic and annotated, providing improved coverage of attribute variations across scenes. It is important to note that the presented generation framework is fundamentally model-agnostic; thus, open- or closed-source models can be incorporated, provided they produce data of sufficient quality and realism.

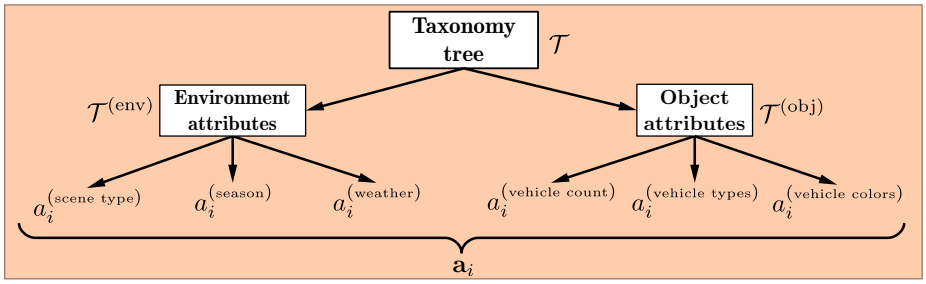

Fig. 3: Attribute taxonomy tree and associated variables $a_i^{(c)}$ that constitute the attribute vector $\mathbf{a}_i$.

Attribute Taxonomy The attribute taxonomy $\mathcal{T}$ (Fig. 3) defines the semantic structure underlying data generation. We deliberately designed this taxonomy to encompass the most critical axes of domain shift and dataset bias known to degrade aerial vision systems. It comprises three primary (environmental) attributes—*scene type*, *season*, and *weather*—and three secondary (object-level) attributes—*vehicle count*, *vehicle color*, and *vehicle type*.

The primary attributes globally determine the image content, enabling testing of macro-level vulnerabilities such as changes in background texture, illumination, and seasonal clutter that are severely imbalanced in standard benchmarks (*e.g.*, the overrepresentation of sunny, urban scenes in DOTAv2). The purpose of the secondary attributes is to ensure object diversity. Potentially, they could also enable fine-grained analysis of local detection challenges and testing the detector's limits with respect to foreground-background contrast, scale variance, and dense object crowding. All attributes are integrated into the diagnostic pipeline described in Sec. 3.2. While we use this specific taxonomy to validate our framework and address known aerial detection bottlenecks, the proposed framework is generic and can be adapted to any taxonomy suitable for a given application.

Generating the Initial Image Set Figure 4-*left* illustrates the procedure for generating the initial image set. First, N attribute tuples $\mathbf{a}_i = \{a_i^{(c)}\}_{c\in\mathcal{T}}$ are uniformly sampled from the taxonomy $\mathcal{T}$. Each tuple is provided to a *Large Language Model* (LLM) f_{LLM}, which composes an aerial-view generation prompt $p_i^{(\text{G})}$ embedding the selected attribute values. The prompt is then passed to a *text-guided diffusion model* (TDM) f_{G} to generate the image x_i.

TDMs often deviate from the intended prompt $p_i^{(\text{G})}$ due to *generation defects* [24,49]. To assess these deviations, we employ a *Multimodal Large Language Model* (MLLM) f_{MLLM} to analyze each image and infer its attribute vector $\tilde{\mathbf{a}}_i$. Because the predicted attributes may not align perfectly with the discrete categories in $\mathcal{T}$, an *attribute validation* procedure is applied to refine them (Algorithm 1). Namely, the text embedding $\tilde{e}_i^{(\text{c})}$ of the element $\tilde{a}_i^{(\text{c})}$ in $\tilde{\mathbf{a}}_i$ is compared

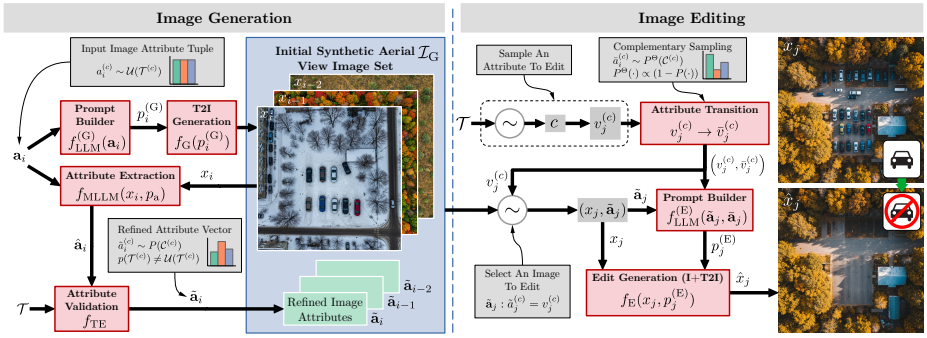

Fig. 4: Overview of the synthetic data generation pipeline. **Left:** Initial image generation using a *text-to-image* foundation model f_{G} conditioned on prompts $p_i^{(\mathrm{G})}$ derived from uniformly sampled attribute values in the taxonomy $\mathcal{T}$. **Right:** Image editing via *image-and-text-to-image* generation to correct distributional biases and enrich underrepresented attribute values.

with the text embeddings of all values $\mathcal{C}^{(c)}$ in $\mathcal{T}$ corresponding to category c. The embeddings are predicted by a text encoder f_{TE}, and cosine similarity is used to perform the comparison. If an available value from $\mathcal{T}$ has cosine similarity with $\tilde{e}_i^{(c)}$ higher than a threshold τ, $\tilde{a}_i^{(c)}$ is swapped with that value. If no similarity higher than τ is found, $\tilde{a}_i^{(c)}$ is added to $\mathcal{T}$ as a new category value. This process ensures semantic consistency between visual content and attribute labels.

However, generation defects introduce imbalances in attribute distributions $P(\mathcal{C}^{(c)})$, despite the uniform sampling of input attributes for image generation. The image editing stage, described next, addresses these biases.

Data Enrichment Through Image Editing The image editing stage employs an *image-and-text-to-image* process to mitigate attribute imbalance within $\mathcal{I}_{\mathrm{G}}$ (Fig. 4-*right*). For each attribute category c, a source value $v_j^{(c)}$ and a target value $\bar{v}_j^{(c)}$ are selected, where the latter is sampled from the complementary distribution $P^{\Theta}(\mathcal{C}^{(c)})$, emphasizing underrepresented attribute values. An image x_j is chosen from $\mathcal{I}_{\mathrm{G}}$ with attribute value $v_j^{(c)}$, and an editing prompt $p_j^{(\mathrm{E})}$ is generated by an LLM $f_{\mathrm{LLM}}^{(\mathrm{E})}$, which explicitly requires the model to preserve unchanged attributes while modifying only the selected one. For more details on the prompts used, refer to the supplementary material. The resulting edited image is analyzed by f_{MLLM} to validate attributes, as in Algorithm 1. Repeating this process M times yields the enriched image set $\mathcal{I}_{\mathrm{E}}$.

Annotations Generative models do not natively produce structured annotations such as bounding boxes. Pre-trained detectors are unsuitable for pseudo-labeling because they rely on real datasets and may inherit their inherent biases. Instead, foundation *Vision–Language Models* (VLMs) and MLLMs are employed

Algorithm 1 Synthetic Image Attribute Validation

Require: Taxonomy tree $\mathcal{T}$, Models: f_{MLLM}, f_{TE}, image x_i, similarity threshold τ
Ensure: Refined attributes $\hat{\mathbf{a}}_i$

$\mathcal{C}^{(\text{env})} \leftarrow \textsc{ChildrenNames}(\mathcal{T}^{(\text{env})})$
$\hat{\mathbf{a}}_i \leftarrow \{\}$
for c in $\mathcal{C}^{(\text{env})}$ **do**
  $\tilde{a}_i^{(c)} \leftarrow \textsc{ExtractAttributeValue}(f_{\text{MLLM}}, x_i, c)$
  $\mathcal{C}^{(c)} \leftarrow \textsc{ChildrenNames}(\mathcal{T}, c)$
  $\mathcal{E}^{(c)} \leftarrow \textsc{ExtractTextEmbeddings}(f_{\text{TE}}, \mathcal{C}^{(c)})$
  $\tilde{\mathbf{e}}_i^{(c)} \leftarrow \textsc{ExtractTextEmbeddings}(f_{\text{TE}}, \tilde{a}_i^{(c)})$
  $\mathcal{S}_i^{(c)} \leftarrow \textsc{CosineSimilarities}(\mathcal{E}^{(c)}, \tilde{\mathbf{e}}_i^{(c)})$
  $k_{\max} \leftarrow \arg\max(\mathcal{S}_i^{(c)})$
  if $\mathcal{S}_i^{(c)}[k_{\max}] \geq \tau$ **then**
    $\hat{a}_i^{(c)} \leftarrow \mathcal{C}^{(c)}[k_{\max}]$
  else
    $\hat{a}_i^{(c)} \leftarrow \tilde{a}_i^{(c)}$
    $\mathcal{C}^{(c)} \leftarrow \mathcal{C}^{(c)} \cup \tilde{a}_i^{(c)}$
  end if
  $\hat{\mathbf{a}}_i \leftarrow \hat{\mathbf{a}}_i \cup \hat{a}_i^{(c)}$
end for
return $\hat{\mathbf{a}}_i$, updated $\mathcal{T}$

in a *zero-shot detection* setting. Prompts instruct these models to output vehicle bounding box coordinates in structured formats (*e.g.*, JSON objects).

Given the limited precision of zero-shot detection in remote sensing imagery [13,46], predictions from multiple models are aggregated and refined through a *human-in-the-loop* process. Human operators verify true-positive detections and discard false positives in a binary decision-making manner. In [32], the authors report that this approach takes, on average, about 1 second per bounding box, whereas drawing bounding boxes takes about 26 seconds per object [47]. Thus, the semi-automatic strategy we employ yields a significant speedup in labeling while maintaining high annotation quality.

3.2 Model Diagnosis

To analyze the behavior of a pre-trained detector, an attribute-driven evaluation strategy is employed. For each selected attribute in the taxonomy $\mathcal{T}$, the synthetic dataset $\mathcal{I}$ is partitioned into subsets according to attribute values. The detector is then evaluated independently on each subset, enabling a fine-grained analysis of its sensitivity to specific visual factors.

4 Experiments

4.1 Evaluation Protocol

We evaluate the proposed method (Sec. 3) using a protocol designed to measure both its diagnostic capability and the extent to which synthetic insights transfer to real data. As the main evaluation metric, we use *Average Precision* (AP) with an *IoU* threshold of 50%, which is widely adopted in the field of object detection in aerial imagery [35, 52]. The evaluation proceeds as follows:

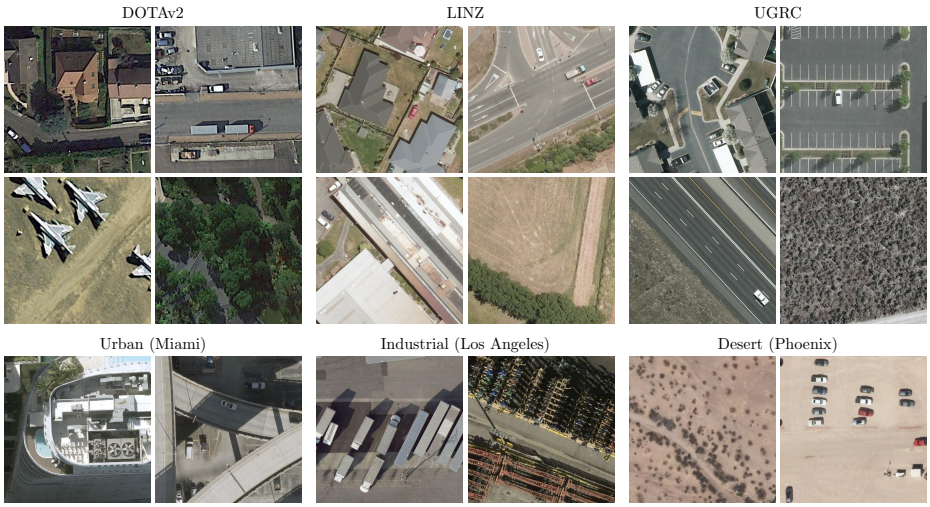

Fig. 5: Samples from the real datasets used in our experiments

1. A pretrained aerial-view vehicle detector is evaluated on the synthetic diagnostic dataset, and both overall and attribute-conditioned APs are computed.
2. Attributes with the lowest AP are identified as underperforming categories.
3. Additional real training images corresponding to these categories are obtained from public sources, labeled, and incorporated into the training set, and the detector is retrained.
4. The updated model is re-evaluated to assess whether the targeted data augmentation leads to the expected performance improvements.

4.2 Datasets

We used three real-world public multi-environment aerial/satellite top-down view image datasets for the initial vehicle detector pre-training: DOTAv2 [11], LINZ [12], and UGRC [12] (Fig. 5). As supplementary datasets, we collected and labeled three new real single-environment datasets: *urban* (Miami, FL, USA), *industrial* (Los Angeles, CA, USA), and *desert* (Phoenix, AZ, USA) (Fig. 5). All real datasets were sampled at 384 px × 384 px with an effective *ground sampling distance* (GSD) of 12.5 cm per px. Since *Imagen 3* produces 1024 px images, they were scaled to 384 px to match the real data. In total, 5,453 synthetic diagnostic images were generated, corresponding to 304 unique combinations of environmental attribute values. More details are provided in Tab. 1 and the supplementary material.

Table 1: Image count per dataset and split

Dataset Name	Image count			
	Train	**Val**	**Test**	**Total**
DOTAv2	28,458	8,807	–	37,265
LINZ	87,654	14,899	27,693	130,246
UGRC	146,993	12,599	8,400	167,992
Urban (Miami)	2,284	–	–	2,284
Industrial (LA)	2,000	–	–	2,000
Desert (Phoenix)	2,000	–	–	2,000
Imagen 3	–	–	5,453	5,453

4.3 Implementation Details

For synthetic data generation (Sec. 3.1), we employ large-scale foundation models accessed via cloud APIs, selected based on image quality and viewpoint consistency for nadir aerial imagery. We evaluated six *text-to-image* models for aerial scene synthesis (Fig. 6). Open-source diffusion models, including *Stable Diffusion 1.5*, *Stable Diffusion XL*, and *Stable Diffusion 3.5* [1, 33, 40], frequently produced unrealistic perspectives, inconsistent object scales, or weak prompt adherence when applied to aerial views without domain-specific fine-tuning. While fine-tuning may mitigate these issues, it reduces direct applicability. Commercial models such as *Gemini 2.5* [8] and *Imagen 3* [2] generated more consistent nadir-view scenes with stronger semantic fidelity (Fig. 6). *DALL-E* [28] showed adequate visual quality but limited structural diversity across prompts. Prior evaluations [42] similarly report stronger photorealism and prompt alignment for *Imagen*. Based on these observations, we adopt *Imagen 3* [2] as the primary model for image generation and editing (*i.e.*, f_{G} and f_{E}), due to its consistent viewpoint control and photorealistic aerial synthesis.

Prompt composition for image generation ($f_{\mathrm{LLM}}^{(\mathrm{G})}$) and editing ($f_{\mathrm{LLM}}^{(\mathrm{E})}$) is performed with *GPT-5* [30] and *Gemini 2.5 Flash Lite* [8], respectively. Image analysis and attribute extraction (f_{MLLM}) are carried out using *Gemini 2.5 Flash* [8], while text embeddings for attribute matching (f_{TE}) are obtained via the *Sentence-BERT* model [36]. Vehicle bounding boxes are extracted using *Gemini 2.5 Flash Lite* and *Moondream 2* [48].

For the downstream evaluation, we consider three widely used aerial-view object detectors [11, 35, 52]: *Faster R-CNN (R50)* [37], *YOLOv8-M* [18], and *ViTDet-B* [22], representing single-stage, two-stage, and transformer-based detection paradigms, respectively. All experiments are implemented using *MMDetection* [5] and *MMYOLO* [9], and are conducted on a server equipped with eight NVIDIA Quadro RTX 6000 GPUs (24 GB VRAM each).

4.4 Aerial-view Object Detectors Diagnosis

Data Annotations The first row of Tab. 2 reports the number of vehicle bounding boxes predicted by *Moondream 2* and *Gemini 2.5 Flash Lite* on our synthetic

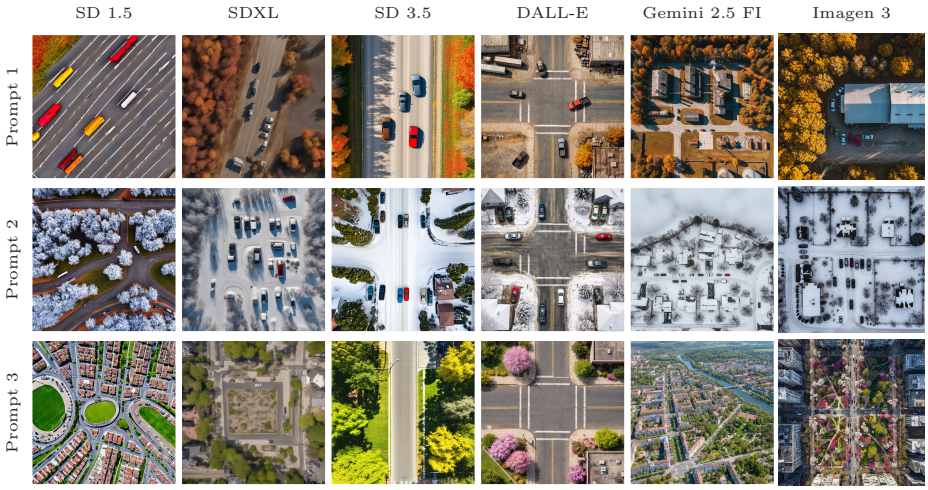

Fig. 6: Qualitative comparison of six *text-to-image* foundation generative models on the aerial-view image generation task. Each row is generated using the same prompt (see the supplementary material for more details). Open-source models: *SD 1.5* [40], *SDXL* [33], and *SD 3.5* [1]. Closed (commercial) models: *DALL-E* [28], *Gemini 2.5 Flash Image* [8], *Imagen 3* [2]. (*Zoom in to see the details better.*)

Table 2: Human operator bounding box approval rates

Dataset Name	Moondream 2	Gemini 2.5 FL	Total	Human Approved
Imagen 3	30,133	51,008	53,885	31,519 (58.5%)
Urban (Miami)	7,606	13,597	14,974	9,741 (65.1%)
Industrial (LA)	3,949	7,850	8,482	4,101 (48.3%)
Desert (Phoenix)	499	1,312	1,429	982 (68.7%)

diagnostic dataset. We merge both prediction sets and average highly overlapping boxes to suppress duplicates, yielding a curated set of pseudo-annotations (reported under *Total*). The last column presents the number and percentage of boxes approved by the human annotator. The same procedure is applied to the three supplementary real datasets, whose statistics are summarized in the bottom rows of Tab. 2. The approval rates underscore the need for human verification and indicate that current foundation models are insufficient for reliable zero-shot detection of aerial vehicles. Nevertheless, their ensemble provides an overcomplete proposal set that substantially accelerates annotation—by approximately 13×, based on reported annotation speeds in [32, 47].

Identifying Lowest-performing Scenes To evaluate the diagnostic capability of our method, we trained three object detectors—*Faster R-CNN*, *YOLOv8*, and *ViTDet*—on three real aerial-view datasets: *DOTAv2*, *LINZ*, and *UGRC*, resulting in nine model–dataset combinations. Each detector was then evalu-

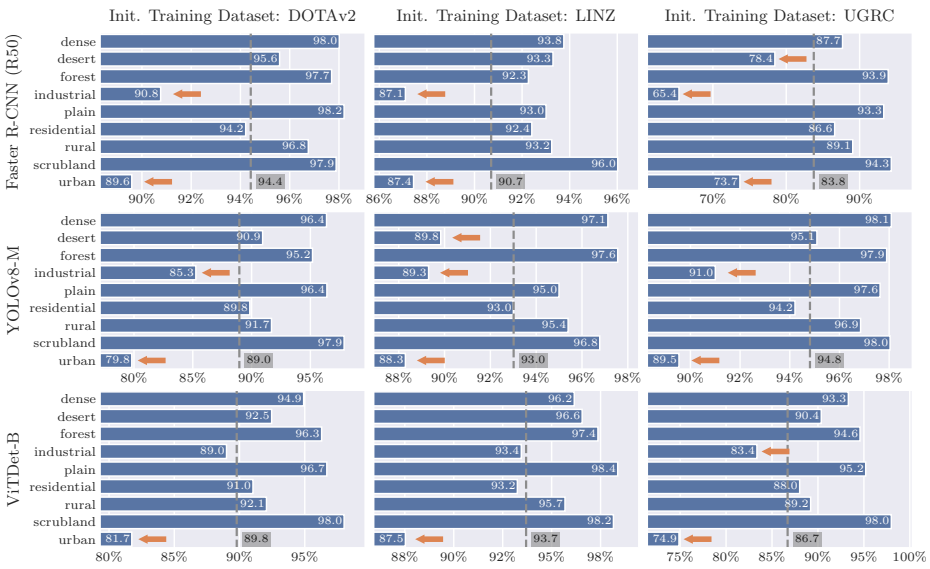

Fig. 7: Scene-wise AP^{50} of three detectors trained on real aerial datasets (DOTAv2, LINZ, UGRC) and evaluated on the synthetic diagnostic set. Rows denote models; columns denote training datasets. Orange arrows highlight scene types below the overall AP (gray dashed line). Best viewed in color.

ated on our synthetic diagnostic dataset to analyze performance across different *scene types*, one of the key image attributes. The dataset includes nine scene categories: *dense*, *desert*, *forest*, *industrial*, *plain*, *residential*, *rural*, *scrubland*, and *urban*. Figure 7 reports the scene-wise *average precision* (AP^{50}) for all models. Orange arrows highlight the *problematic* scene types where the detector performance drops by more than one percentage point relative to the dataset-level AP^{50} (gray dashed line). The *urban* scene has been marked as *problematic* in all nine tests, while the *industrial* scene in seven, and the *desert* scene in two, respectively.

Improvement Through Targeted Supplementation After identifying underperforming scene categories as described in the previous section, we conduct a targeted data supplementation experiment to validate the diagnostic findings. Specifically, we introduce three additional real-world datasets, each representing a single-scene environment (*i.e.*, *urban*, *industrial*, and *desert*), selected from the problematic scene set. These datasets are used to augment the initial training data (*DOTAv2*, *LINZ*, and *UGRC*) and assess whether targeted real-data supplementation—guided by insights from our synthetic diagnostic analysis—leads to measurable improvements in detector performance.

Table 3 summarizes the impact of targeted supplementation across all nine model–dataset combinations. It is important to note that each supplementation dataset is at least an order of magnitude smaller than the training splits of the

Table 3: Effect of supplementing training data with additional urban, industrial, and desert imagery.

Initial Training Dataset	Model	Supplementary Data			Synthetic Data Testing (AP)		
		Urban	Industrial	Desert	Baseline	With suppl.	Gain
DOTAv2	Faster R-CNN R50	✓	✓		93.8%	94.7%	+0.9%
	YOLOv8-M	✓	✓		87.2%	92.3%	+5.1%
	ViTDet-B	✓			86.3%	95.9%	+9.6%
LINZ	Faster R-CNN R50	✓	✓		87.7%	95.1%	+7.4%
	YOLOv8-M	✓	✓	✓	92.7%	95.8%	+3.1%
	ViTDet-B	✓			91.5%	96.1%	+4.6%
UGRC	Faster R-CNN R50	✓	✓	✓	80.9%	93.5%	+12.6%
	YOLOv8-M	✓	✓		94.2%	95.9%	+1.7%
	ViTDet-B	✓	✓		81.6%	95.0%	+13.4%

three initial datasets (Tab. 1). In eight of the nine cases, augmenting the training data with real images from the identified weak scene types led to higher dataset-level AP^{50}, with improvements of up to 13%. Only the *Faster R-CNN* model trained on *DOTAv2* showed no notable change.

Figure 8 provides a finer breakdown of scene-wise AP improvements. In most cases, the supplemented scene categories exhibit the largest relative performance gains, confirming that the insights derived from synthetic data effectively guide real-data collection.

4.5 Targeted vs. Random Supplementation Experiments

Figure 9 compares the AP gains achieved by targeted versus random (non-targeted) supplementation when *DOTAv2* is used as the primary training dataset. To ensure a strictly controlled data budget, we randomly subsample two datasets—*LINZ* (aerial) and *COCO* [23] (non-aerial)—to exactly match the image counts of our targeted supplementary datasets (Tab. 1).

As the results demonstrate, targeted supplementation guided by our diagnostic pipeline consistently outperforms random sampling. For standard architectures such as *Faster R-CNN* and *YOLOv8*, injecting randomly sampled data degrades performance across most scene types (as indicated by negative AP gains). This highlights the danger of blind data scaling, which often introduces domain interference and uninformative samples. In contrast, using the exact same data budget to target identified failure modes yields high-magnitude, positive improvements (e.g., up to +11.7% for *YOLOv8* in urban scenes). While the transformer-based *ViTDet* exhibits greater baseline robustness to random data shifts, our targeted approach still achieves superior peak gains, proving the high data efficiency and practical value of the diagnostic loop.

5 Conclusions

This work demonstrated that foundation generative models can serve as diagnostic instruments for aerial-view object detection systems. By constructing a

Fig. 8: AP^{50} gains per scene type provided by the data supplementation, described in Tab. 3. The values outlined in the orange box correspond to the scene types for which additional supplement data were used during supplementation. Best viewed in color.

synthetic, attribute-controlled testbed, we identified systematic scene-dependent weaknesses of detectors trained on real datasets and showed that synthetic scene-wise trends closely reflect real-world performance gaps. Guided by these diagnostics, targeted supplementation with small, scene-specific real datasets improved performance by up to 13% AP^{50}.

Although instantiated using commercial generative models, the proposed diagnostic loop is model-agnostic. Any generative and multimodal systems capable of producing high-quality, prompt-aligned aerial imagery can be integrated into the pipeline. We selected closed-source models for their current maturity and viewpoint consistency, without task-specific fine-tuning. However, reliance on proprietary systems raises reproducibility considerations, as such models may evolve or become unavailable. Open-source alternatives, while desirable for accessibility, often require substantial computational resources and domain adaptation to achieve comparable realism. As open-weight generative models mature, integrating them into this framework will further improve transparency and long-term reproducibility.

Our experiments focus on RGB nadir aerial imagery and vehicle detection. Extending synthetic diagnostic probing to additional viewpoints and sensing modalities—such as multispectral, thermal, or SAR imagery—remains an important direction. Similarly, while we defined a foundational taxonomy of scene attributes (scene type, season, weather) and performed diagnosis primarily along scene-type variations, broader, more granular studies of attributes are needed to assess cross-factor interactions and task generality.

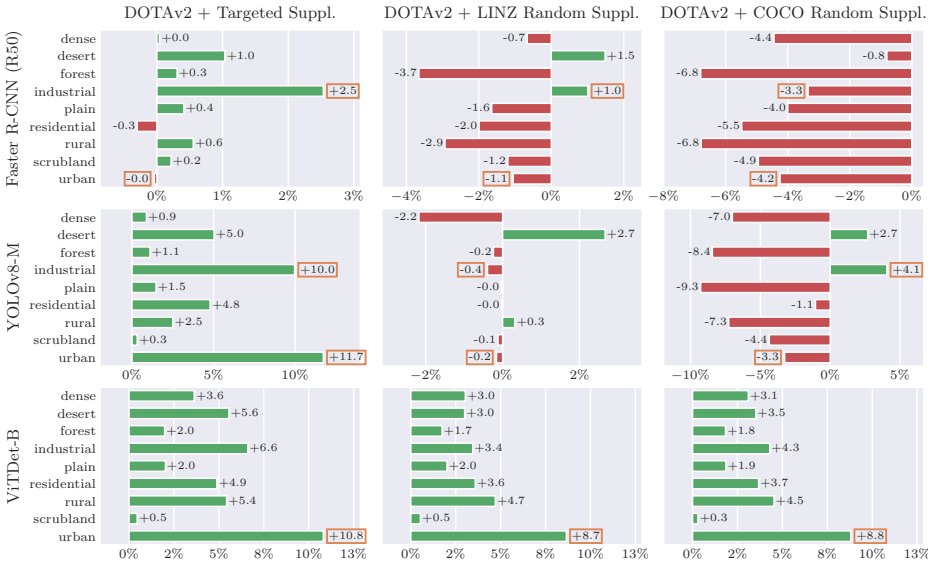

Fig. 9: Targeted vs. random (non-targeted) supplementation AP gains

More broadly, this study highlights controlled synthetic probing as a practical mechanism for improving remote sensing systems. Rather than replacing real data, synthetic diagnostics enable informed data acquisition by identifying where additional coverage is most impactful. In high-stakes Earth observation applications—where balanced data collection is costly, and deployment conditions are variable—such targeted guidance offers a principled path toward more reliable and resource-efficient aerial detection pipelines.

Acknowledgments

This work has been funded by the DEVCOM Army Research Lab, USA.

References

1. AI, S.: Stable diffusion 3.5 large. Machine learning model (2024), available from https://huggingface.co/stabilityai/stable-diffusion-3.5-large
2. Baldridge, J., Bauer, J., Bhutani, M., Brichtova, N., Bunner, A., Castrejon, L., Chan, K., Chen, Y., Dieleman, S., Du, Y., et al.: Imagen 3. arXiv preprint arXiv:2408.07009 (2024)
3. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023)
4. Bolya, D., Foley, S., Hays, J., Hoffman, J.: Tide: A general toolbox for identifying object detection errors. In: European Conference on Computer Vision. pp. 558–573. Springer (2020)

5. 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., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
6. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3339–3348 (2018)
7. Clark, K., Jaini, P.: Text-to-image diffusion models are zero shot classifiers. Advances in Neural Information Processing Systems **36**, 58921–58937 (2023)
8. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
9. Contributors, M.: MMYOLO: OpenMMLab YOLO series toolbox and benchmark. `https://github.com/open-mmlab/mmyolo` (2022)
10. Deng, J., Li, W., Chen, Y., Duan, L.: Unbiased mean teacher for cross-domain object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4091–4101 (2021)
11. Ding, J., Xue, N., Xia, G.S., Bai, X., Yang, W., Yang, M.Y., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges . IEEE Transactions on Pattern Analysis & Machine Intelligence **44**(11), 7778–7796 (Nov 2022). `https://doi.org/10.1109/TPAMI.2021.3117983`, `https://doi.ieeecomputersociety.org/10.1109/TPAMI.2021.3117983`
12. Fang, X., Jeon, M., Qin, Z., Panev, S., De Melo, C., Hu, S., Chakraborty, S., De la Torre, F.: Adapting vehicle detectors for aerial imagery to unseen domains with weak supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8088–8099 (2025)
13. Fiaz, M., Debary, H., Fraccaro, P., Paudel, D., Van Gool, L., Khan, F., Khan, S.: Geovlm-r1: Reinforcement fine-tuning for improved remote sensing reasoning. arXiv preprint arXiv:2509.25026 (2025)
14. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4340–4349 (2016)
15. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of the International Conference on Learning Representations (2019)
16. Hoiem, D., Chodpathumwan, Y., Dai, Q.: Diagnosing error in object detectors. In: European conference on computer vision. pp. 340–353. Springer (2012)
17. Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G.: Gaia-1: A generative world model for autonomous driving (2023), `https://arxiv.org/abs/2309.17080`
18. Jocher, G., Chaurasia, A., Qiu, J.: Yolov8: Ultralytics real-time object detection (2023), `https://github.com/ultralytics/ultralytics`, version 8, available at `https://github.com/ultralytics/ultralytics`
19. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
20. Li, A.C., Prabhudesai, M., Duggal, S., Brown, E., Pathak, D.: Your diffusion model is secretly a zero-shot classifier. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2206–2217 (2023)

21. Li, W., Saeedi, S., McCormac, J., Clark, R., Tzoumanikas, D., Ye, Q., Huang, Y., Tang, R., Leutenegger, S.: Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. In: British Machine Vision Conference (BMVC) (2018)
22. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: European conference on computer vision. pp. 280–296. Springer (2022)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
24. Liu, Q., Kortylewski, A., Bai, Y., Bai, S., Yuille, A.: Discovering Failure Modes of Text-guided Diffusion Models via Adversarial Search. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=TOWdQQgMJY
25. Midjourney, Inc.: Midjourney. https://www.midjourney.com (2022), accessed: 2025-11-13
26. Müller, M., Casser, V., Lahoud, J., Smith, N., Ghanem, B.: Sim4cv: A photo-realistic simulator for computer vision applications. International Journal of Computer Vision **126**(9), 902–919 (2018)
27. NVIDIA, :, Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., Dworakowski, D., Fan, J., Fenzi, M., Ferroni, F., Fidler, S., Fox, D., Ge, S., Ge, Y., Gu, J., Gururani, S., He, E., Huang, J., Huffman, J., Jannaty, P., Jin, J., Kim, S.W., Klár, G., Lam, G., Lan, S., Leal-Taixe, L., Li, A., Li, Z., Lin, C.H., Lin, T.Y., Ling, H., Liu, M.Y., Liu, X., Luo, A., Ma, Q., Mao, H., Mo, K., Mousavian, A., Nah, S., Niverty, S., Page, D., Paschalidou, D., Patel, Z., Pavao, L., Ramezanali, M., Reda, F., Ren, X., Sabavat, V.R.N., Schmerling, E., Shi, S., Stefaniak, B., Tang, S., Tchapmi, L., Tredak, P., Tseng, W.C., Varghese, J., Wang, H., Wang, H., Wang, H., Wang, T.C., Wei, F., Wei, X., Wu, J.Z., Xu, J., Yang, W., Yen-Chen, L., Zeng, X., Zeng, Y., Zhang, J., Zhang, Q., Zhang, Y., Zhao, Q., Zolkowski, A.: Cosmos world foundation model platform for physical ai (2025), https://arxiv.org/abs/2501.03575
28. OpenAI: Dall·e 3. https://platform.openai.com/docs/guides/images (2023), accessed: 2025-11-13
29. OpenAI: Gpt-image-1. https://platform.openai.com/docs/guides/images (2024), accessed: 2025-11-13
30. OpenAI: Gpt-5: Large multimodal model. https://openai.com/gpt-5 (2025), available at https://openai.com/gpt-5
31. Paolo, F., Lin, T.t.T., Gupta, R., Goodman, B., Patel, N., Kuster, D., Kroodsma, D., Dunnmon, J.: xview3-sar: Detecting dark fishing activity using synthetic aperture radar imagery. Advances in Neural Information Processing Systems **35**, 37604–37616 (2022)
32. Papadopoulos, D.P., Clarke, A.D.F., Keller, F., Ferrari, V.: Training object class detectors from eye tracking data. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 361–376. Springer International Publishing, Cham (2014)
33. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., Sun, Y. (eds.) International Conference on Learning Representations. vol. 2024, pp. 1862–1874 (2024), https://proceedings.iclr.cc/paper_files/paper/2024/file/081b08068e4733ae3e7ad019fe8d172f-Paper-Conference.pdf

34. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. ArXiv **abs/2209.14988** (2022), https://api.semanticscholar.org/CorpusID:252596091
35. Randieri, C., Ganesh, S.V., Raj, R.D.A., Yanamala, R.M.R., Pallakonda, A., Napoli, C.: Aerial autonomy under adversity: Advances in obstacle and aircraft detection techniques for unmanned aerial vehicles. Drones **9**(8) (2025). https://doi.org/10.3390/drones9080549, https://www.mdpi.com/2504-446X/9/8/549
36. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). p. 3982. Association for Computational Linguistics (2019)
37. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
38. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
39. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: European conference on computer vision. pp. 102–118. Springer (2016)
40. 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. pp. 10684–10695 (2022)
41. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3234–3243 (2016). https://doi.org/10.1109/CVPR.2016.352
42. Saharia, C., Chan, W., Saxena, S., Lit, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Gontijo-Lopes, R., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22, Curran Associates Inc., Red Hook, NY, USA (2022)
43. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
44. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision **128**(2), 336–359 (Oct 2019). https://doi.org/10.1007/s11263-019-01228-7, http://dx.doi.org/10.1007/s11263-019-01228-7
45. Shermeyer, J., Hossler, T., Van Etten, A., Hogan, D., Lewis, R., Kim, D.: Rareplanes: Synthetic data takes flight. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 207–217 (2021)
46. Soni, S., Dudhane, A., Debary, H., Fiaz, M., Munir, M.A., Danish, M.S., Fraccaro, P., Watson, C.D., Klein, L.J., Khan, F.S., et al.: Earthdial: Turning multi-sensory earth observations to interactive dialogues. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14303–14313 (2025)

47. Su, H., Deng, J., Fei-Fei, L.: Crowdsourcing annotations for visual object detection. In: Human Computation - Papers from the 2012 AAAI Workshop, Technical Report. pp. 40–46. AAAI Workshop - Technical Report (2012), 2012 AAAI Workshop ; Conference date: 23-07-2012 Through 23-07-2012
48. Vikhyat, K.: Moondream 2, rev. 2025-06-21 (2024), https://huggingface.co/vikhyatk/moondream2, hugging Face. DOI: 10.57967/hf/6762
49. Wang, R., Chen, Z., Chen, C., Ma, J., Lu, H., Lin, X.: Compositional text-to-image synthesis with attention map control of diffusion models. In: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. pp. 5544–5552 (2024)
50. Wu, W., Zhao, Y., Chen, H., Gu, Y., Zhao, R., He, Y., Zhou, H., Shou, M.Z., Shen, C.: Datasetdm: Synthesizing data with perception annotations using diffusion models. Advances in Neural Information Processing Systems **36**, 54683–54695 (2023)
51. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection (2022), https://arxiv.org/abs/2203.03605
52. Zhang, X., Zhang, T., Wang, G., Zhu, P., Tang, X., Jia, X., Jiao, L.: Remote sensing object detection meets deep learning: A metareview of challenges and advances. IEEE Geoscience and Remote Sensing Magazine **11**(4), 8–44 (2023). https://doi.org/10.1109/MGRS.2023.3312347