RarePlanes: Synthetic Data Takes Flight

Jacob Shermeyer¹, Thomas Hossler², Adam Van Etten¹, Daniel Hogan¹, Ryan Lewis¹, and Daeil Kim²

¹IQT - CosmiQ Works, [jshermeyer, avanetten, dhogan, rlewis]@iqt.org
²AI.Reverie, [thomas.hossler, daeil]@aireverie.com

Abstract

RarePlanes is a unique open-source machine learning dataset that incorporates both real and synthetically generated satellite imagery. The RarePlanes dataset specifically focuses on the value of synthetic data to aid computer vision algorithms in their ability to automatically detect aircraft and their attributes in satellite imagery. Although other synthetic/real combination datasets exist, RarePlanes is the largest openly-available very-high resolution dataset built to test the value of synthetic data from an overhead perspective. Previous research has shown that synthetic data can reduce the amount of real training data needed and potentially improve performance for many tasks in the computer vision domain. The real portion of the dataset consists of 253 Maxar WorldView-3 satellite scenes spanning 112 locations and 2,142 km² with 14,700 hand-annotated aircraft. The accompanying synthetic dataset is generated via AI.Reverie’s simulation platform and features 50,000 synthetic satellite images simulating a total area of 9331.2 km² with ~ 630,000 aircraft annotations. Both the real and synthetically generated aircraft feature 10 fine grain attributes including: aircraft length, wingspan, wing-shape, wing-position, wingspan class, propulsion, number of engines, number of vertical-stabilizers, presence of canards, and aircraft role. Finally, we conduct extensive experiments to evaluate the real and synthetic datasets and compare performances. By doing so, we show the value of synthetic data for the task of detecting and classifying aircraft from an overhead perspective.

1. Introduction

Over the last decade, computer vision research and the development of new algorithms has been driven largely by permissively licensed open datasets. Datasets such as ImageNet [9], MSCOCO [34], and PASCAL VOC [14] (among others) remain critical drivers for advancement. Convolutional neural networks (CNNs), currently the leading class of algorithms for most vision tasks [45, 71], require a large amount of annotated observations. However, the development of such datasets is often manually intensive, time-consuming, and costly to create. An alternative approach to manually annotating training data is to create computer generated images and annotations (referred to as synthetic data). After creating realistic 3D environments, one can then generate thousands of images at virtually no cost. Such data has been shown to be effective for augmenting and replacing real data, thus reducing the burden of dataset creation. Synthetic datasets continue to be developed and have been notably helpful in various domains including: autonomous driving [46, 47, 48, 17], optical flow [35, 46, 31], facial recognition [30, 10, 29], and domain adaptation [8, 26, 24, 59] (see Section 2.1 for further detail).

Although synthetic datasets continue to become more prevalent, no expansive permissively licensed synthetic datasets exist in the context of overhead observation. Overhead imagery presents unique challenges for computer vision models such as: the detection of small visually-heterogeneous objects, varying look angles or lighting conditions, and different geographies that can feature distinct seasonal variability. As such, creating synthetic datasets from an overhead perspective is a significant challenge and simulators must attempt to closely mimic the complexities of a spaceborne or aerial sensor as well as the Earth’s ever-changing conditions. For example, to create a large and heterogeneous synthetic dataset, one must account for each sensor’s varying spatial resolution, changes in sensor look angle, the time of day of collection, shadowing, and changes in illumination due to the sun’s location relative to the sensor. Furthermore, the simulator must be able to account for other factors such as the ground appearance due to seasonal change, weather conditions, and varying geographies or biomes.

While synthetic datasets certainly have the potential to be beneficial, they require a paired real dataset with shared features to baseline performance and quantitatively...
Figure 1. Example of the real and synthetic datasets present in RarePlanes. The top two rows feature the real Maxar WorldView-3 satellite imagery and the bottom two rows show the synthetic data. The dataset features variable weather conditions, biomes, and ground surface types.

test value. However, few permissively licensed overhead datasets [13, 64, 51] exist that focus on detection or segmentation tasks and feature very-high resolution real imagery from an overhead perspective. Overhead datasets remain one of the best avenues for developing new computer vision methods that can adapt to limited sensor resolution, variable look angles, and locate tightly grouped, cluttered objects. Such methods can extend beyond the overhead space and be helpful in other domains such as face-id, autonomous driving, and surveillance.

To address the limitations described above, we introduce the RarePlanes dataset. This dataset focuses on the detection of aircraft and their fine-grain attributes from an overhead perspective. It consists of both an expansive synthetic and real dataset. We use AI.Reverie’s novel platform which incorporates unreal engine [12] to develop realistic synthetic data based off of real world airports. The platform ingests real world metadata and overhead images to procedurally generate 3D environments of real world locations. The weather, time of collection, sunlight intensity, look angle, biome, and distribution of aircraft model are among the multiple parameters that the simulator can modify to create diverse and heterogeneous data. The synthetic portion of RarePlanes consists of 50,000 images simulating a total area of 9331.2 km$^2$ and $\sim$ 630,000 annotations. The real portion consists of 253 Maxar WorldView-3 satellite images spanning 112 locations in 22 countries and 2,142 km$^2$ with $\sim$ 14,700 hand annotated aircraft. Examples of the synthetic and real images are shown in Figure 1.

RarePlanes also provides fine-grain labels with 10 distinct aircraft attributes and 33 different sub-attribute choices labeled for each aircraft. These include: aircraft length, wingspan, wing-shape, wing-position, Federal Aviation Administration (FAA) wingspan class [19], propulsion, number of engines, number of vertical-stabilizers, canards, and aircraft type or role. Although several other overhead detection datasets exist [13, 64, 51, 32, 20, 62], no others have multiple fine-grain attributes that detail specific object features. Such fine-grain attributes have been particularly helpful for zero-shot learning applications [16] and enable end users to create diverse custom classes. Using these combined attributes, anywhere from 1 to 110
classes can be created for individual research purposes. The dataset is available for free download through Amazon Web Services’ Open Data Program, with download instructions available at [https://www.cosmiqworks.org/RarePlanes](https://www.cosmiqworks.org/RarePlanes).

**Contributions**

- An expansive real and synthetic overhead computer vision dataset focused on the detection of aircraft and their features.

- Annotations with fine-grain attributions that enable various CV tasks such as: detection, instance segmentation, or zero-shot learning.

- Extensive experiments to evaluate the real and synthetic datasets and compare performances. By doing so, we show the value of synthetic images for the task of detecting and classifying aircraft from an overhead perspective.

## 2. Related Work

RarePlanes sits at the intersection of three distinct computer vision dataset domains: synthetic datasets, geospatial datasets, and fine-grain attribution datasets. These three domains are cornerstones around which computer vision research has continued to rapidly advance and grow. We summarize the key characteristics of modern synthetic, geospatial, and attribute datasets in Table 1 and compare them to the RarePlanes dataset.

### 2.1. Synthetic Datasets

Synthetic data has become prevalent across many computer vision domains and has shown value as a replacement for real data or to augment existing training datasets [47, 29, 48, 1, 42]. Many synthetic datasets focus on the autonomous driving domain; including the Synthia [48], GTA [46, 47], and vKITTI [17] datasets. These synthetic datasets are often paired with real-world data such as Cityscapes [7], CamVid [3], or KITTI [18] to benchmark the value of synthetic data. Other notable synthetic datasets such as SUNCG [54] or Matterport3D [5] focus on indoor scenes and include RGB-D data for depth estimation. Moreover, other datasets focus on addressing challenging occlusion (amodal) problems such as the expansive SAIL-VOS [25] and DvCE [11]. Finally, the Synthinel-1 [28] Dataset is the only other dataset that bridges the synthetic/satellite domain. It features synthetic data from an overhead perspective with binary pixel masks of building footprints. Overall, combined synthetic and real datasets, similar to RarePlanes, have been helpful with several different tasks including: enhancing object detection [56, 40, 42], semantic segmentation [48, 21, 50], or instance segmentation performance [63, 1]. Furthermore, such datasets continue to inspire new domain adaptation (DA) techniques [8, 26, 24, 59, 56]. Such DA techniques could be particularly valuable for overhead applications as there remains a dearth of openly available training data and models trained on one location often do not generalize well to new areas.

![Three annotation styles within RarePlanes. The dataset features three annotation styles including: 'Bounding Box', 'Diamond Polygon', and 'Full Instance Segmentation' (synthetic only).](image)

### 2.2. Geospatial Datasets

Geospatial and very-high resolution remote sensing datasets have continued to draw increased interest due to their relevancy to many computer vision challenges. Such datasets contain lower resolution images with tiny, closely grouped objects with varying aspect ratios, arbitrary orientations and high annotation density. The lessons learned from such datasets continue to inspire new computer vision approaches related to detection [69, 61, 58], segmentation [27], super-resolution [52, 41], and even bridges to natural language processing [57]. Some notable datasets include SpaceNet [13, 64, 51] and xBD [20], which focus on foundational mapping and instance/semantic segmentation for problems such as building footprint and road network extraction or building damage assessment. Others such as xView [32], A large-scale dataset for object detection in aerial images (DOTA) [66] and A Large-scale Dataset for Instance Segmentation in Aerial Images (iSAID) [63] focus on overhead object detection or instance segmentation, featuring multiple classes of different object types. The Functional Map of the World (FMOW) [6] dataset centers on the task of classification of smaller image chips from an overhead perspective. Other datasets such as CORE3D [4], MVS [2], the ISPRS benchmark [49] and SatStereo [38] target height extraction and 3D mapping from space. RarePlanes builds upon these existing datasets and contributes both synthetic and real data. Furthermore, RarePlanes adds 10 unique object attributes, which enable customizable classes, as well as three annotation styles per object (Bounding Box, Diamond Polygon, and Full-Instance (Synthetic Only)).

### 2.3. Fine-Grain Attribute Datasets

Many datasets focus on identifying general objects in imagery, however, several others take an alternative approach and label unique attributes of each object. As previ-
Table 1. Comparison with other synthetic, attribute and overhead imagery datasets. Our dataset has a similar scale as modern computer vision datasets and provides both a real and synthetic component. For SpaceNet (Buildings + Road Speed), xBD (Building Damage Scale), and RarePlanes we report the range of possible customizable classes that end-users can create using varieties of the dataset attributes.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Gigapixels</th>
<th>Classes</th>
<th>Attributes</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Real</td>
<td>Synthetic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SpaceNet [13, 64, 51]</td>
<td>100.1</td>
<td>1 to 8</td>
<td>1</td>
<td>859,982</td>
</tr>
<tr>
<td>xBD [20]</td>
<td>9.8</td>
<td>1 to 4</td>
<td>1</td>
<td>850,736</td>
</tr>
<tr>
<td>xView [32]</td>
<td>56.0</td>
<td>60</td>
<td>0</td>
<td>1,000,000</td>
</tr>
<tr>
<td>iSAID [62]</td>
<td>44.9</td>
<td>15</td>
<td>0</td>
<td>655,451</td>
</tr>
<tr>
<td>FMOW [6]</td>
<td>1084.0</td>
<td>63</td>
<td>0</td>
<td>132716</td>
</tr>
<tr>
<td>Cityscapes [7] + GTA [46]</td>
<td>537.5</td>
<td>30/19</td>
<td>0</td>
<td>210,179</td>
</tr>
<tr>
<td>COCOA [72] + SAIL-VOS [25]</td>
<td>115.7</td>
<td>-/163</td>
<td>0</td>
<td>46,314</td>
</tr>
<tr>
<td>Animals with Attributes 2 [67]</td>
<td>24.7</td>
<td>50</td>
<td>85</td>
<td>37,322</td>
</tr>
<tr>
<td>CompCars [68]</td>
<td>86.1</td>
<td>1,716</td>
<td>13</td>
<td>136,726</td>
</tr>
<tr>
<td>RarePlanes (Ours)</td>
<td>187.1</td>
<td>1 to 110</td>
<td>10</td>
<td>14,707</td>
</tr>
</tbody>
</table>

3. The RarePlanes Dataset and Statistics

3.1. Annotations, Features, and Attributes

The RarePlanes dataset contains 14,707 real and 629,551 synthetic annotations of aircraft. Each aircraft is labeled in a diamond style with the nose, left-wing, tail, and right-wing being labeled in successive order (Figure 2). This annotation style has the advantage of being: simplistic, easily reproducible, convertible to a bounding box, and ensures that aircraft are consistently annotated (other hand-annotated formats can often lead to imprecise labeling). Furthermore, this annotation style enables the calculation of aircraft length and wingspan by measuring between the first annotation node to the third and from the second to the fourth. We employ a professional labeling service to produce high-quality annotations for the real portion of the dataset. Two rounds of quality control are included in the process, a first one by the professional service and a second by the authors.

After each aircraft is annotated in the diamond format, an expert geospatial team labels aircraft features. The features include attributes of aircraft wings, engines, fuselage, tail, and role (Figure 7). We ultimately chose these attributes as they were visually distinctive from an overhead perspective and have been shown to be helpful in aiding to visually identifying the type or make of aircraft [45].

- **Wings:** We label aircraft Wing Shape: (‘straight’, ‘swept’, ‘delta’, and ‘variable-swept’), Wing Position: (‘high mounted’ and ‘mid/low mounted’), Wingspan in Meters: (‘float’), and the FAA Aircraft Design Group Wingspan Class: [19] (‘1’ to ‘6’) which determines which airports can accommodate different sized aircraft. Examples of wing-shape and position can be seen in figure 3.

![Wing Shapes](image)

- **Fuselage:** We label aircraft Length in Meters: (‘float’) and if the plane has Canards: (‘yes’ or ‘no’). Canards are small fore-wings that are added to planes to increase maneuverability or reduce the load/airflow on the main wing.

- **Engines:** we label the Number of Engines: (‘0’ to ‘4’) and the Type of Propulsion: (‘unpowered’, ‘jet’, ‘propeller’).

![Canards and Engines](image)
• **Tail:** We label the **Number of Vertical Stabilizers:** (‘1’ or ‘2’) or tail fins that each aircraft possesses.

![Canards 2 Tail Fins 1 Tail Fin](image)

Figure 5. Examples of fuselage and tail attributes including canards and number of vertical stabilizers [55] present in the RarePlanes dataset.

• **Role:** After labeling each attribute, we then use these attributes to classify the **Role or Type:** of an aircraft into seven unique classes. These include: ‘Civil Transport/Utility’ (‘Small’, ‘Medium, and ‘Large’ based upon wingspan), ‘Military Transport/Utility/ AWAC’, ‘Military Bomber’, ‘Military Fighter/Interceptor/Attack’, and ‘Military Trainer’. Further detail on role definitions and can be found in the RarePlanes User Guide, hosted on AWS with the dataset.

![Aircraft by role at 1:150 scale. Note the heterogeneous aircraft styles and variation in sizes. Top Row: Small, Medium and Large Civil Transports. Bottom Row: Military Fighter, Bomber, and Transport.](image)

Figure 6. Aircraft by role at 1:150 scale. Note the heterogeneous aircraft styles and variation in sizes. Top Row: Small, Medium and Large Civil Transports. Bottom Row: Military Fighter, Bomber, and Transport.

3.2. Real Imagery and Locations

All electro-optical imagery is provided by the Maxar Worldview-3 satellite with a maximum ground sample distance (GSD) of 0.31 to 0.39 meters depending upon sensor look-angle. The dataset consists of 253 unique scenes, spanning 2,142 km$^2$ with 112 locations in 22 countries. Locations were chosen by performing a stratified random sampling of OpenStreetMap [36] aerodromes of area $\geq 1$ km$^2$ across the US and Europe using the Köppen climate zone as the stratification layer. We stratify by climate to increase seasonal diversity and geographic heterogeneity. Seven additional locations were manually chosen as they overlap with preexisting datasets [64, 13], and we considered further revisits over these locations to potentially have additional value. We then chose individual satellite scenes by attempting to select scenes from different seasons for each location. Many locations have several scenes taken at different points in time, which may enable future investigation.
on the value of annotating the same areas using multiple images. The imagery is collected from variable look-angles (3.2 to 29.6°), target azimuth angles (1.8 to 359.7°), and sun elevation angles (10.7 to 79.0°). Imagery is collected from all four seasons, with scenes featuring instances of cloud cover (12.6%), snow (9.1%) and clear skies (78.3%). Combined together, this leads to high variability in illumination, shadowing, and lighting conditions. Consequently, the dataset should help to improve generalizability to new areas. Finally, background surfaces are quite diverse with grass, dirt, concrete, and asphalt surface types. The collection is composed of three different sets of data with different spatial resolutions: one panchromatic band (0.31 – 0.39m), eight multi-spectral (coastal to NIR (400 – 954μm)) bands (1.24 – 1.56m), and three RGB (448 – 692μm) pan-sharpened bands (0.31 – 0.39m). Each data product is atmospherically compensated to surface-reflectance values by Maxar’s AComp [37] and orthorectified using the SRTM DEM. RGB data is also converted to 8-bit. Areas containing non-valid imagery are set to 0. We distribute both 512 × 512 pixel tiles (20% overlap) that contain aircraft as well as as the full images, cropped to the extent of the area of annotation.

### 3.3. Synthetic Data

The synthetic dataset contains 629,551 annotations of aircraft across 50,000 images with an extent of 1920 × 1080 pixels at 15 distinct locations, simulating a total area of 9331.2 km². All synthetic data is created via AI.Reverie’s novel simulator software in a multi-step process. First, individual real-world locations are simulated using a proprietary procedural GIS framework integrated with Houdini [53]. The framework ingests geospatial vector data (OpenStreetMap [50]) and overhead images to procedurally generate building and airport terminal models. Airport ground and runways are then generated through simple planar shape and texture projections. Once the simulated locations are generated, we import each location into unreal engine 4 [12]. Capture modules and settings are configured to spawn aircraft and simulate the environmental factors and biomes. Each image features a simulated GSD of 0.3 meters to closely approximate our real imagery and is collected from variable look-angles ranging between 5.0 to 30.0° off-nadir. The imagery is evenly split across 5 distinct biomes including: ‘Alpine’, ‘Arctic’, ‘Temperate Evergreen Forests’, ‘Grasslands’, and ‘Tundra’. The biome parameter controls the type of vegetation, its density, as well as the ground textures. Four unique weather conditions are also evenly distributed across the dataset including: ‘Overcast’, ‘Clear Sky’, ‘Snow’, and ‘Rain’. Other parameters include the sunlight intensity, weather intensity, and the time of the day. Ultimately, this produces an expansive heterogeneous dataset with a wide variety of backgrounds. We believe that this dataset will be helpful in improving model generalizability to new areas and developing new algorithmic approaches that could move beyond aircraft detection.

### 4. Experiments, Results, and Discussion

In this section, we validate the synthetic dataset by running three experiments for two tasks: object detection and instance segmentation. For each task, we train a benchmark network on three subsets of data: on the real data only, on the synthetic data only, and perform a fine tuning experiment training on the synthetic data and then a portion (∼ 10%) of the real dataset. Each experiment is validated on the test real dataset and the results are shown Table [2]. We ran these experiments for two attributes: aircraft (detection of an aircraft without classifying it) and civil role.

#### 4.1. Training and Testing Splits

For the real world data, given the size of the raw satellite scenes, we adopted a tiling approach. Each scene is split into 512x512 tiles containing at least one aircraft. Furthermore, we ensure that the training and test split contains at
Table 2. Results of the object detection and segmentation experiments. We report models performance trained on the real dataset (Real) and the synthetic dataset (Synth.) as well as the fine tuning experiment (FT) using only 10% of the real training dataset. We show the results of the single class experiments (‘aircraft’) and the three classes experiment: small ($C_S$), medium ($C_M$), and large ($C_L$) civil transport aircraft. Performance is evaluated using the mean average precision (mAP) (IOU@[0.5:0.95]), the mAP50 (IOU@0.5) and the average recall (AR) metrics, as well as the class APs when applicable. For the Mask R-CNN instance segmentation experiments, we only report the segmentation AP. Each value reported is an average of 5 runs. The standard deviations for mAP, mAP50, and AR are also indicated.

<table>
<thead>
<tr>
<th>Network</th>
<th>Attribute</th>
<th>Dataset</th>
<th>$C_S$</th>
<th>$C_M$</th>
<th>$C_L$</th>
<th>mAP</th>
<th>mAP50</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN</td>
<td>aircraft</td>
<td>Real</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>73.32 (0.34)</td>
<td>96.80 (0.02)</td>
<td>77.16 (0.21)</td>
</tr>
<tr>
<td></td>
<td>aircraft</td>
<td>Synth.</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>54.86 (0.25)</td>
<td>87.03 (0.53)</td>
<td>60.67 (0.27)</td>
</tr>
<tr>
<td></td>
<td>aircraft</td>
<td>FT</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>69.16 (0.69)</td>
<td>95.29 (0.41)</td>
<td>73.03 (0.57)</td>
</tr>
<tr>
<td></td>
<td>role</td>
<td>Real</td>
<td>66.68</td>
<td>70.26</td>
<td>67.68</td>
<td>68.21 (0.4)</td>
<td>92.16 (0.23)</td>
<td>75.39 (0.40)</td>
</tr>
<tr>
<td></td>
<td>role</td>
<td>Synth.</td>
<td>27.70</td>
<td>37.09</td>
<td>42.85</td>
<td>35.88 (2.26)</td>
<td>59.09 (2.9)</td>
<td>53.82 (1.28)</td>
</tr>
<tr>
<td></td>
<td>role</td>
<td>FT</td>
<td>56.73</td>
<td>66.05</td>
<td>66.52</td>
<td>63.10 (0.78)</td>
<td>89.15 (0.22)</td>
<td>71.06 (0.75)</td>
</tr>
<tr>
<td>Mask R-CNN</td>
<td>aircraft</td>
<td>Real</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>73.67 (0.17)</td>
<td>96.81 (0.03)</td>
<td>76.46 (0.20)</td>
</tr>
<tr>
<td></td>
<td>aircraft</td>
<td>Synth.</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>56.28 (0.46)</td>
<td>87.54 (0.69)</td>
<td>60.71 (0.51)</td>
</tr>
<tr>
<td></td>
<td>aircraft</td>
<td>FT</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>70.51 (0.34)</td>
<td>94.73 (0.53)</td>
<td>73.72 (0.26)</td>
</tr>
<tr>
<td></td>
<td>role</td>
<td>Real</td>
<td>65.60</td>
<td>72.13</td>
<td>70.97</td>
<td>69.57 (0.47)</td>
<td>91.89 (0.55)</td>
<td>76.16 (0.30)</td>
</tr>
<tr>
<td></td>
<td>role</td>
<td>Synth.</td>
<td>29.12</td>
<td>41.78</td>
<td>47.47</td>
<td>39.46 (3.20)</td>
<td>62.31 (4.51)</td>
<td>57.33 (1.96)</td>
</tr>
<tr>
<td></td>
<td>role</td>
<td>FT</td>
<td>58.96</td>
<td>70.02</td>
<td>72.33</td>
<td>67.11 (0.46)</td>
<td>90.03 (0.52)</td>
<td>74.40 (0.58)</td>
</tr>
</tbody>
</table>

least one satellite scene per country to maximize geographic diversity. As the dataset contains satellite images captured over the same location at different dates, an airfield can appear in both splits but at different points in time. Moreover, we created a subset of the real training split for the fine tuning experiments. This subset contains roughly 10% of the images of the training split, created by drawing a 10% random sample of image tiles by location. All synthetic data is included in the training split to maintain a domain gap. For this study we train on 45,000 of the 50,000 images, reserving 5,000 images for cross-validation purposes. Note these cross-validation results are not reported here. Further details on the initial training and testing splits used in this study can be found at [https://www.cosmigworks.org/RarePlanes](https://www.cosmigworks.org/RarePlanes). Ultimately end users are encouraged to reorganize and create their own splits for their specific research or use-case(s).

4.2. Implementation

In our experiments, we used a Resnet-50 [23] and FPN [33] as the backbone for the Faster R-CNN [44] detection network. A similar backbone was used for the Mask R-CNN [22] instance segmentation network. Backbones are pre-trained using ImageNet [9] weights and all experiments are conducted with the Detectron2 framework [65], using the default configurations for each network. The network was optimized with Stochastic Gradient Descent (SGD) using a learning rate of 0.001, weight decay of 0.0001 and a momentum of 0.9. Additionally, we used a linear warmup period over 1K iterations. We maintain a consistent learning rate for the fine tuning experiments. We found that decreasing the learning rate or freezing some of the layers in the backbone did not improve performance. The networks were trained on a NVIDIA Tesla V100 GPU with 12GB memory. Each network was trained until convergence, which was reached after around 60K iterations. We also applied basic pixel level augmentations, such as blurring and modifying the contrast or the brightness. Finally, we performed random cropping (512x512) when training on the synthetic dataset.

4.3. Results and Discussion

We evaluated our network performances using the COCO average precision (AP) metric. Table 2 reports the average precision for each class as well as the mAP, mAP50, and average recall (AR). Qualitative results are shown in Figure 9.

In the first set of experiments, we focused on the performance of the synthetic dataset only. As expected, we observe a drop in performances when training on the synthetic data only, due to the domain gap between the real and synthetic datasets. We observe that the model trained on the synthetic dataset tends to mislabel clutter or nearby objects as aircraft, as shown in Figure 9. Additionally, snow patches, ground markings, airport vehicles are sometimes detected as aircraft. This leads to a significantly lower AP (55% to 75% of the real AP) when models are trained on the synthetic dataset only. However, the AR is not as sensitive
Figure 9. Example of aircraft detection results. (a) ground truth, (b) model trained real dataset (c) model trained on synthetic dataset (d) model fine tuned on real subset.

to the domain gap (70% to 80% of the real AR), meaning that the majority of aircraft are still detected when only the synthetic dataset is used. Similarly, we observe that the drop in AP50 is also lower relative to the AP metric. Ultimately, the AP50 metric may be more informative as we are most interested in accurately counting aircraft, rather than how well they are localized.

Most importantly, when a small subset (∼10%) of real data is added for fine tuning, we observe a significant gain in mAP, leading to similar performance (91 to 96% of real mAP) to the models trained on the real dataset only. We hypothesize that the synthetic data helps to build a prior model for aircraft detection and eases transfer learning, thus greatly reducing the need for annotated real data. In Figure 9 we see how fine tuning on the real subset removes some of the false positive predictions versus training on the synthetic dataset only. However, the false positive detection rate still remains slightly higher compared to training on the entire real training set. It’s important to note that the goal of these experiments is to define a baseline for future experimentation for other algorithms to improve upon, particularly within the area of domain adaptation.

5. Conclusions

We present RarePlanes, a unique machine learning dataset that incorporates both real and synthetically generated satellite imagery. This dataset is critical for expanding and examining the value of synthetic data for overhead applications. Our benchmark experiments using the dataset found that blends of 90% synthetic to 10% real can deliver nearly equivalent performance as 100% real data for the task of aircraft identification. Additionally, we believe that RarePlanes could be particularly valuable for developing new and extensible domain adaptation approaches. Finally, the detailed fine-grain attribution and annotation contained in the dataset may enable various CV tasks and future experiments in the fields of detection, instance segmentation, or zero-shot learning.

Acknowledgments

The authors thank the whole AI.Reverie team for making the creation of the synthetic dataset possible. We would especially like to thank Danny Gillies and Natasha Ruiz for their devoted help. Additionally we thank IQT’s Christyn Zehnder for her contributions in launching RarePlanes.
References


[53] SideFX. Houdini.


