SKYSCENES: A Synthetic Dataset for Aerial Scene Understanding

Sahil Khose(✉), Anisha Pal, Aayushi Agarwal, Deepanshi, Judy Hoffman, and Prithvijit Chattopadhyay

Georgia Institute of Technology, Atlanta, USA
{sahil.khose,apal72,judy,prithvijit3}@gatech.edu
https://huggingface.co/datasets/hoffman-lab/SkyScenes

Abstract. Real-world aerial scene understanding is limited by a lack of datasets that contain densely annotated images curated under a diverse set of conditions. Due to inherent challenges in obtaining such images in controlled real-world settings, we present SkyScenes, a synthetic dataset of densely annotated aerial images captured from Unmanned Aerial Vehicle (UAV) perspectives. We carefully curate SkyScenes images from Carla to comprehensively capture diversity across layouts (urban and rural maps), weather conditions, times of day, pitch angles and altitudes with corresponding semantic, instance and depth annotations. Through our experiments using SkyScenes, we show that (1) models trained on SkyScenes generalize well to different real-world scenarios, (2) augmenting training on real images with SkyScenes data can improve real-world performance, (3) controlled variations in SkyScenes can offer insights into how models respond to changes in viewpoint conditions (height and pitch), weather and time of day, and (4) incorporating additional sensor modalities (depth) can improve aerial scene understanding. Our dataset and associated generation code are publicly available at: https://hoffman-group.github.io/SkyScenes/

Keywords: Aerial scene understanding · Synthetic-to-Real generalization · Segmentation · Domain Generalization · Synthetic Data

1 Introduction

Aerial imagery provides a unique perspective that is invaluable for a wide range of applications, including surveillance [27,32], mapping [2,30], urban planning [12, 20], environmental monitoring [6,24], and disaster response [18,25]. These applications rely on accurate and detailed analysis of aerial images to make informed

S. Khose, A. Pal, A. Agarwal and Deepanshi—Equal Contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-72986-7_2.

A. Leonardis et al. (Eds.): ECCV 2024, LNCS 15137, pp. 19–35, 2025.
https://doi.org/10.1007/978-3-031-72986-7_2

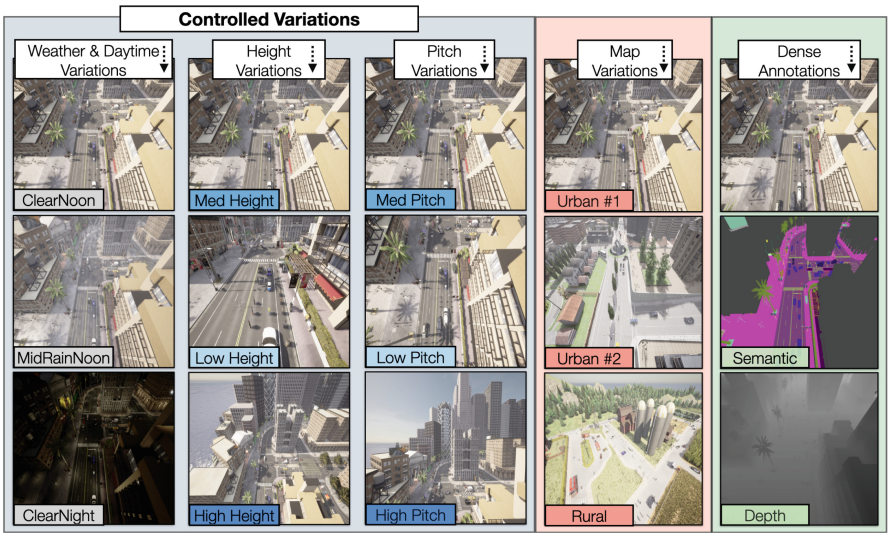


Fig. 1. SkyScenes comprises of 33.6k aerial images curated from aerial oblique viewpoints with ***controlled variations*** facilitating reproducibility of viewpoints across different weather and daytime conditions (col 1), different flying altitudes (col 2) and different viewpoint pitch angles (col 3), across different map layouts (rural and urban, col 4) with dense pixel-level semantic, instance and depth annotations (col 5).

decisions and effectively address various challenges. Naturally, training effective aerial scene-understanding models requires access to large-scale annotated exemplar data that have been *carefully curated* under diverse conditions. Capturing such images not only allows training models that can be robust to anticipated test-time variations but also allows assessing model susceptibility to changing conditions. However, carefully curating and annotating such images in the real-world can be prohibitively expensive due to various reasons. First, densely annotating every pixel of high-resolution real-world aerial images is expensive – for instance, densely annotating a single 4K image in UAVid [22] can take up to 2 hours! Second, although diversity in training data is vital for developing robust generalization algorithms and for model sensitivity assessment, expanding the real set to include widespread variations (weather, time of day, pitch, altitude) would be uncontrolled (*i.e.,* we can't guarantee the same viewpoint under different conditions as the real world is not static), and hence would require re-annotating newly captured frames. Synthetic data curated from simulators can help counter both these issues as (1) labels are automatic and cheap to obtain and (2) it is possible to recreate same viewpoint (with same scene layout and actor instances, `vehicles`, `humans`, etc. in the scene) under differing conditions.

The unique challenges introduced by outdoor aerial imagery setting such as variability in altitude and angle of image capture, skewed representation for classes with smaller object sizes (`humans`, `vehicles`), size and occlusion variations of object classes in the same image, etc., make it a relatively difficult

problem compared to ground-view imagery setting. This is illustrated by the observation that the current developments in syn-to-real generalization methods [14,45] have resulted in a significant reduction in the syn-to-real generalization gap in ground-view settings to as minimal as 20%[1], compared to the comparatively higher margin observed in aerial imagery scenarios, reaching up to 50% (see Sec. 4.3–4.5 in supplementary). Unlike synthetic ground plane view datasets (especially for autonomous driving [7,26,35,38,47]), synthetic datasets for aerial imagery (see Table 1, rows 4-10) have received relatively less attention [4,11,19,21,34,39,44]. While existing synthetic aerial imagery datasets have tried to close this gap, they are often found lacking in a few different aspects – complementary metadata to reproduce existing frame viewpoints under different conditions, limited diversity, availability of dense annotations for a wide vocabulary of classes and image capture (height, pitch) conditions (see Table 1 for an exhaustive summary). These synthetic aerial datasets rarely allow for reproducing the exact viewpoint with different variations based on detailed scene metadata (see Table 1 Controlled Variations column), a key aspect for evaluating deep learning models' responses to changing conditions and assessing sensitivity to single-variable alterations (e.g., weather, time of day, sensor angle).

We cover all these aspects by introducing SkyScenes, a synthetic dataset containing 33.6k densely annotated aerial scenes, leveraging the Carla [9] simulator to capture diverse *layout* (urban and rural), *weather*, *daytime*, *pitch* and *altitude* conditions. Our pipeline meticulously generates scenes with precise actor locations and orientations, ensuring each scene's reproducibility with a comprehensive metadata store and rigorous consistency checks.

SkyScenes encompasses detailed semantic, instance segmentation (28 classes), and depth annotations across 8 distinct map layouts across 5 different weather and daytime conditions each over a combination of 3 altitude and 4 pitch variations (see Fig. 1 for examples). While doing so, we keep several important desiderata in mind. First, we ensure that stored snapshots are not correlated, to promote diverse viewpoints within a town and facilitate model training. Second, we store all metadata associated with the position of actors, camera, and other scene elements to be able to reproduce the same viewpoints under different weather and daytime conditions. Thirdly, we ensure that the generated data mimics real-world imperfections by introducing variations in sensor locations, such as adding jitter to specified height and pitch values.[2] Finally, since Carla [9] by default does not spawn a lot of pedestrians in a scene, we propose an algorithm to ensure adequate representation of `humans` in the scene while curating images (see Sect. 3.1 and Sect. 3.2 for a detailed discussion).

Our experiments across 3 different real datasets and 3 increasingly competitive semantic segmentation architectures consistently demonstrate that SkyScenes outperforms its closest synthetic counterpart dataset, SynDrone [34]. However, despite these dataset-level improvements, the syn-to-real gener-

[1] metric of choice: mIoU.

[2] Moreover, through rigorous validations, we ensure this process is consistent and yields error-free re-generations. See Sect. 3.1.

alization gap persists, indicating that algorithmic improvements developed for ground-view imagery fail to translate effectively to aerial imagery. This underscores the urgent need for specialized algorithmic development in this area.

Empirically, we demonstrate the utility of SkyScenes in several different ways. First, we show that SkyScenes is a valuable pre-training dataset for real-world aerial scene understanding by, (1) demonstrating that models trained on SkyScenes generalize well to multiple real-world datasets and (2) demonstrating that SkyScenes pretraining improves real-world performance in low-shot regimes. Second, we show that controlled variations in SkyScenes can serve as a diagnostic test-bed to assess model sensitivity to weather, daytime, pitch, altitude, and layout conditions – by testing SkyScenes trained models in unseen SkyScenes conditions. Finally, we show that SkyScenes can enable developing multi-modal segmentation models with improved aerial-scene understanding capabilities when additional sensors, such as Depth, are available. To summarize, we make the following contributions:

- We introduce, SkyScenes, a densely-annotated dataset of 33.6k synthetic aerial images. SkyScenes contains images from different altitude and pitch settings, encompassing different layouts, weather, and daytime conditions with corresponding dense annotations and viewpoint metadata.
- We demonstrate that SkyScenes pre-trained models generalize well to real-world scenes and that SkyScenes data can effectively augment real-world training data for improved performance. We also bring attention to the point that while the synthetic-to-real gap has considerably narrowed for ground-view datasets, the same algorithms are unable to bridge this gap in aerial imagery.
- We show that our unique ability to generate controlled variations enables SkyScenes to serve as a diagnostic test-bed to assess model sensitivity to changing weather, daytime, pitch, altitude, and layout conditions.
- Finally, we show that incorporating additional modalities (depth) while training aerial scene-understanding models can improve aerial scene recognition, enabling further development of multi-modal segmentation models.

2 Related Work

Ground-view Synthetic Datasets. Real-world ground-view scene-understanding datasets (Cityscapes [7], Mapillary [26], BDD-100K [47], Dark Zurich [37]) fail to capture the full range of variations that exist in the world. Synthetic data is a popular alternative for generating diverse and bountiful views. GTAV [38], Synthia [35], and VisDA-C [31] are some of the widely-used synthetic datasets. These datasets can be curated using underlying simulators, such as GTAV [38] game engine and Carla [9] simulator and offer a cost-effective and scalable way to generate large amounts of labeled data under diverse conditions. Similar to SELMA [42] and SHIFT [41], we use Carla [9] as the underlying simulator for SkyScenes.

Table 1. SkyScenes compared with other Real and Synthetic Datasets. We compare SkyScenes (row 11) with other real (rows $1-3$) and synthetic (rows $4-10$) aerial datasets across several axes: (i) **Controlled Variations** – the ability to reproduce the exact viewpoint under different variations from fine-grained scene metadata, (ii) **Diversity** – diversity of map layouts (rural, urban), weather and daytime conditions in the provided images, (iii) **Annotation Diversity** – supporting dense annotations across depth (D), semantic(S) and instance segmentation (I) tasks, (iv) **Altitude** – altitude of image capture; Low is < 30m, Med is $\in [30, 50]$m and High is > 50m, (v) **Perspective** – UAV pitch angle during image capture; Fwd. is forward view with $\theta = 0°$, Obl. is oblique view with $\theta \in (0°, 90°)$ and Nad. is nadir view with $\theta = 90°$ (θ is pitch), (vi) **Resolution** – resolution of the images, (vii) **Scale** – number of images. We see that while existing datasets might be lacking in a subset of criteria, SkyScenes fulfills all of these.

Dataset	Controlled Variations	Diversity			Annotation Diversity	Altitude	Perspective	Resolution	Scale
		Town	Daytime	Weather					
Real									
1 UAVid [22]	✗	✗	✗	✗	S	Med	Obl.	3840 × 2160	0.42k
2 AeroScapes [28]	✗	✗	✗	✗	S	(Low, Med)	(Obl., Nad.)	1280 × 720	3.27k
3 ICG Drone [17]	✗	✗	✗	✗	S	Low	Nad.	6000 × 4000	0.6k
Synthetic									
4 Espada [21]	✗	✓	✗	✗	D	(Med, High)	Nad.	640 × 480	80k
5 UrbanScene3D [19]	✗	✓	✗	✗	-	Med	Obl.	6000 × 4000	128k
6 SynthAer [39]	✓	✗	✓	✗	S	(Low, Med)	Obl.	1280 × 720	∼ 0.77k
7 MidAir [11]	✗	✓	✓	✓	(S,D)	Low	(Obl., Nad.)	1024 × 1024	119k
8 TartanAir [44]	✗	✓	✓	✓	(S,D)	Low	(Fwd., Obl.)	640 × 480	∼ 1M
9 VALID [4]	✗	✓	✓	✓	(S,I,D)	(Low, Med, High)	Nad.	1024 × 1024	6.7k
10 SynDrone [34]	✗	✓	✗	✗	(S,D)	(Low, Med, High)	(Obl., Nad.)	1920 × 1080	72k
11 SkyScenes	✓	✓	✓	✓	(S,I,D)	(Low, Med, High)	(Fwd., Obl., Nad.)	2160 × 1440	33.6k

Real-World Aerial Datasets. To support remote sensing applications, it is crucial to have access to datasets that offer aerial-specific views. Datasets such as GID [43], DeepGlobe [8], ISPRS2D [36], and FloodNet [33] primarily provide nadir perspectives and are designed for scene-recognition and understanding tasks. However, this study specifically focuses on lower altitudes, which are more relevant to UAVs, enabling object identification. Unfortunately, there is a scarcity of high-resolution real-world datasets based on UAV imagery emphasizing object identification. xisting urban scene datasets, like Aeroscapes [28], UAVid [22], VDD [1], UDD [5], UAVDT [10], VisDrone [48], Semantic Drones [17] and others, suffer from limited sizes and a lack of diverse images under different conditions. This limitation raises concerns regarding model robustness and generalization.

Synthetic Aerial Datasets. Simulators can facilitate affordable, reliable, and quick collection of large synthetic aerial datasets, which aids in fast prototyping, improves real-world performance by enhancing robustness, and enables controlled studies on varied conditions. One such high-fidelity simulator, AirSim [40], used for development and testing of autonomous systems (in particular, aerial vehicles), is the foundation of several synthetic UAV-based datasets – MidAir [11], Espada [21], Tartan Air [44], UrbanScene3D [19] and VALID [4]. Carla [9] is another such open-source simulator that is the foundation of datasets like SynDrone [34]. However, these datasets fall short in capturing real-world irregularities, lack deterministic re-generation capabilities, controlled diversity in weather and daytime conditions, and exhibit skewed representation for certain classes (differ-

(a) SkyScenes Ground View → (Oblique) Aerial View. **(b)** Pixel proportions

Fig. 2. Ground View → (Oblique) Aerial View. (a) The same scene viewed in Ground View vs Aerial View exhibits a significant difference in pixel proportion especially across the tail classes (`vehicle`, `human`) (b) For a subset of commonly annotated classes across CityScapes [7] (red), UAVid [22] (dark blue) , we show the percentage of pixels occupied by different classes. Aerial scenes (in UAVid) have significant under-representation of tail classes (`vehicle`, `human`).

ences summarized in Table1). This restricts their ability to generalize well to real-world datasets and their usage as a diagnostic tool for studying the controlled effect of diversity on the performance of computer vision perception tasks. To enable such studies, SkyScenes offers images featuring varied scenes, diverse weather, daytime, altitude, and pitch variations while incorporating real-world irregularities and addressing skewed class representation along with simultaneous depth, semantic, and instance segmentation annotations.

3 SkyScenes

We curate SkyScenes using, Carla [9][3] 0.9.14, which is a flexible and realistic open-source autonomous vehicle simulator. The simulator offers a wide range of sensors, environmental configurations, and varying rendering configurations. As noted earlier, we take several important considerations into account while curating SkyScenes images. These include strategies for obtaining diverse synthetic data and embedding real-world irregularities, avoiding correlated images, addressing skewed class representations, and more. In this section, we first discuss such desiderata and then describe our procedural image curation algorithm. Finally, we describe different aspects of the curated dataset.

3.1 (Synthetic) Aerial Image Desiderata

Before investigating the image curation pipeline, we first outline a set of desiderata taken into account while curating synthetic aerial images in SkyScenes.

1. **Viewpoint Reproducibility:** Critical to understanding how models respond to changing conditions is the ability to evaluate them under scenarios where only one variable is altered. However, any effort to do so in the

[3] https://carla.org/.

real-world would be uncontrolled, due to its dynamic (constantly changing) nature. In contrast, simulated data allows us to do so by providing control over image generation conditions. Unlike certain existing aerial datasets that do not support this feature (see Table. 1), we do so in SkyScenes by additionally storing comprehensive metadata for each viewpoint (and image), including details about camera world coordinates, orientation, and all movable/immovable actors and objects in the scene. We couple this with rigorous consistency checks for image generation that verify the number of actors, their location, sensor height, pitch, etc. This meticulous approach enables us to reproduce the same viewpoint under multiple conditions effortlessly.

2 . **Adequate Representation of Tail Classes:** Unlike ground-view datasets, pixel distribution of classes in aerial images is substantially more long-tailed (see Fig. 2 (a); classes with smaller object size, `humans`). This substantial difference in class proportions severely affects the performance of tail classes in aerial datasets when compared to ground-view datasets (see Fig. 2 (b)), thus making visual recognition tasks harder. To counter this, we consider structured spawning of `humans` to ensure adequate representation (see Sect. 3.2).

3. **Adequate Height Variations:** Aerial images are captured at different altitudes to meet specific needs. Lower altitudes (5–15 m) are optimal for high-resolution photography and detailed inspections. Altitudes ranging from 30 m–50 m strike a balance between fine-grained detail and a broader perspective, making them ideal for surveillance. Altitudes above 50m are suitable for capturing extensive areas, making them ideal for surveying and mapping. Existing datasets (synthetic or real) often focus on "specific" altitude ranges (see Table 1, Image Capture columns), limiting their adaptability to different scenarios. With SkyScenes, our aim is to provide flexibility in altitude sampling, thus accommodating various real-world requirements. We curate SkyScenes images at heights of 15m, 35m, and 60m. Additionally, recognizing imperfections in real-world actuation, we induce slight jitter in the height values ($\Delta h \sim \mathcal{N}(1, 2.5\text{m})$) to simulate realistic data sampling.

4. **Adequate Pitch Variations:** Similar to height, aerial images can be captured from 3 primary perspectives or pitch angles (θ): nadir ($\theta = 90°$), oblique ($\theta \in (0°, 90°)$), or forward ($\theta = 0°$) views (see Table. 1, Image Capture columns). The nadir view (directly perpendicular to the ground plane), preserves object scale while forward views are well-suited for tasks like UAV navigation and obstacle detection. Oblique views, on the other hand, capture objects from a side profile, aiding object recognition and providing valuable context and depth perspective often lost in nadir and forward views. To ensure widespread utility, SkyScenes data generation process is designed to support all these viewing angles, with a particular emphasis on oblique views (the most common one). Similar to height, pitch variations allow models trained on SkyScenes to generalize to different viewpoint variations. We use $\theta = 45°$ and $60°$ for oblique-views and introduce jitter ($\Delta\theta \sim \mathcal{N}(1, 5°)$) to mimic real-world data sampling.

5. **Adequate Map Variations:** In addition to sensor locations, it is equally important to curate aerial images across diverse scene layouts. To ensure

adequate map variations, we gather images from 8 different CARLA [9] towns (can be categorized as *urban* or *rural*), which provide substantial variations in the observed scene. These towns differ in layouts, size, road map design, building design, and vegetation cover. Figure 4 illustrates how images curated from different towns in CARLA [9] differ in class distributions.

6. **Adequate Weather & Daytime Variations:** Training robust perception models using SKYSCENES that generalize to unforeseen environmental conditions necessitates the curation of annotated images encompassing various weather and daytime scenarios. To accomplish this, we generate SKYSCENES images from identical viewpoints under 5 different variations – ClearNoon, ClearSunset, MidRainNoon, ClearNight, and CloudyNoon.[4] Generating images in different conditions from the same perspectives allows us to (1) leverage diverse data for improved generalization and (2) systematically investigate the susceptibility of trained models to variations in daytime and weather conditions.
7. **Fine-grained Annotations:** To support a host of different computer vision tasks (segmentation, detection, multimodal recognition), we curate all SKYSCENES images with dense semantic, instance segmentation and depth annotations. We provide semantic annotations for a wide vocabulary of 28 classes to support broad applicability (see Fig. 1 column 4 for an example).

3.2 SKYSCENES Image Generation

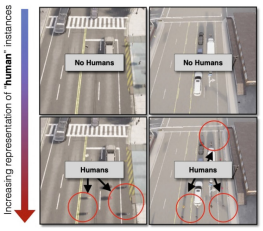
We generate SKYSCENES images from CARLA [9] by taking the previously mentioned considerations into account. Curating images from CARLA [9] broadly consists of two key steps: (1) positioning the agent camera in an aerial perspective and (2) procedurally guiding the agent within the scene to capture images. We accomplish the first by mimicking a UAV perspective in CARLA [9] by positioning the ego vehicle (with RGB, semantic and depth sensors) based on specified (high) altitude (h) and pitch (θ) values to generate aerial views (see Fig. 2a).[5] Once positioned, the agent is translated by fixed amounts to traverse the scene and capture images from various viewpoints (detailed in Sec. 2.1 in the supplementary). Initially, we generate 70 datapoints for each of the 8 town variations under ClearNoon conditions using the baseline $h = 35\,\text{m}$, $\theta = 45°$ setting. Subsequently, following the traversal algorithm (see Sec. 2.1 in the supplementary), we re-generate these datapoints across 5 weather conditions and 12 height/pitch variations, resulting in $70 \times 8 \times 5 \times 12 = 33,600$ images.

Checks and Balances. Additionally, we ensure the following checks and balances while curating SKYSCENES images.

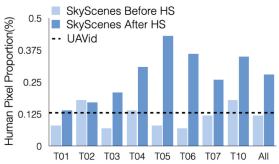
▷ **Avoiding Overly Correlated Frames for Viewpoints.** CARLA [9] uses a traffic manager with a PID controller to control the egocentric vehicle based on

[4] Note that CARLA [9] provides 14 such conditions but we use only 5 such conditions in this preliminary version of SKYSCENES.

[5] This also requires setting other scenes – weather, daytime, etc. – and camera (notably the `FoV` $= 110°$ (field of view) and image resolution $= 2160 \times 1440$) parameters.



(a) HumanSpawn() (HS)



(b) Human Pixel Proportion

Eval Data	mIoU(↑) human	All
SkyScenes		
1 Before HS	43.03	80.87
2 After HS	**61.79**	**84.07**
SkyScenes → UAVid		
3 Before HS	4.71	45.11
4 After HS	**10.21**	**47.09**

(c) mIoU improvement w/ HS

Fig. 3. SkyScenes w/ HumanSpawn() increases representation of humans and improves SkyScenes→UAVid(S→U) performance. (a) Incorporating HumanSpawn() in the image generation pipeline for SkyScenes increases the proportion of **humans** in snapshots (**[Top]→[Bottom]**). (b) Increased representation of **humans** across all the layout variations in SkyScenes after HumanSpawn(), with the dotted line representing the proportion of **humans** in UAVid (c) Training on HumanSpawn (HS) SkyScenes images improves the model's ability to recognize **humans** (improved mIoU). T = Town.

current pose, speed, and a list of waypoints at every pre-defined time step. Curating images at every time step (or tick) results in highly correlated frames with little change in object position. Since overly correlated frames are not very useful when training models for static scene understanding, we move the camera by a fixed distance multiple times before saving a frame. This also helps with moving dynamic actors by a considerable amount in the scene. Additionally, pedestrian objects are regenerated before saving an image, which adds randomness to the spawning and placement of pedestrians, reducing the correlation between frames.

▷ **Adequate Representation of humans.** Real-world scenes often exhibit a long-tailed distribution in pixel proportions, particularly in aerial images where variations in object sizes and camera positions contribute to significant under-representation of the tail classes (in Fig. 2, for the shared set of classes across UAVid [22] (aerial) and Cityscapes [7] (ground), we can see that the class distributions are different and aerial images are significantly more heavy-tailed). As a result, naively spawning **humans** (rarest class) in Carla [9] is detrimental for eventual task performance – for the **human** class, a SkyScenes trained DAFormer [14] (with HRDA [15] source training; MiT-B5 [46] backbone) model leads to an in-distribution performance of 43.03 mIoU and out-of-distribution (SkyScenes →UAVid [22]) performance of 4.71 mIoU. To counter this under-representation issue, we design an algorithm, HumanSpawn() (see Sec. 2.1 in supplementary), to explicitly spawn more **human** instances while curating SkyScenes images. HumanSpawn() increases **human** instances by $40 - 200$ per snapshot, improving the proportion of densely annotated **humans** in SkyScenes by approximately 10 times (see Fig. 3 (a) & Fig. 3 (b)). This improvement in **human** representation is also evident in eventual task performance, with in-distribution and out-of-distribution mIoUs for **humans** increasing from 43.03 to 61.79 (+18.76) and 4.71 to 10.21 (+5.50) respectively (see Table. 3 (c)).

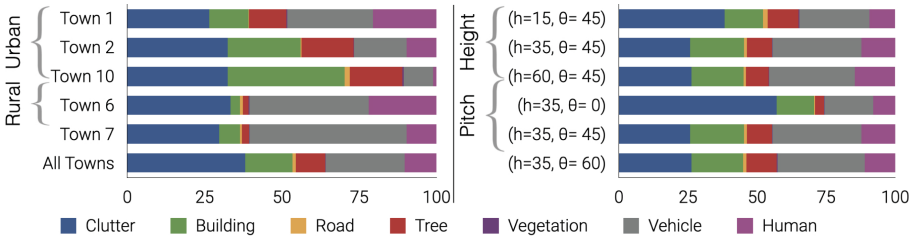


Fig. 4. Class-distribution Diversity in SkyScenes. We show how the distribution of densely-annotated pixels varies across different SkyScenes conditions. **[Left]** Class distribution varies substantially within and across urban and rural map layouts. **[Right]** Similarly, for the same SkyScenes layouts (and viewpoints) class distribution varies substantially across different height and pitch values.

3.3 SkyScenes: Dataset Details

Annotations. We provide semantic, instance and depth annotations for every image in SkyScenes. Semantic annotations in SkyScenes by default are across 28 classes. These are building, fence, pedestrian, pole, roadline (markings on road), road, sidewalk, vegetation, cars, wall, traffic sign, sky, bridge, railtrack, guardrail, traffic light, water, terrain, rider, bicycle, motorcycle, bus, truck and others (see Fig. 1 for an example and Sec. 2.2 in supplementary for definitions)

Training, Validation and Test Splits. SkyScenes has 70 images per town (across 8 towns) for each of the 5 weather and daytime conditions, and 12 height & pitch combinations, resulting in a total of 33,600 images. We use 80% (26,880 images) of the dataset for training models, with 10% (3,360 images) each for validation and testing (see Sec. 3.3 in supplementary).

Class Distribution(s). In Fig. 4, we highlight how the distribution of classes changes across variations within SkyScenes– rural and urban map layouts and height and pitch specifications. SkyScenes exhibits substantial diversity in class distributions across such conditions, allowing these individual conditions to serve as diagnostic splits to assess model sensitivity (see Sect. 4.2).

4 Experiments

We conduct semantic segmentation experiments with SkyScenes to assess a few different factors. First, we check if training on SkyScenes is beneficial for real-world transfer. Second, we check if SkyScenes can augment real-world training data in low and full shot regimes. Third, we check if variations in SkyScenes can be used to assess the sensitivity of trained models to changing conditions. Finally, we check if using additional modality information (depth) can help improve aerial scene understanding.

Synthetic and Real Datasets. We compare real-world generalization performance of training on SkyScenes with SynDrone [34], a recently proposed

Table 2. Models trained on SkyScenes generalize well to the real-world. We train semantic segmentation models (DeepLabv2 [3], DAFormer [14], Rein [45]) on SkyScenes, SynDrone [34], and real datasets and show how training models on SkyScenes provides better out-of-the-box generalization to multiple real-world datasets.

Source	(Target) Real-World mIoU (↑)		
	UAVid	AeroScapes	ICG Drone
DeepLabv2 (R-101) [3]			
1 SynDrone	39.86	24.50	8.20
2 SkyScenes	**41.82**	**26.94**	**15.14**
DAFormer (MiT-B5) [15]			
3 SynDrone	42.31	30.53	15.92
4 SkyScenes	**47.09**	**40.72**	**25.91**
Rein (DINOv2)			
5 SynDrone	**54.92**	40.28	20.01
6 SkyScenes	54.19	**43.96**	**28.10**

Table 3. SkyScenes training exhibits strong real-world generalization for tail classes. We show how DAFormer [14] and Rein [45] models trained on SkyScenes exhibit improved real-world generalization compared to those trained on SynDrone [34] for under-represented tail classes (`vehicles` and `humans`). SkyScenes training facilitates better recognition of tail class instances.

Source	(Target) Real-World IoU (↑)					
	UAVid		AeroScapes		ICG Drone	
	vehicle	human	vehicle	person	vehicle	person
DAFormer (MiT-B5) [15]						
1 SynDrone	42.52	8.27	49.77	0.77	0.24	0.38
2 SkyScenes	**63.64**	**10.21**	**80.99**	**3.09**	**39.71**	**45.89**
Rein (DINOv2) [45]						
3 SynDrone	68.68	21.6	84.2	10.29	7.91	0
4 SkyScenes	**75.14**	**25.52**	**87.71**	**21.67**	**50.91**	**77.93**

synthetic aerial dataset also curated from Carla [9] featuring 3 different (h, θ) conditions across 8 different map layouts. We assess performance on 3 real-world aerial datasets – UAVid [22], AeroScapes [28], ICG Drone [17]. Since different datasets have different class vocabularies and definitions, for our experiments, we adapt the class vocabulary of the synthetic source dataset to that of the target real-world datasets (see Sec. 3.1 in supplementary for class merging and assignment schemes). Additionally, since different real aerial datasets have been captured from different heights and pitch angles, we train models on (h, θ) subsets of synthetic datasets that are aligned with corresponding real data (h, θ) conditions. We provide additional details for the real aligned synthetic data selection and model evaluation in Sec. 4 in the supplementary.

Models. We use (1) CNN – DeepLabv2 [3] (ResNet-101 [13]), (2) transformer – DAFormer [14] (with HRDA [15] source training; MiT-B5 [46] backbone) and (3) Vision Foundation Model – Rein [45] (LoRA [16] fine-tuned Dino-V2 [29] backbone) based semantic segmentation architectures for our experiments. We provide implementation details for our experiments in Sec. 3.5 in supplementary (Fig. 5).

4.1 SkyScenes → Real Transfer

▷ **SkyScenes Trained Models Generalize Well to Real-Settings.** In Table 2, we show how models trained on SkyScenes exhibit strong out-of-the box generalization performance on multiple real world datasets. We find that SkyScenes pretraining exhibits stronger generalization compared to SynDrone [34] across both CNN and transformer segmentation backbones. In

Table 4. SkyScenes augmented real data improves performance in low shot regimes. We compare DeepLabv2 [3], DAFormer [14], and Rein [45] models trained using varying percentages of labeled UAVid [22] images. Models are either trained jointly on SkyScenes and UAVid (JT) or pretrained on SkyScenes and finetuned on UAVid (FT). Augmenting real data with SkyScenes enhances real-world generalization in low-shot scenarios.

Source	(Target) Real World mIoU (↑)				
	5%	10%	25%	50%	100%
DeepLabv2 (R-101) [3]					
1 Only Real	48.25	55.29	62.86	66.81	68.53
2 SkyScenes + Real (JT)	**59.27**	**64.15**	**68.11**	**70.18**	69.51
3 SkyScenes + Real (FT)	53.67	60.61	65.57	68.54	**69.70**
DAFormer (MiT-B5) [15]					
4 Only Real	60.59	65.63	70.31	72.16	72.47
5 SkyScenes + Real (JT)	**62.97**	**67.58**	70.20	71.83	72.25
6 SkyScenes + Real (FT)	60.90	66.79	**70.41**	**72.63**	**73.02**
Rein (DINOv2) [45]					
7 Only Real	64.04	71.87	73.87	76.05	76.55
8 SkyScenes + Real (JT)	69.15	73.54	**75.07**	76.08	76.54
9 SkyScenes + Real (FT)	**70.07**	**73.99**	75.01	**76.44**	**76.89**

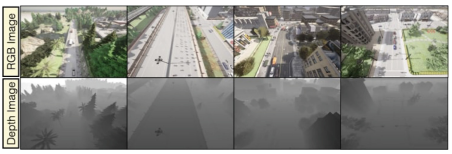


Fig. 5. SkyScenes RGB Images and corresponding depth images generated using depth sensor for $h = 35$, $\theta = 45°$, and ClearNoon setting across four different town layouts.

Table 5. Multi-modal Segmentation in SkyScenesbf . We evaluate M3L [23] multimodal segmentation architectures with MiT-B5 [46] backbones using RGB and RGB+D data in SkyScenes. Additional sensors improve aerial scene understanding significantly across various classes in UAVid [22].

Sensors	SkyScenes Test IoU (↑)							
	clutter	building	road	tree	low-veg.	vehicle	human	Avg
1 RGB	87.80	94.54	94.07	88.03	69.37	82.89	43.35	80.01
2 RGB+D	**90.64**	**95.97**	**94.87**	**89.41**	**74.36**	**86.87**	**50.47**	**83.22**

Table 3, we show how generalization improvements are more pronounced for under-represented tail classes (`vehicles` and `humans`). Comparison across all classes is provided in Tables 5–7 in supplementary.

▷ **SkyScenes Can Augment Real Training Data.** In addition to zero-shot real-world generalization, akin to other synthetic aerial datasets, we also show how SkyScenes is useful as additional training data when labeled real-world data is available. In Table 4, for SkyScenes →UAVid [22], we compare models trained only using $5\%, 10\%, 25\%$, 50%, 100% of the 200 UAVid [22] training images with counterparts that were either pretrained using SkyScenes data or additionally supplemented with SkyScenes data at training time. We find that in low-shot regimes (when little "real" world data is available), SkyScenes data (either explicitly via joint training or implicitly via finetuning) is beneficial in improving recognition performance (see Sec. 4.5 of supplementary).

4.2 SkyScenes as a Diagnostic Framework

As noted earlier, the images we curate in SkyScenes contain several variations – ranging from 5 different weather and daytime conditions, rural and urban map layouts, and 12 different height and pitch combinations (see Fig. 4 for variations in class distributions). We curate images under such diverse conditions in a *controlled manner* – ensuring the same spatial coordinates for (h, θ) variations, same spatial coordinates and (h, θ) settings across different weather and daytime conditions, the same number of images across layouts.

Table 6. Model Sensitivity to Changing Conditions. We show how changing conditions (weather, daytime, map, viewpoint) in SkyScenes can serve as diagnostic test splits to assess the sensitivity of trained DAFormer [14] semantic segmentation models. In (a) and (b), we evaluate models trained under different weather and daytime conditions across the same conditions. In (c), we evaluate models trained on rural and urban scenes across the same layouts. In (d), we evaluate a model trained on moderate height, pitch settings ($h = 35\text{m}, \theta = 45°$) across different h, θ variations. Best numbers across each row condition is highlighted in blue.

Train	Test mIoU (↑)		
	Clear	Cloudy	Rainy
1 Clear	73.91	73.59	69.95
2 Cloudy	69.60	74.02	69.14
3 Rainy	69.00	73.36	72.62

(a) Weather Variation

Train	**Test mIoU** (↑)		
	Noon	Sunset	Night
1 Noon	73.91	71.16	35.60
2 Sunset	63.16	66.53	39.36
3 Night	52.00	57.35	70.36

(b) Daytime Variation

Train	Test mIoU (↑)	
	Rural	Urban
1 Rural	58.00	35.90
2 Urban	38.99	73.16

(c) Map Variation

Height	Test mIoU (↑)			
	Pitch			
	$\theta = 0°$	$\theta = 45°$	$\theta = 60°$	$\theta = 90°$
1 $h = 15\,\text{m}$	48.50	50.71	45.22	42.21
2 $h = 35\,\text{m}$	50.49	55.74	57.11	52.19
3 $h = 60\,\text{m}$	45.33	49.79	50.37	44.62

(d) Height & Pitch Variation

This allows us to assess the sensitivity of trained models to one factor of variation (h, θ, daytime, weather, map layout) by changing that specific aspect. We summarize some takeaways from such experiments in Table 6.

In Table 6 (a), we show how models trained in a certain weather condition are best at generalizing to the same condition at test-time. We make similar observations for daytime variations in Table 6 (b). In Table 6 (c), we show how models trained in rural conditions fail to perform well in urban test-time conditions and vice-versa. In Table 6 (d), we evaluate a model trained under moderate ($h = 35\text{m}, \theta = 45°$) conditions under different (h, θ) variations. We find that as altitudes increase, trained models are better at recognizing objects from oblique ($\theta \in (0°, 90°)$) viewpoints. We provide exhaustive quantitative comparisons in Sec. 4.6 in the supplementary (Fig. 6).

4.3 SkyScenes Enables Multi-modal Dense Prediction

Sensors on UAVs in deployable settings often include modalities beyond RGB cameras, such as depth sensors. These additional modalities can significantly enhance aerial scene understanding. In Table. 5, we investigate the impact of augmenting RGB data with depth observations from SkyScenes viewpoints on aerial semantic segmentation using M3L [23], a multimodal segmentation model. Similar to our DAFormer [14] experiments, we consider a SegFormer equivalent version of M3L [23] (with an MiT-B5 [46] backbone). We test RGB and RGB+D models trained under ($h = 35, \theta = 45°$) (moderate viewpoint) conditions on SkyScenes and find that incorporating additional Depth observations can substantially improve recognition performance. This demonstrates that images in SkyScenes can be used to train multimodal scene-recognition models.

5 Conclusion

We introduce SkyScenes, a large-scale densely-annotated dataset of synthetic aerial scene images curated from unmanned aerial vehicle (UAV) perspectives.

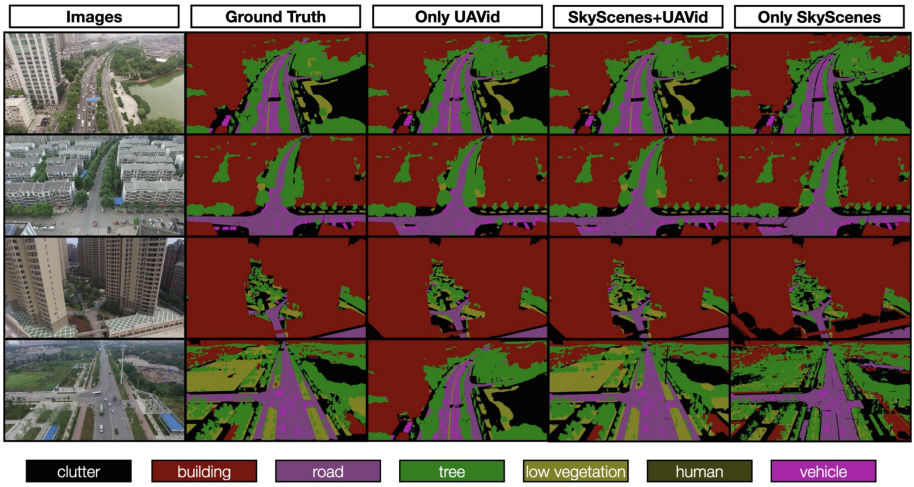


Fig. 6. UAVid, SkyScenes + UAVid and SkyScenes → UAVid semantic segmentation predictions Predictions on randomly selected UAVid [22] validation images by a Rein [45] model trained on UAVid and SkyScenes. Columns 1 and 2 show the original image and its ground truth. Columns 3, 4, and 5 display predictions from models trained exclusively on UAVid, jointly on SkyScenes and UAVid, and exclusively on SkyScenes, respectively.

SkyScenes images are generated using Carla by situating an agent aerially and procedurally tele-operating it through the scene to capture frames with semantic, instance, and depth annotations. Our careful curation process ensures that SkyScenes images span across diverse weather, daytime, map, height, and pitch conditions, with accompanying metadata that enables reproducing the same viewpoint (spatial coordinates and perspective) under differing conditions.

Through our experiments, we demonstrate that: (1) SkyScenes-trained models generalize well to real-world settings, (2) SkyScenes augments labeled real-world data in low-shot scenarios, (3) SkyScenes serves as a diagnostic tool for assessing model sensitivity to varied conditions, and (4) incorporating additional sensors like depth enhances multi-modal aerial scene understanding.

We aim to enhance SkyScenes with improved realism, additional anticipated edge cases, and support for 3D perception tasks aligning with advancements in our simulator (additional details in Sec. 6 of supplementary) We have publicly released the dataset and associated generation code and hope that our experimental findings encourage further research using SkyScenes for aerial scenes.

Acknowledgements. We would like to thank Sean Foley for his contributions to the early efforts and discussions of this project. This work has been partially sponsored

by NASA University Leadership Initiative (ULI) #80NSSC20M0161, ARL, and NSF #2144194.

References

1. Cai, W., Jin, K., Hou, J., Guo, C., Wu, L., Yang, W.: Vdd: varied drone dataset for semantic segmentation (2023)
2. Chauhan, A., et al.: Chapter 10 - earth observation applications for urban mapping and monitoring: research prospects, opportunities and challenges. In: Kumar, A., Srivastava, P.K., Saikia, P., Mall, R.K. (eds.) Earth Observation in Urban Monitoring, pp. 197–229. Earth Observation, Elsevier (2024). https://doi.org/10.1016/B978-0-323-99164-3.00007-0. https://www.sciencedirect.com/science/article/pii/B9780323991643000070
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs (2017)
4. Chen, L., Liu, F., Zhao, Y., Wang, W., Yuan, X., Zhu, J.: Valid: a comprehensive virtual aerial image dataset. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 2009–2016 (2020). https://doi.org/10.1109/ICRA40945.2020.9197186
5. Chen, Yu., Wang, Y., Lu, P., Chen, Y., Wang, G.: Large-scale structure from motion with semantic constraints of aerial images. In: Lai, J.H., et al. (eds.) PRCV 2018. LNCS, vol. 11256, pp. 347–359. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-03398-9_30
6. Chiang, C.Y., Barnes, C., Angelov, P., Jiang, R.: Deep learning-based automated forest health diagnosis from aerial images. IEEE Access **8**, 144064–144076 (2020). https://doi.org/10.1109/ACCESS.2020.3012417
7. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
8. Demir, I., et al.: Deepglobe 2018: a challenge to parse the earth through satellite images. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 172–172 (2018)
9. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: an open urban driving simulator. In: Conference on Robot Learning, pp. 1–16. PMLR (2017)
10. Du, D., et al.: The unmanned aerial vehicle benchmark: object detection and tracking. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 370–386 (2018)
11. Fonder, M., Droogenbroeck, M.V.: Mid-air: a multi-modal dataset for extremely low altitude drone flights. In: Conference on Computer Vision and Pattern Recognition Workshop (CVPRW) (2019)
12. Frueh, C., Sammon, R., Zakhor, A.: Automated texture mapping of 3d city models with oblique aerial imagery. In: Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004, pp. 396–403 (2004). https://doi.org/10.1109/TDPVT.2004.1335266
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
14. Hoyer, L., Dai, D., Gool, L.V.: Daformer: improving network architectures and training strategies for domain-adaptive semantic segmentation (2022)

15. Hoyer, L., Dai, D., Gool, L.V.: Hrda: context-aware high-resolution domain-adaptive semantic segmentation (2022)
16. Hu, E.J., et al.: Lora: low-rank adaptation of large language models (2021)
17. Institute of Computer Graphics and Vision, Graz University of Technology: Semantic drone dataset. http://dronedataset.icg.tugraz.at
18. Kedys, J., Tchappi, I., Najjar, A.: Uavs for disaster management - an exploratory review. Procedia Comput. Sci. **231**, 129–136 (2024). https://doi.org/10.1016/j.procs.2023.12.184. https://www.sciencedirect.com/science/article/pii/S1877050923021968
19. Lin, L., Liu, Y., Hu, Y., Yan, X., Xie, K., Huang, H.: Capturing, reconstructing, and simulating: the urbanscene3d dataset. In: ECCV 2022, pp. 93–109. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-20074-8_6
20. Liu, T., Yang, X.: Monitoring land changes in an urban area using satellite imagery, gis and landscape metrics. Appl. Geogr. **56**, 42–54 (2015). https://doi.org/10.1016/j.apgeog.2014.10.002. https://www.sciencedirect.com/science/article/pii/S0143622814002306
21. Lopez-Campos, R., Martinez-Carranza, J.: Espada: extended synthetic and photogrammetric aerial-image dataset. IEEE Rob. Autom. Lett. **6**(4), 7981–7988 (2021). https://doi.org/10.1109/LRA.2021.3101879
22. Lyu, Y., Vosselman, G., Xia, G.S., Yilmaz, A., Yang, M.Y.: Uavid: a semantic segmentation dataset for UAV imagery. ISPRS J. Photogramm. Remote. Sens. **165**, 108–119 (2020)
23. Maheshwari, H., Liu, Y.C., Kira, Z.: Missing modality robustness in semi-supervised multi-modal semantic segmentation (2023)
24. Morgan, G.R., Wang, C., Li, Z., Schill, S.R., Morgan, D.R.: Deep learning of high-resolution aerial imagery for coastal marsh change detection: a comparative study. ISPRS Int. J. Geo-Inf. **11**(2) (2022). https://doi.org/10.3390/ijgi11020100. https://www.mdpi.com/2220-9964/11/2/100
25. Munawar, H.S., Ullah, F., Qayyum, S., Khan, S.I., Mojtahedi, M.: Uavs in disaster management: application of integrated aerial imagery and convolutional neural network for flood detection. Sustainability **13**(14) (2021). https://doi.org/10.3390/su13147547. https://www.mdpi.com/2071-1050/13/14/7547
26. Neuhold, G., Ollmann, T., Bulò, S.R., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 5000–5009 (2017). https://doi.org/10.1109/ICCV.2017.534
27. Nguyen, K., et al.: The state of aerial surveillance: a survey (2022)
28. Nigam, I., Huang, C., Ramanan, D.: Ensemble knowledge transfer for semantic segmentation. In: Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision, pp. 916–924. IEEE (2018)
29. Oquab, M., et al.: Dinov2: learning robust visual features without supervision (2024)
30. Otal, H.T., Zavar, E., Binder, S.B., Greer, A., Canbaz, M.A.: Harnessing deep learning and satellite imagery for post-buyout land cover mapping (2024)
31. Peng, X., et al.: Visda: the visual domain adaptation challenge. In: IEEE International Conference on Computer Vision, pp. 1685–1692 (2017)
32. Prokaj, J., Medioni, G.: Persistent tracking for wide area aerial surveillance. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)

33. Rahnemoonfar, M., Chowdhury, T., Sarkar, A., Varshney, D., Yari, M., Murphy, R.: Floodnet: a high resolution aerial imagery dataset for post flood scene understanding (2020)
34. Rizzoli, G., Barbato, F., Caligiuri, M., Zanuttigh, P.: Syndrone–multi-modal uav dataset for urban scenarios. arXiv preprint arXiv:2308.10491 (2023)
35. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3234–3243 (2016)
36. Rottensteiner, F., et al.: The isprs benchmark on urban object classification and 3d building reconstruction. ISPRS Ann. Photogram. Remote Sens. Spatial Inf. Sci. **I-3** (2012).https://doi.org/10.5194/isprsannals-I-3-293-2012
37. Sakaridis, C., Dai, D., Van Gool, L.: Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
38. Sankaranarayanan, S., Balaji, Y., Castillo, C.D., Chellappa, R.: Generate to adapt: aligning domains using generative adversarial networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
39. Scanlon, M.: Semantic Annotation of Aerial Images using Deep Learning, Transfer Learning, and Synthetic Training Data. Ph.D. thesis, University of Galway (09 2018)
40. Shah, S., Dey, D., Lovett, C., Kapoor, A.: Airsim: high-fidelity visual and physical simulation for autonomous vehicles. In: Field and Service Robotics (2017). https://arxiv.org/abs/1705.05065
41. Sun, T., et al.: Shift: a synthetic driving dataset for continuous multi-task domain adaptation (2022)
42. Testolina, P., Barbato, F., Michieli, U., Giordani, M., Zanuttigh, P., Zorzi, M.: Selma: semantic large-scale multimodal acquisitions in variable weather, daytime and viewpoints (2022)
43. Tong, X.Y., et al.: Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sens. Environ. **237**, 111322 (2020)
44. Wang, W., et al.: Tartanair: a dataset to push the limits of visual slam (2020). https://arxiv.org/abs/2003.14338
45. Wei, Z., et al.: Stronger, fewer, & superior: harnessing vision foundation models for domain generalized semantic segmentation (2024)
46. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: simple and efficient design for semantic segmentation with transformers. Adv. Neural. Inf. Process. Syst. **34**, 12077–12090 (2021)
47. Yu, F., et al.: Bdd100k: a diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2636–2645 (2020)
48. Zhu, P., et al.: Detection and tracking meet drones challenge. IEEE Trans. Pattern Anal. Mach. Intell. (2021). https://doi.org/10.1109/TPAMI.2021.3119563