SKYSCENES: A Synthetic Dataset for Aerial Scene Understanding

Sahil Khose* Anisha Pal* Aayushi Agarwal* Deepanshi Deepanshi*
Judy Hoffman Prithvijit Chattopadhyay
Georgia Tech
{sahil.khose,anisha_pal,prithvijit3,judy}@gatech.edu
{aayushi.agarwal007,deepanshi.asr.21}@gmail.com

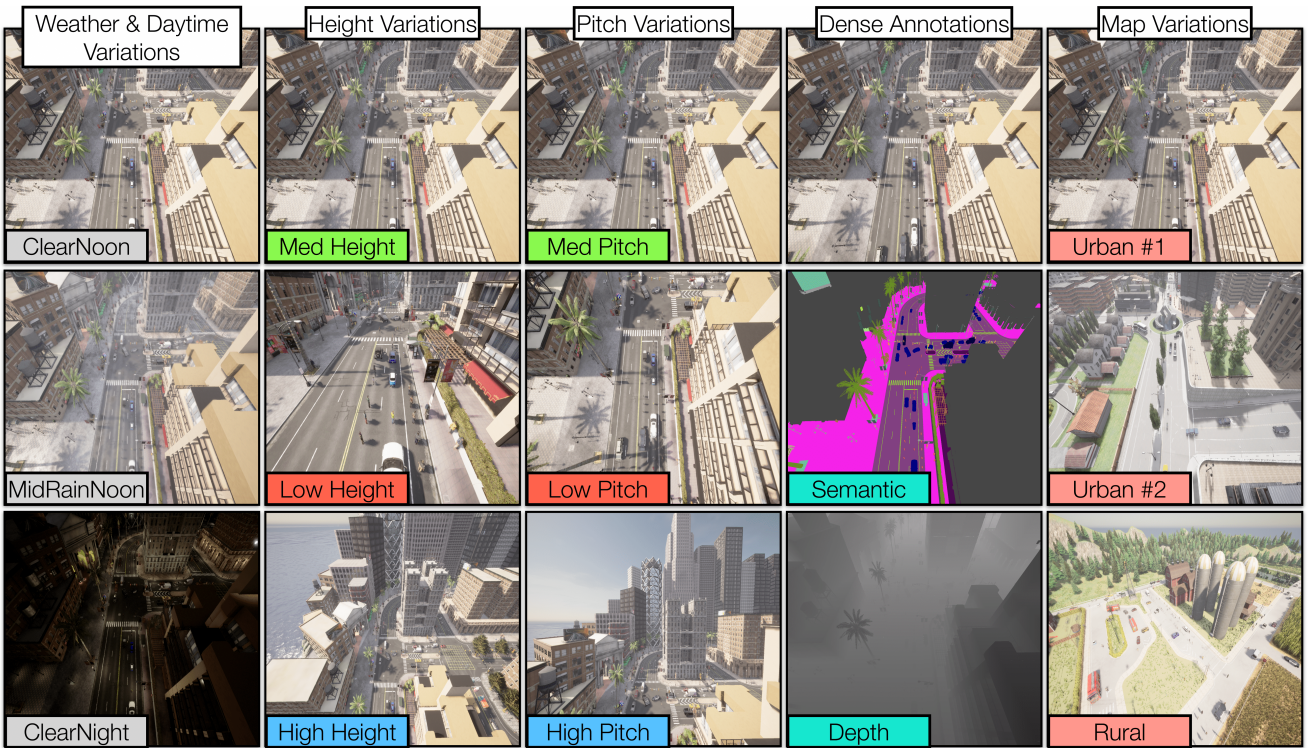

Figure 1. **Overview.** SKYSCENES comprises of 33.6K aerial images curated from UAV perspectives under different weather and daytime conditions (col 1), different flying altitudes (col 2), different viewpoint pitch angles (col 3), different map layouts (rural and urban, col 5) with supporting dense pixel level semantic, and depth annotations (col 4). SKYSCENES not only serves as a synthetic source dataset to train real-world generalizable models, but can also augment real data for improved real-world performance.

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 layout (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, and (4) incorporating additional sensor modalities (depth) can improve aerial scene understanding. We plan on publicly releasing the dataset and associated generation code.*

Dataset	Reproducibility		Diversity			Annotations			Image Capture			Scale
	Metadata	Contr. Var.	Town	Daytime	Weather	Semantic	Instance	Depth	Altitude	Perspective	Resolution	
Real												
1 UAVid [16]	✗	✗	✓	✗	✗	✓	✗	✗	Med	Obl.	3840 × 2160	0.42K
2 AeroScapes [19]	✗	✗	✗	✗	✗	✓	✗	✗	(Low, Med)	(Obl., Nad.)	1280 × 720	3.27K
3 ICG Drone [13]	✗	✗	✗	✗	✗	✓	✗	✗	Low	Nad.	6000 × 4000	0.6K
Synthetic												
4 MidAir [9]	✓	partial	✓	✓	✓	✓	✗	✓	Low	(Obl., Nad.)	1024 × 1024	119K
5 VALID [2]	✗	✗	✓	✓	✓	✓	✓	✓	(Low, Med, High)	Nad.	1024 × 1024	6.7K
6 Espada [15]	✗	✗	✓	✗	✗	✗	✗	✓	(Med, High)	Nad.	640 × 480	80K
7 TartanAir [32]	✓	*	✓	✓	✓	✓	✗	✓	Low	(Fwd., Obl.)	640 × 480	∼ 1M
8 UrbanScene3D [14]	*	*	✓	*	*	✗	✗	✗	Med	Obl.	6000 × 4000	128K
9 SynthAer [27]	✓	✓	✗	✓	✗	✓	✗	✗	(Low, Med)	Obl.	1280 × 720	∼ 0.77K
10 SynDrone [22]	✓	✗	✓	✗	✗	✓	✗	✓	(Low, Med, High)	(Obl., Nad.)	1920 × 1080	72K
11 SkyScenes	✓	✓	✓	✓	✓	✓	✓	✓	(Low, Med, High)	(Fwd., Obl., Nad.)	2160 × 1440	33.6K

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) **Reproducibility** – the ability to reproduce the same exact viewpoint under different *Controlled Variations (Contr. Var.)* from fine-grained scene *Metadata*, (ii) **Diversity** – diversity of map layouts (rural, urban), weather and daytime conditions in the provided images, (iii) **Annotations** – supporting dense annotations for images, (iv) **Image Capture** – conditions under which images are captured, and (v) **Scale** – number of images. We see that while existing datasets might be lacking in a subset of criteria, SkyScenes fulfills all of these. In reference to altitude, Low is < 30m, Med is $\in [30, 50]$m and High is > 50m. Similarly for perspective, 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). * indicates lack of adequate information about the dataset (from the corresponding publication or source).

1. Introduction

We introduce SkyScenes, a densely annotated dataset of synthetic scenes captured from aerial (UAV) viewpoints under diverse *layout* (urban and rural), *weather*, *daytime*, *pitch* and *altitude* conditions. The setting of outdoor aerial imagery provides unique challenges for scene understanding: 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, changes in weather or daytime conditions, etc. 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 high-resolution aerial images is expensive – for instance, densely annotating a single 4K image in UAVid [16] can take up to 2 hours! Second, any effort to expand the real set to include widespread variations (weather, time of day, pitch, altitude) would be uncontrolled (*i.e.,* can't guarantee the same viewpoint under different conditions as real world is not static) and additionally would require re-annotating newly captured frames. Synthetic data curated from simulators can help counter both of these issues as – (1) labels are automatic and cheap to obtain and (2) it is possible to recreate the same viewpoint (with the same actor instances – `vehicles`, `humans`, etc. in the scene) under differing conditions.

Unlike synthetic ground plane view datasets (especially for autonomous driving [5, 18, 23, 26, 34]), synthetic datasets for aerial imagery (see Table. 1, rows 4-10) relatively have received less attention [2, 9, 14, 15, 22, 27, 32] Existing synthetic datasets for aerial imagery can be 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). We cover all these aspects by introducing SkyScenes, a synthetic dataset containing 33.6K densely annotated aerial scenes captured from Carla [7]. We curate SkyScenes images by re-purposing the Carla [7] simulator for aerial viewpoints by re-positioning the agent camera to a desired altitude and pitch to obtain an oblique perspective. SkyScenes images are coupled with dense semantic, instance segmentation (28 classes) and depth annotations.

We carefully curate SkyScenes images by procedurally teleoperating the aerially situated camera through 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 is *physically* realistic. This involves introducing vari-

ations in sensor locations, such as adding jitter to specified height and pitch values, to mimic real-world conditions.[1] Finally, since CARLA [7] 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 Sec. 3.1 and Sec. 3.2 for a detailed discussion).

Empirically, we demonstrate the utility of SKYSCENES in several different ways. First, we show that SKYSCENES is a "good" 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) that SKYSCENES pretraining can help reduce data requirements from the real-world (improved 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 show that SKYSCENES alone can 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 development of multi-modal segmentation models.

2. Related Work

Ground-view Synthetic Datasets. Real world ground-view scene-understanding datasets (Cityscapes [5], Mapillary [18], BDD-100K [34], Dark Zurich [25]) 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 [26], Synthia [23] and VisDA-C [20] are some of the widely-used synthetic datasets. These datasets can be curated using underlying simulators, such as the GTAV [26] game engine and CARLA [7] simulator, and offer a cost-effective and scalable way to generate large amounts of labeled data under diverse conditions. Similar to SELMA [30] and SHIFT [29], we use CARLA [7] as the underlying simulator for SKYSCENES.

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 [31], DeepGlobe [6], ISPRS2D [24], and FloodNet [21] 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 that emphasize on object identification. Many existing urban scene datasets, like Aeroscapes [19], UAVid [16], VDD [1], UDD [4], UAVDT [8], VisDrone [35], Semantic Drones [13] 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 [28], used for the development and testing of autonomous systems (in particular, aerial vehicles), is the foundation of several synthetic UAV-based datasets like MidAir [9], Espada [15], Tartan Air [32], UrbanScene3d [14] and VALID [2]. CARLA [7] is another such open-source simulator that is the foundation of datasets like SynDrone[22]. 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 (differences summarized in Table. 1). 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 [7][2] 0.9.14, which is a flexible and realistic open-source autonomous vehicle

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

[2] https://carla.org/

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

1. **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 30m-50m 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 (jitter $\Delta h \sim \mathcal{N}(1, 2.5m)$) to simulate realistic data sampling.

2. **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 (most common one). Similar to height, pitch variations allow models trained on SKYSCENES to generalize to different real-world conditions. We use $\theta = 45°$ and $60°$ for oblique-views and introduce jitter (jitter $\Delta\theta \sim \mathcal{N}(1, 5°)$) to mimic real-world data sampling.

3. **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 [7] 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. Fig. 5 illustrates how images curated from different towns in CARLA [7] differ in class distributions.

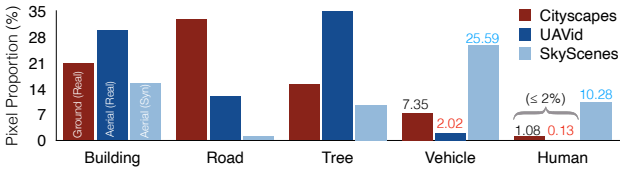

Figure 2. **Ground-View vs Aerial-View Pixel Proportions.** For a subset of commonly annotated classes across Cityscapes [5] (red), UAVid [16] (dark blue) and SKYSCENES (light blue), we show the percentage of pixels occupied by different classes. Aerial scenes (in UAVid) have significant under-representation of tail classes (`vehicle`, `human`). SKYSCENES image curation (for the same viewpoint settings as UAVid) helps counter this discrepancy via synthetic aerial scenes.

4. **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.[3] 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.

5. **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).

6. **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 real-world would be uncontrolled, due to it's 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

[3]Note that CARLA [7] provides 14 such conditions but we use only 5 such conditions in this preliminary version of SKYSCENES.

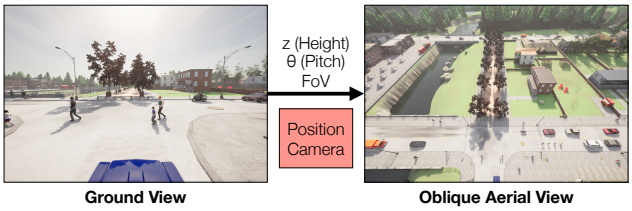

Figure 3. **SKYSCENES Ground View → (Oblique) Aerial View.** Upon initializing a CARLA [7] scene (for a Town and Variation), we re-position the camera associated with the actor to obtain oblique aerial views from ground views.

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.

7 . **Adequate Representation of Tail Classes:** Unlike ground-view datasets, pixel distribution of classes in aerial images is substantially more long-tailed (see Fig. 2; classes with smaller object size, `humans`), making visual recognition tasks harder. This problem is further exacerbated by differing class distributions across synthetic and real data (see Fig. 5) – especially severe for instances of `human` class. To counter this, we consider structured spawning of `humans` to ensure adequate representation.

3.2. SKYSCENES Image Generation

We generate SKYSCENES images from CARLA [7] by taking the previously mentioned considerations into account. Curating images from CARLA [7] 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 [7] 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. 3).[4] Once positioned, the agent is translated by fixed amounts to traverse the scene and capture images from various viewpoints (detailed in Algo 1 in supplementary). Initially, we generate 70 datapoints for each of the 8 town variations under ClearNoon conditions using the baseline $h = 35, \theta = 45°$ setting. Subsequently, following the traversal algorithm (see 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 [7] uses a traffic manager with a PID controller to control the egocentric vehicle based on 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 and further reduces 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 [16] (aerial) and Cityscapes [5] (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 [7] is detrimental for eventual task performance – for the `human` class, a SKYSCENES trained DAFormer [11](with HRDA [12] source training; MiT-

[4]This also requires setting other scene – weather, daytime, etc. – and camera – notably the `FoV` $= 110°$ (field of view) and $\{H, W\} = \{2160, 1440\}$ (image resolution) – parameters.

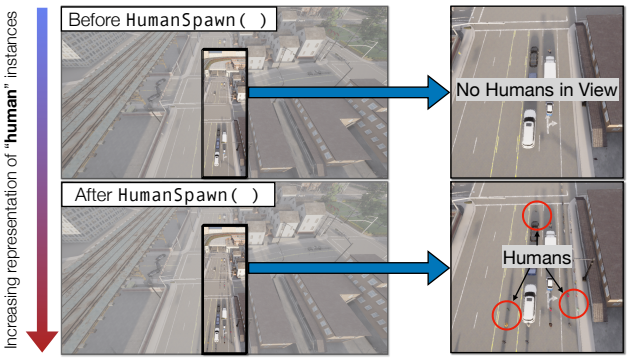

Figure 4. **SKYSCENES HumanSpawn() Effect.** Incorporating HumanSpawn() in the image generation pipeline for SKYSCENES increases the proportion of `humans` in snapshots (**[Top]→[Bottom]**).

Layout	RS (%)	HS (%)
1 Town01	0.08	0.14
2 Town02	0.18	0.17
3 Town03	0.07	0.21
4 Town04	0.14	0.31
5 Town05	0.08	0.43
6 Town06	0.07	0.36
7 Town07	0.12	0.26
8 Town10HD	0.18	0.35
9 All	0.12	0.28
10 UAVid [16]	0.13	0.13

Table 2. **Increase in the `human` representation.** HumanSpawn (HS) improves the representation of `humans` in SKYSCENES. RS = Random Spawn.

		mIoU(↑)	
Eval Data	HS	`human`	All
1 S		43.03	80.87
2 S	✓	**61.79**	**84.07**
3 S→U		4.71	45.11
4 S→U	✓	**10.21**	**47.09**

Table 3. **Improved `human` recognition.** Training on HumanSpawn (HS) SKYSCENES images improves the model's ability to recognize `humans` (improved mIoU). S = SKYSCENES, U = UAVid [16].

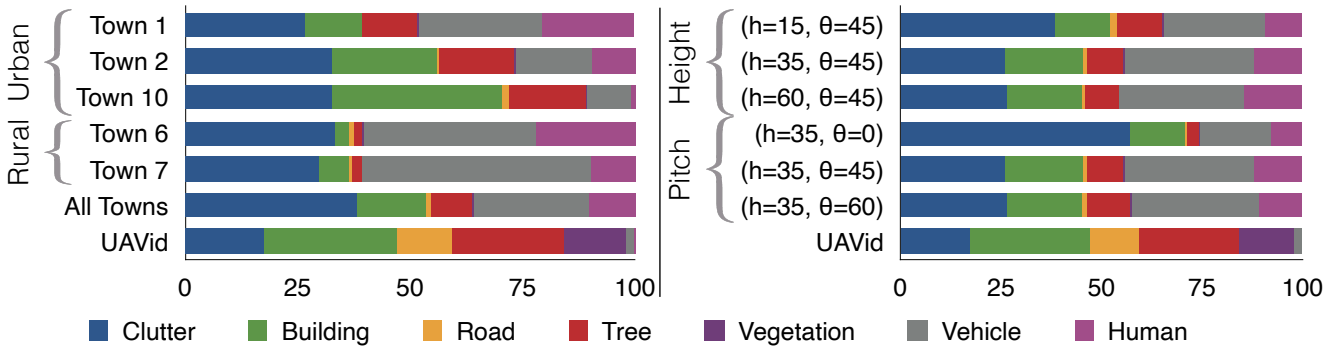

Figure 5. **Class-distribution Diversity in SkyScenes.** We show how the distribution of densely-annotated pixels varies across different SkyScenes conditions. **[Left]** Class-distribution varies substantially within and across urban and rural map layouts. **[Right]** Similarly, for same SkyScenes layouts (and viewpoints) class distribution varies substantially across different height and pitch values.

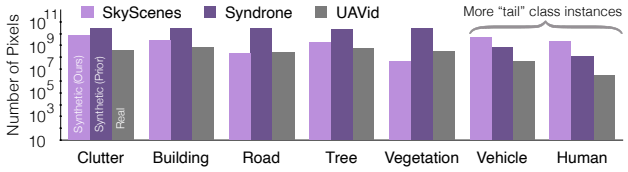

Figure 6. **SkyScenes Per-Class Pixel Counts.** We compare the number of densely annotated pixels per-class for SkyScenes (ours), Syndrone [22] (another synthetic aerial dataset) and UAVid [16] (real aerial dataset). We can see how compared to both synthetic and real counterparts, SkyScenes provides increased representation of tail classes (`vehicles`, `humans`).

B5 [33] backbone) model leads to an in-distribution performance of 43.03 mIoU and out-of-distribution (SkyScenes →UAVid [16]) performance of 4.71 mIoU. To counter this under-representation issue, we design an algorithm, HumanSpawn()[5], 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 Table. 2 & Fig. 4). 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.)

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).[6]

[5]More details in the supplementary.

[6]We provide detailed definitions in the supplementary.

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. While creating train, val and test splits, we collect equal number of samples from each town by dividing each town-specific traversal sequence into 3 segments: the initial 80% for training, the next 10% for testing, and the final 10% of the segment for validation. Moreover, within each split, we ensure that every viewpoint is accompanied by its 60 different variations across weather, daytime, height, and pitch settings. This safeguards against any potential cross-contamination across different splits while ensuring fair representation and equal distributions of all variations.

Class Distribution(s). In Fig. 6, we compare the number of densely annotated pixels in SkyScenes with those in Syndrone [22] (another synthetic dataset) and UAVid [16] (a real aerial dataset). Compared to both the real and prior synthetic counterparts, we show that SkyScenes is specifically curated to ensure adequate tail class representation (more `human` and `vehicle` pixels) to facilitate model learning. Additionally, in Fig. 5, 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 Sec. 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 as-

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**
3 Train-on-Target	68.53	68.59	73.12
DAFormer (MiT-B5) [12]			
4 SynDrone	42.31	30.53	15.92
5 SkyScenes	**47.09**	**40.72**	**25.91**
6 Train-on-Target	72.47	77.80	76.44

Table 4. **Models trained on SkyScenes generalize well to the real-world.** We train semantic segmentation models (DeepLabv2 [3], DAFormer [11]) on SkyScenes, SynDrone and real datasets [22] and show how training models on SkyScenes provides better out-of-the-box generalization to multiple real-world datasets. Rows in gray represent an upper-bound oracle performance.

sess 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 [22], a recently proposed synthetic aerial dataset also curated from Carla [7]. We assess performance on 3 real-world aerial datasets – UAVid [16], AeroScapes [19], ICG Drone [13]. 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.[7] 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 the supplementary.

Models. We use both CNN – DeepLabv2 [3] (ResNet-101 [10]) – and transformer – DAFormer [11] (with HRDA [12] source training; MiT-B5 [33] backbone) – based semantic segmentation architectures for our experiments. We provide implementation details surrounding our experiments in supplementary.

4.1. SkyScenes → Real Transfer

▷ **SkyScenes trained models generalize well to real-settings.** In Table. 4, 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 [22] across both CNN and transformer segmentation backbones. In Table. 5, we show how generalization improvements are more pronounced for under-represented tail classes (`vehicles` and `humans`)[8].

[7] We detail class merging and assignment schemes used for these experiments in supplementary.

Source	(Target) Real-World mIoU (↑) UAVid		AeroScapes		ICG Drone	
	vehicle	human	vehicle	person	vehicle	person
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**
3 Train-on-Target	80.70	41.25	87.58	57.99	94.08	83.64

Table 5. **SkyScenes training exhibits strong real-world generalization for tail classes.** We show how DAFormer [11] models trained on SkyScenes exhibit improved real-world generalization compared to those trained on SynDrone for under-represented tail classes (`vehicles` and `humans`). SkyScenes training facilitates better recognition of tail class instances. Rows in gray are meant to represent oracle numbers, indicating an upper bound on attainable performance.

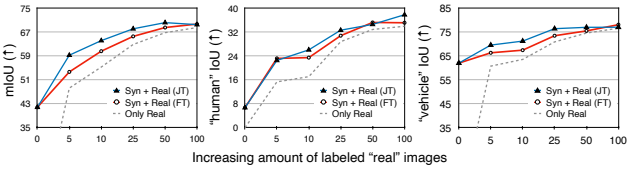

Figure 7. **SkyScenes can augment "real" training data.** We show how SkyScenes can additionally augment real (UAVid [16]) training data. We compare DeepLabv2 [3] models trained using only $5\%, 10\%, 25\%, 50\%, 100\%$ of labeled UAVid [16] images with counterparts that were either (1) pretrained on SkyScenes, and finetuned on UAVid [16] (FT) or (2) trained jointly on SkyScenes and UAVid [16] (JT). We find that **[Left]** additionally augmenting training data with SkyScenes and help improve real-world generalization in low-shot regimes, **[Middle, Right]** especially for under-represented classes.

▷ **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 Fig. 7, for SkyScenes →UAVid [16], we compare models trained only using $5\%, 10\%, 25\%, 50\%, 100\%$ of the 200 UAVid [16] 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. We find this to be especially beneficial for under-represented classes in aerial imagery (such as `humans` and `vehicles`). We discuss more such results across other real-world datasets in 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. 5 for

[8] Model performance comparisons for all the classes are provided in the supplementary.

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 Variations**

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 Variations**

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

(c) **Map Variations**

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

(d) **Height & Pitch Variations**

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 [11] 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 ($h = 35, \theta = 45°$) across different h, θ variations. We observe that, as the altitude increases, oblique viewpoints are best suited for recognition. Best numbers across each row condition is highlighted in blue.

Sensors	SKYSCENES Test mIoU (↑)							
	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 7. **Multi-modal Segmentation in SKYSCENES.** We show how SKYSCENES enables developing multi-modal segmentation models. We evaluate M3L multimodal segmentation architectures (with MiT-B5 backbones) with RGB and RGB+D observations and show additional sensors help substantially improve aerial scene understanding. We consider the broad set of UAVid class palette for this experiment.

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, same number of images across layouts. 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 $(h = 35, \theta = 45°)$ 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 supplementary.

4.3. SKYSCENES Enables Multi-modal Recognition

Sensors on UAVs in deployable settings are not limited to RGB cameras. It is common to have UAVs deployed in the real-world with additional modality sensors, such as depth. Additional sensor modalities can also potentially help improve aerial scene understanding. In Table. 7, we check if assisting RGB with Depth observations for SKYSCENES viewpoints can help improve aerial semantic segmentation using M3L [17], a model capable of multimodal segmentation. Similar to our DAFormer [11] experiments, we consider a SegFormer equivalent version of M3L [17] (with an MiT-B5 [33] 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 annotated 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 scenes curated from unmanned aerial vehicle (UAV) perspectives. We collect SKYSCENES images from CARLA by first aerially situating an agent (to get an aerial perspective) and then procedurally teleoperating the agent through the scene to capture aerial frames with corresponding semantic, instance and depth annotations. Our careful curation process ensures that SKYSCENES images are carefully curated 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 how (1) SKYSCENES trained models can generalize to real-world settings, (2) SKYSCENES images can augment labeled real-world data in low-shot regimes, (3) SKYSCENES can serve as a diagnostic framework to assess model sensitivity to changing conditions and (4) additional sensors, such as Depth, in SKYSCENES can facilitate development of multi-modal aerial scene understanding models.

Lastly, we plan on updating SKYSCENES with evolving considerations for real-world aerial scene-understanding – improved realism, additional anticipated edge cases – as more and more features are supported in the underlying simulator. We intend to release both the dataset and associated generation code for SKYSCENES publicly, and hope that our experimental findings encourage further research using SKYSCENES for aerial scenes.

References

[1] Wenxiao Cai, Ke Jin, Jinyan Hou, Cong Guo, Letian Wu, and Wankou Yang. Vdd: Varied drone dataset for semantic segmentation, 2023. 3

[2] Lyujie Chen, Feng Liu, Yan Zhao, Wufan Wang, Xiaming Yuan, and Jihong Zhu. Valid: A comprehensive virtual aerial image dataset. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2009–2016, 2020. 2, 3

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, 2017. 7

[4] Yu Chen, Yao Wang, Peng Lu, Yisong Chen, and Guoping Wang. Large-scale structure from motion with semantic constraints of aerial images. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 347–359. Springer, 2018. 3

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 3, 4, 5

[6] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Bo Huang, Saikat Basu, Forest Hughes, Devis Tuia, Radim Raska, Abigail Kressner, et al. Deepglobe 2018: A challenge to parse the earth through satellite images. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–172, 2018. 3

[7] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 2, 3, 4, 5, 7

[8] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–386, 2018. 3

[9] Michael Fonder and Marc Van Droogenbroeck. Mid-air: A multi-modal dataset for extremely low altitude drone flights. In *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2019. 2, 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[11] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation, 2022. 5, 7, 8

[12] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation, 2022. 5, 7

[13] Institute of Computer Graphics and Vision, Graz University of Technology. Semantic drone dataset. http://dronedataset.icg.tugraz.at. 2, 3, 7

[14] Liqiang Lin, Yilin Liu, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang. Capturing, reconstructing, and simulating: the urbanscene3d dataset. In *ECCV*, pages 93–109, 2022. 2, 3

[15] Rafael Lopez-Campos and Jose Martinez-Carranza. Espada: Extended synthetic and photogrammetric aerial-image dataset. *IEEE Robotics and Automation Letters*, 6(4):7981–7988, 2021. 2, 3

[16] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS journal of photogrammetry and remote sensing*, 165:108–119, 2020. 2, 3, 4, 5, 6, 7

[17] Harsh Maheshwari, Yen-Cheng Liu, and Zsolt Kira. Missing modality robustness in semi-supervised multi-modal semantic segmentation, 2023. 8

[18] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5000–5009, 2017. 2, 3

[19] Ishan Nigam, Chen Huang, and Deva Ramanan. Ensemble knowledge transfer for semantic segmentation. In *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision*, pages 916–924. IEEE, 2018. 2, 3, 7

[20] Xi Peng, Bingyi Usman, Karthik Kaushik, Judy Hoffman, Dequan Wang, Kate Saenko, and Yan Zhang. Visda: The visual domain adaptation challenge. In *IEEE International Conference on Computer Vision*, pages 1685–1692, 2017. 3

[21] Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding, 2020. 3

[22] Giulia Rizzoli, Francesco Barbato, Matteo Caligiuri, and Pietro Zanuttigh. Syndrone–multi-modal uav dataset for urban scenarios. *arXiv preprint arXiv:2308.10491*, 2023. 2, 3, 6, 7

[23] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. 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*, pages 3234–3243, 2016. 2, 3

[24] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sébastien Bénitez, and U Breitkopf. The isprs benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, I-3, 2012. 3

[25] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[26] Swami Sankaranarayanan, Yogesh Balaji, Carlos D. Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3

[27] Maria Scanlon. *Semantic Annotation of Aerial Images using Deep Learning, Transfer Learning, and Synthetic Training Data*. PhD thesis, 2018. 2

[28] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*, 2017. 3

[29] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. Shift: A synthetic driving dataset for continuous multi-task domain adaptation, 2022. 3

[30] Paolo Testolina, Francesco Barbato, Umberto Michieli, Marco Giordani, Pietro Zanuttigh, and Michele Zorzi. Selma: Semantic large-scale multimodal acquisitions in variable weather, daytime and viewpoints, 2022. 3

[31] Xin-Yi Tong, Gui-Song Xia, Qikai Lu, Huanfeng Shen, Shengyang Li, Shucheng You, and Liangpei Zhang. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237:111322, 2020. 3

[32] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. 2020. 2, 3

[33] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 6, 7, 8

[34] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 2, 3

[35] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 3