UAV (UNMANNED AERIAL VEHICLES): DIVERSE APPLICATIONS OF UAV DATASETS IN SEGMENTATION, CLASSIFICATION, DETECTION, AND TRACKING

Md. Mahfuzur Rahman
Software Engineer
Silicon Orchard Research and Analytics Lab
Dhaka, Bangladesh
mahim1066@gmail.com

Sunzida Siddique
Dept. of Computer Science
Daffodil International University
Dhaka, Bangladesh
sunzida15-9667@diu.edu.bd

Marufa Kamal
Dept. of CSE
BRAC University
Dhaka, Bangladesh
marufa.kamal1@g.bracu.ac.bd

Rakib Hossain Rifat
Dept. of Computer Science
Texas Tech University
Lubbock, TX 79409
rrifat@ttu.edu

Kishor Datta Gupta
Cyber Physical Systems
Clark Atlanta University
Atlanta, GA
kgupta@cau.edu

ABSTRACT

Unmanned Aerial Vehicles (UAVs), have greatly revolutionized the process of gathering and analyzing data in diverse research domains, providing unmatched adaptability and effectiveness. This paper presents a thorough examination of Unmanned Aerial Vehicle (UAV) datasets, emphasizing their wide range of applications and progress. UAV datasets consist of various types of data, such as satellite imagery, images captured by drones, and videos. These datasets can be categorized as either unimodal or multimodal, offering a wide range of detailed and comprehensive information. These datasets play a crucial role in disaster damage assessment, aerial surveillance, object recognition, and tracking. They facilitate the development of sophisticated models for tasks like semantic segmentation, pose estimation, vehicle re-identification, and gesture recognition. By leveraging UAV datasets, researchers can significantly enhance the capabilities of computer vision models, thereby advancing technology and improving our understanding of complex, dynamic environments from an aerial perspective. This review aims to encapsulate the multifaceted utility of UAV datasets, emphasizing their pivotal role in driving innovation and practical applications in multiple domains.

Keywords UAV (Unmanned Aerial Vehicle) · UAV datasets · object detection · semantic segmentation · action recognition · event recognition · aerial · surveillance

1 Introduction

Unmanned Aerial Vehicles (UAVs)[1], commonly referred to as drones, have revolutionized the way we collect and analyze data from above, offering unparalleled versatility and efficiency across various research fields. This review paper aims to explore the "Multiple Uses of UAV Datasets" by examining the diverse applications and advancements facilitated by these datasets. UAV datasets encompass a wide array of data types, including satellite imagery, drone-captured images, and videos, as well as images from other aerial vehicles like helicopters. These datasets can be unimodal, focusing on a single type of data, or multimodal, integrating multiple data types to provide deeper, more comprehensive insights.

UAV datasets have proven to help assess disaster damage because they allow for the classification of damage from natural disasters using sophisticated semantic segmentation and annotation techniques. By training computer vision models with these datasets, researchers can automate the aerial scene classification of disaster events, significantly

enhancing response and recovery efforts. The ability to extract information and detect objects from UAV-captured data is pivotal for tasks such as action recognition, where human behavior is analyzed from aerial imagery, including recognizing aerial gestures and classifying disaster events.

A critical application of UAV datasets lies in 'Aerial Surveillance'[2], which supports advanced research at the intersection of computer vision, robotics, and surveillance. These datasets are used for event recognition in aerial videos, aiding in the monitoring of urban environments and traffic systems. The use of pre-trained models and transfer learning techniques further amplifies the utility of UAV datasets, allowing for the rapid deployment of sophisticated models for event recognition and tracking.

In the context of urban surveillance, UAV datasets enhance object recognition capabilities by providing comprehensive views from both top-down and side perspectives. This facilitates tasks such as categorization, verification, object detection, and tracking of individuals and vehicles. Moreover, UAV datasets contribute significantly to understanding and managing forest ecosystems by addressing the challenge of segmenting individual trees, which is crucial for sustainable forest management.

The versatility of UAV datasets extends to various domains, including developing speech recognition systems for UAV control using video capture and object tracking in low-light conditions, which is essential for night-time surveillance operations. Innovative UAV designs, such as bionic drones with flapping wings, have also led to specialized video datasets used for single object tracking (SOT)[3][4], demonstrating the broad scope and potential of UAV datasets in enhancing real-time object tracking under varying lighting conditions.

Overall, UAV datasets represent a cornerstone for cutting-edge research and practical applications across multiple disciplines. This review will delve into the specific uses and benefits of these datasets, highlighting their role in advancing technology and improving our understanding of complex, dynamic environments from an aerial perspective.

The subsequent sections provide a comprehensive exposition of the contributions made by our study, which can be stated as follows:

- Our study is driven by the increasing importance of UAV datasets in several research domains such as object detection, traffic monitoring, action identification, surveillance in low-light conditions, single object tracking, and forest segmentation utilizing point cloud or LiDAR point process modeling. Through an in-depth analysis of current datasets, their uses, and prospects, this paper intends to provide valuable insights that will assist researchers in harnessing these resources for creative solutions. Furthermore, they will acquire knowledge of existing constraints and prospective opportunities, enhancing their research endeavors.
- We conducted an extensive analysis of a dataset consisting of 15 Unmanned Aerial Vehicles (UAVs), showcasing its diverse applications in research.
- We emphasized the applications and advancements of several novel methods utilizing these datasets based on unmanned aerial vehicles (UAVs).
- Our study also delved into the potential for future research and the feasibility of utilizing these UAV datasets, engaging in in-depth discussions on these topics.

2 Literature Review

An unmanned aircraft or UAV, functions without a human pilot on board and can be operated remotely by a human controller or independently by onboard computers. Drones are a common term used to describe UAVs. Drones are employed for various purposes, including surveillance, aerial photography, agriculture, environmental monitoring, and military operations. However, within the UAV dataset context, the term encompasses more than just drones. UAV datasets encompass not only drone image and video datasets, but also include satellite imagery. Table1 and 2 shows the summary of the literature review performed.

These papers were reviewed to determine the definition and range of applications of UAVs in computer vision.

2.1 RescueNet

Maryam Rahnemoonfar, Tashnim Chowdhury, and Robin Murphy presented the RescueNet[5] dataset in their paper, which focuses on post-disaster scene understanding using UAV imagery. The dataset contains high-resolution images with detailed pixel-level annotations for ten classes of objects, including buildings, roads, pools, and trees, which were collected by sUAVs following Hurricane Michael. The authors employed state-of-the-art segmentation models like Attention UNet[6], PSPNet[7], and DeepLabv3[8], achieving superior performance with attention-based and transformer-based methods. The findings demonstrated RescueNet's effectiveness in improving damage assessment and

response strategies, with transfer learning outperforming other datasets like FloodNet[9]. The dataset was observed to have limited generalization to other domains and to require a time-consuming annotation process, despite its detailed annotations.

2.2 UAV-Human

Tianjiao Li et al. developed the UAV-Human[10] dataset, a comprehensive benchmark for improving human behavior understanding with UAVs. The dataset contains 67,428 multi-modal video sequences with 119 subjects for action recognition, 22,476 frames for pose estimation, 41,290 frames for person re-identification with 1,144 identities, and 22,263 frames for attribute recognition, all captured over three months in various urban and rural locations under varying conditions. The data encompasses RGB videos, depth maps, infrared sequences, and skeleton data. The authors used methods such as HigherHRNet[11], AlphaPose[12], and the Guided Transformer I3D framework to recognize actions while addressing fisheye video distortions[13][14] and leveraging multiple data modalities. The results demonstrated the dataset's effectiveness in improving action recognition, pose estimation, and re-identification tasks, with models showing significant performance improvements. The UAV-Human dataset stands out as a reliable benchmark, encouraging the creation of more effective UAV-based human behavior analysis algorithms.

2.3 AIDER

Christos Kyrkou and Theocharis Theocharides introduced the AIDER[15] dataset, which is intended for disaster event classification using UAV aerial images. The dataset contains 2,565 images of Fire/Smoke, Flood, Collapsed Building/Rubble, Traffic Accidents, and Normal cases, which were manually collected from various sources, mainly from UAVs. To increase variability and combat overfitting, images were randomly augmented with rotations, translations, and color shifting. The paper presents ERNet, a lightweight CNN designed for efficient classification on embedded UAV platforms. ERNet, which uses components from architectures such as VGG16[16], ResNet[17], and MobileNet[18], incorporates early downsampling to reduce computational costs. When tested on both embedded platforms attached to UAVs and desktop CPUs, ERNet achieved almost perfect accuracy (90%) while running three times faster on embedded platforms. This showed that it is a good choice for real-time applications that do not need a lot of memory. The study emphasizes the benefits of combining ERNet with other detection algorithms to improve situational awareness in emergency response.

2.4 AU-AIR

In their paper Ilker Bozcan and Erdal Kayacan present the AU-AIR[19] dataset, a comprehensive UAV dataset designed for traffic surveillance. The dataset comprises 32,823 labeled video frames with annotations for eight traffic-related object categories, along with multi-modal data including GPS coordinates, altitude, IMU data[20], and velocity. To establish a baseline for real-time performance in UAV applications, the authors train and evaluate two mobile object detectors on this dataset: YOLOv3-Tiny[21] and MobileNetv2-SSDLite[22]. The findings highlight the difficulties of object detection in aerial images, emphasizing the importance of datasets tailored to mobile detectors. The study highlights the dataset's potential for furthering research in computer vision, robotics, and aerial surveillance, while also acknowledging limitations and suggesting future improvements for broader applicability.

2.5 ERA

Lichao Mou et al. introduced the ERA[23] dataset, a comprehensive collection of 2,864 labeled video snippets for 24 event classes and 1 normal class, designed for event recognition in UAV videos. The videos, sourced from YouTube, are 5 seconds long, 640×640 pixels, and run at 24 fps, ensuring a diverse dataset that includes both high-quality and extreme condition footage. The paper employs various deep learning models, including VGG-16, ResNet-50, DenseNet-201[24], and video classification models like I3D-Inception-v1, to benchmark event recognition. DenseNet-201 achieved the highest performance with an overall accuracy of 62.3% in single-frame classification. The findings highlight the difficulties of recognizing events in a variety of environments and scales, noting that while models can identify specific events such as traffic congestion and smoke, they struggle with conditions such as night and snow scenes, indicating the need for improved attribute recognition and temporal cue exploitation in future research.

2.6 UAVid

Ye Lyu et al. introduced the UAVid[25] dataset in their paper which addresses the need for semantic segmentation in urban scenes from the perspective of UAVs. The UAVid dataset consists of 30 video sequences with 4K high-resolution images, which capture top and side views for improved object recognition and include 8 labeled classes. The

Table 1: Summary of Research and Findings of UAV Datasets discussed

Dataset	Dataset Details	Paper Findings	Limitations	Future Work
RescueNet[5]	High-resolution images with pixel-level annotations for 10 classes, collected via UAVs after Hurricane Michael.	Attention-based and transformer-based methods performed best. Transfer learning from RescueNet to FloodNet improved segmentation.	Time-consuming annotation process and potential lack of comprehensive post-disaster elements.	Further evaluation across different disaster scenarios to enhance robustness.
UAV-Human[10]	67,428 multi-modal video sequences for action recognition, pose estimation, person re-identification, and attribute recognition.	Highest action recognition accuracies with night-vision and IR videos. Pose estimation methods achieved mAP scores of 56.5 and 56.9.	Potential overfitting, lack of subject diversity, and constrained capturing conditions.	Increase sample size and diversity, and capture conditions to enhance model robustness and generalization.
AIDER[15]	2565 manually gathered images of disaster events with augmentations.	Development of ERNet, achieving near state-of-the-art accuracy (90%) and over 50 fps on a CPU platform.	Does not extensively discuss real-time implementation challenges, robustness in diverse conditions, or hyperparameter tuning.	Integrate ERNet with algorithms for detecting people and vehicles, use additional modalities like infrared cameras, and optimize the model for improved generalization and accuracy.
AU-AIR[19]	32,823 labeled video frames with object annotations and flight data.	YOLOv3-tiny and MobileNetv2-SSD Lite for real-time object detection on UAVs showed potential for onboard computer applicability.	Focus on traffic surveillance may limit applicability to other scenarios, lacks advanced baselines for tasks like UAV navigation.	Enhance dataset diversity, incorporate more environmental contexts, and develop additional baselines leveraging sensor data for broader applications.
ERA[23]	2,864 videos capturing events in a wide range of settings and sizes.	DenseNet-201 achieved the highest accuracy of 62.3% in single-frame classification.	Dataset size, class imbalance, and challenge of distinguishing events from normal videos.	Focus on attribute recognition, temporal cue exploitation, and addressing challenging cases like human action recognition.
UAVid[25]	30 video sequences featuring high-resolution 4K images with 8 labeled classes for semantic segmentation.	Multi-Scale-Dilation Net achieved an average IoU score of around 50%.	Class imbalance, particularly in urban street scenes, potentially affecting model performance and generalization.	Balance method complexity with practical implementation, expand dataset size and object categories, address class imbalance, and explore other applications like object detection and tracking.
VRAI[26]	137,613 images of 13,022 vehicles with detailed annotations captured by two UAVs.	Outperforms existing methods in vehicle ReID techniques using GANs and attention models.	Comparison scope, domain specificity, annotation complexity, scalability, and real-world deployment insights.	Explore transfer learning, enhance scalability, integrate advanced techniques, focus on real-world applications, and improve annotation strategies.
VERI-Wild[27]	Over 400,000 images of 40,671 vehicle IDs captured from a real CCTV camera system over one month.	FDA-Net outperforms existing methods, achieving highest Rank-1 and Rank-5 accuracies.	Potential biases due to urban district focus and dataset-specific adversarial scheme.	Explore more challenging real-world factors, generate comprehensive datasets, and leverage GANs to improve cross-view ReID performance.

Table 2: Summary of Research and Findings of UAV Datasets discussed

Dataset	Dataset Details	Paper Findings	Limitations	Future Work
UAV-Assistant[28]	Data synthesis pipeline combining egocentric UAV views and exocentric user views with smooth silhouette loss.	Smooth silhouette loss enhances 3D pose estimation accuracy.	Lack of real-world data poses a challenge to generalizability, and determining optimal kernel size for smoothing filter.	Optimize parameters, explore additional loss functions, and validate approach in real-world scenarios.
KITE[29]	Focus on UAV control speech recognition with multimodal systems.	Recurrent neural networks (RNNs) for language modeling and visual cues integration.	Imperfect command-image associations, biases from semi-automatic methods for training data generation.	Address biases, enhance dataset generalizability, and explore other architectural decisions.
UAV-Gesture[30]	119 high-definition video clips of 13 gestures for UAV navigation and command.	Annotates body joints and gesture classes in 37,151 frames using an extended version of VATIC.	Limited gesture set and non-expert actors may affect dataset quality.	Leverage dataset for gesture and action recognition in UAV control, expand and refine dataset for broader research applications.
DarkTrack 2021[31]	110 annotated sequences totaling over 100,000 frames for low-light UAV tracking.	SCT demonstrated significant performance gains for nighttime UAV tracking.	Comparisons with daytime tracking scenarios needed to be improved.	Explore advanced transformer architectures, attention mechanisms, noise reduction strategies, and real-world validation.
UAVDark 135[32]	Over 125k manually annotated frames for dark tracking methods.	ADTrack demonstrates superiority in bright and dark conditions.	Lacks broader comparison with other state-of-the-art trackers.	Further research on real-time tracking algorithms, new image enhancement methods, multi-sensor fusion techniques, and hardware optimization strategies.
BioDrone[33]	600 videos annotated and labeled at the frame level for single object tracking using bionic drone-based systems.	Comprehensive evaluation platform for robust vision research.	Focus on bionic UAVs may limit generalization, potential biases in annotations.	Improve tracking algorithms, address computational complexity and real-time performance.
FOR-Instance[34]	Five collections from around the world for individual tree segmentation from UAV-based laser scanning data.	Supports both instance and semantic segmentation, adaptable to deep learning frameworks.	Potential overfitting, lack of generalizability to other forest types, challenges with unclassified points.	Incorporate more data types, develop advanced deep learning architectures, study tree species classification, and conduct longitudinal studies on forest changes.

paper highlights the challenges of large-scale variation, moving object recognition, and temporal consistency. The effectiveness of deep learning techniques, such as the Multi-Scale-Dilation net which is a novel technique proposed by the author, was evaluated and resulted in an average Intersection over Union[35] (IoU) score of approximately 50%. Further enhancements were observed by employing spatial-temporal regularization methods like FSO[36] and 3D CRF[37]. The dataset's applicability extends to traffic monitoring, population density analysis, and urban greenery monitoring, showcasing its potential for diverse urban surveillance applications. The paper also discusses the dataset's class imbalance and suggests future expansions and optimizations to enhance its utility for semantic segmentation and other UAV-based tasks.

2.7 VRAI

Peng Wang et al. introduced the VRAI[26] dataset, the largest vehicle re-identification (ReID) dataset with over 137,613 images of 13,022 vehicles. This UAV-based dataset includes annotations for unique IDs, color, vehicle type, attributes, and distinguishing features, capturing a wide range of view angles and poses from UAVs flying between 15m and 80m. The study devised an innovative vehicle ReID algorithm that utilizes weight matrices, weighted pooling, and comprehensive annotations to identify distinctive components. This algorithm surpasses both the baseline and the most advanced techniques currently available. The paper utilizes a comprehensive strategy to perform vehicle ReID using aerial images, showcasing its effectiveness through a range of experiments. Ablation study results demonstrate that the novel Multi-task + DP model, which integrates attribute classification and additional triplet loss on weighted features, exhibits superior performance compared to less complex models. The proposed method outperforms ground-based methods such as MGN[38], RNN-HA[39], and RAM[40], because it can easily handle different view angles in UAV images. Weighted feature aggregation improves performance, as evidenced by the enhanced mean average precision (mAP) and cumulative match characteristic (CMC) metrics. Human performance evaluation highlights the algorithm's strength in fine-grained recognition, though humans still excel in detailed tasks. The study suggests further research to improve flexibility, scalability, and real-world application of the algorithm.

2.8 FOR-Instance

For semantic and instance segmentation of individual trees, Stefano Puliti et al. presented the FOR-Instance[34] dataset in their paper "FOR-Instance: a UAV laser scanning benchmark dataset for semantic and instance segmentation of individual trees." This dataset fills a gap in the market for ML-ready datasets and standardized benchmarking infrastructure by offering publicly accessible annotated forest data for point cloud segmentation[41] tasks. The primary goal is to use data from unmanned aerial vehicle (UAV) laser scanning to precisely identify and separate individual trees. The dataset includes extensive annotations that are used for training and evaluation, and it is composed of five carefully chosen collections from different types of forests worldwide. In the context of deep learning, the dataset is divided into separate sets for the purpose of training and validation. In image segmentation research, rasterized canopy height models are utilized, along with either unprocessed point clouds or two-dimensional projections. The FOR-Instance dataset was found to be useful for studying and testing advanced segmentation methods. This highlights the significance of comprehending forest ecosystems and formulating sustainable management techniques. The standardization of the dataset in 3D forest scene segmentation research helps to address current methodological limitations, such as overfitting and lack of comparability.

2.9 VERI-Wild

Yihang Lou et al. presented the VERI-Wild[27] dataset, the largest vehicle ReID dataset to date, in their paper. Over 400,000 photos of 40,000 vehicle IDs are included in the dataset, which was collected over the course of a month in an urban district using 174 CCTV cameras. The dataset poses a formidable challenge for ReID algorithms due to its inclusion of diverse conditions such as varying backgrounds, lighting, obstructions, perspectives, weather, and vehicle types. The authors introduced FDA-Net, a novel technique for vehicle ReID, to enhance the model's ability to distinguish between different vehicles. FDA-Net combines a feature distance adversary network with a hard negative generator and embedding discriminator. After being tested on the VERI-Wild dataset and other established datasets, FDA-Net surpassed various standard methods, achieving higher accuracies in Rank-1 and Rank-5. This demonstrates the effectiveness of FDA-Net in vehicle ReID tasks. The method's ability to generate hard negatives significantly improved model performance, highlighting its potential for advancing vehicle ReID research in real-world scenarios.

2.10 UAV-Assistant

G. Albanis and N. Zioulis et al. introduced the UAV-Assistant[28] (UAVA) dataset in their paper. The dataset was created using a data synthesis pipeline to generate realistic multimodal data, including exocentric and egocentric views

from UAVs. The dataset can be utilized to train a model that can estimate the pose of an individual by incorporating a novel smooth silhouette loss in addition to a direct regression objective. The dataset can be used to train a model that can accurately determine the position of a person by incorporating a unique smooth silhouette loss along with a direct regression objective. It also uses differentiable rendering techniques to help the model learn from both real and fake data. The study highlights the critical role of tuning the kernel size for the smoothing filter to optimize model performance. The suggested smooth silhouette loss surpasses conventional silhouette loss functions by reducing discrepancies and enhancing the accuracy of 3D pose estimation. This approach specifically tackles the lack of available data for estimating the three-dimensional position and orientation of unmanned aerial vehicles (UAVs) in non-hostile environments. It is different from existing datasets that primarily focus on remote sensing or drones with malicious intent. The paper underscores the need for further research on rendering techniques, parameter optimization, and real-world validations to enhance the model's generalizability and robustness.

2.11 KITE

The KITE[29] dataset, created to improve speech recognition systems for UAV control, was presented by Dan Oneata and Horia Cucu in their paper. The KITE eval dataset is a comprehensive collection that includes 2,880 spoken commands, along with corresponding audio and images. It is specifically designed for UAV operations and covers a range of commands related to movement, camera usage, and specific scenarios. The authors employed time delay neural networks[42] (which is implemented in Kaldi[43]) and recurrent neural networks to perform language modeling. They initialized the models with out-of-domain datasets and subsequently fine-tuned them for UAV tasks. The study emphasizes the efficacy of customizing language models for UAV-specific instructions, showcasing substantial enhancements in speech recognition precision through domain adaptation. Future directions include grounding uttered commands in images for enhanced context understanding and improving the acoustic model's robustness to outdoor noises.

2.12 UAV-Gesture

A. Perera et al. introduced the UAV-Gesture[30] dataset, which addresses the lack of research on gesture-based UAV control in outdoor settings. This dataset aims to fill the existing research gap, as most studies in this field are focused on indoor environments. The dataset consists of 119 high-definition video clips, totaling 37,151 frames, captured in an outdoor setting using a 3DR Solo UAV and a GoPro Hero 4 Black camera. The dataset comprises annotations of 13 body joints and gesture classes for all frames, encompassing gestures appropriate for UAV navigation and command. The dataset was captured with variations in phase, orientation, and camera movement to augment realism. The authors employed an extended version of the VATIC[44] tool for annotation and utilized a Pose-based Convolutional Neural Network[45] (P-CNN) for gesture recognition. This approach resulted in a baseline accuracy of 91.9%. This dataset facilitates extensive research in gesture recognition, action recognition, human pose recognition, and UAV control, showcasing its efficacy and potential for real-world applications.

2.13 UAVDark135

In their research Bowen Li et al. presented the UAVDark135[32] dataset and the ADTrack algorithm. Their work aimed to tackle the challenge of achieving reliable tracking of unmanned aerial vehicles (UAVs) under different lighting conditions. UAVDark135 is the inaugural benchmark specifically developed for tracking objects during nighttime. It consists of more than 125,000 frames that have been manually annotated, addressing a deficiency in current benchmarks. The paper details the ADTrack algorithm, a discriminative correlation filter-based tracker with illumination adaptive and anti-dark capabilities, utilizing image illuminance information and an image enhancer for real-time, all-day tracking. ADTrack performs better in both bright and dark environments, as evidenced by extensive testing on benchmarks such as UAV123@10fps[46], DTB70[47], and UAVDark135—achieving over 30 FPS on a single CPU. While effective, the paper recommends broader comparisons with other state-of-the-art trackers and future research on image enhancement, multi-sensor fusion, and UAV hardware optimization.

2.14 DarkTrack2021

Junjie Ye et al. presented the DarkTrack2021[31] dataset to tackle the difficulty of tracking unmanned aerial vehicles (UAVs) in low-light situations. The dataset consists of 110 annotated sequences containing more than 100,000 frames, providing a varied evaluation platform for tracking UAVs during nighttime. The researchers created an effective low-light enhancer called the Spatial-Channel Transformer (SCT), which combines a spatial-channel Transformer with a robust non-linear curve projection model to effectively enhance low-light images. The Spatial-Channel Attention Module (SCT) employs a technique that effectively combines global and local information, resulting in enhanced

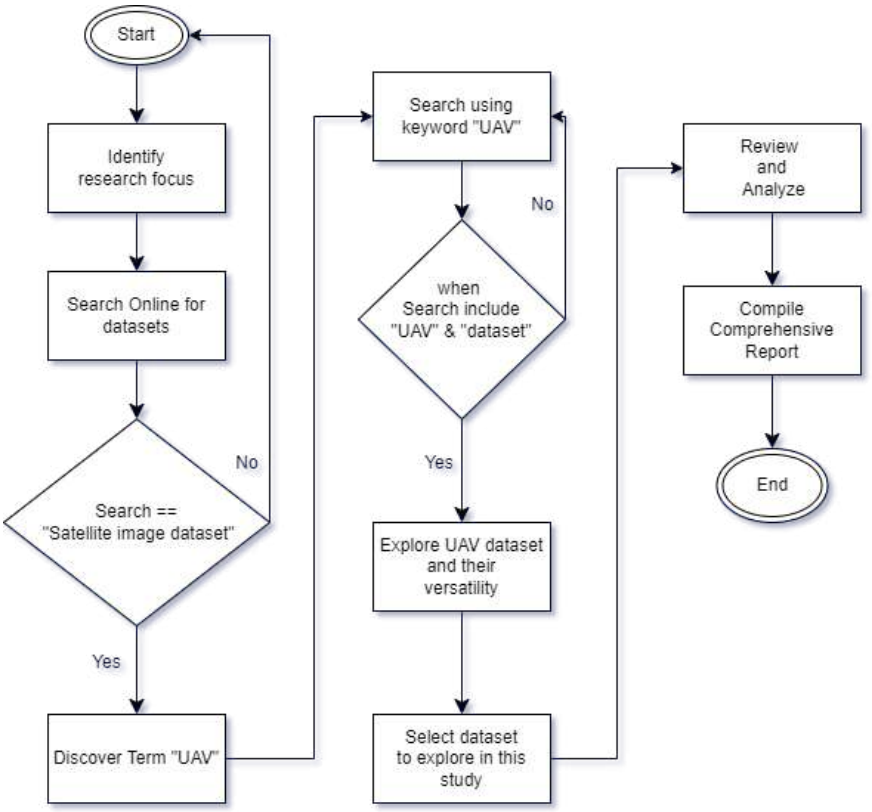


Figure 1: Workflow of this Study

image quality by reducing noise and improving illumination in nighttime scenes. This study utilizes the proposed ADTrack algorithm together with 16 state-of-the-art handmade correlation filter (CF)-based trackers to evaluate their performance on tracking benchmarks UAV123@10fps, DTB70, and UAVDark135. The aim is to demonstrate the comprehensive robustness of the proposed ADTrack algorithm in all-day UAV tracking. Evaluations conducted on the public UAVDark135 and the new DarkTrack2021 benchmarks demonstrated that SCT exhibited superior performance compared to existing methods in tracking UAVs during nighttime. The practicality of the approach has been confirmed through real-world tests. The DarkTrack2021 dataset and SCT code are openly accessible on GitHub for additional research and experimentation.

2.15 BioDrone

Xin Zhao et al. presented the BioDrone[33] dataset. BioDrone is a pioneering visual benchmark for Single Object Tracking[48] (SOT) that utilizes bionic drones. It specifically tackles the difficulties associated with tracking small targets that undergo significant changes in appearance, which are common in flapping-wing UAVs. The dataset consists of 600 videos containing 304,209 frames that have been manually labeled. Additionally, there are automatically generated labels for ten challenge attributes at the frame level. The study presents a new baseline method, UAV-KT, optimized from KeepTrack[49], and evaluates 20 SOT models, ranging from traditional approaches like KCF[50] to sophisticated models combining CNNs and SNNs. The results of comprehensive experiments demonstrate that UAV-KT outperforms other methods in handling challenging vision tasks with resilience. The paper emphasizes BioDrone's potential for advancing SOT algorithms and encourages future research to address remaining challenges, such as camera shake and dynamic visual environments.

3 Methodology

The term UAV (Unmanned Aerial Vehicle) encompasses a diverse range of applications, requiring a thorough investigation to examine and define the extensive utilization of UAV datasets. We aimed to comprehend how these datasets can be employed in different research and project scenarios. To accomplish this, we implemented an exhaustive search for

UAV datasets, initially narrowing our focus to the keyword "satellite or drone image datasets". The initial search led to the identification of "UAV datasets". After acknowledging the potential of UAV datasets, we conducted further research in this field, identifying their diverse applications in object detection, tracking, and event detection, as well as semantic segmentation and single object tracking.

To gather relevant UAV datasets, we conducted systematic searches on the Internet, employing a range of keywords and search terms related to UAVs and their applications. We specifically looked for datasets that showed off the adaptability of UAVs, choosing those that researchers had proposed and used in other research contexts. This approach ensured that the datasets we included were novel and provided diverse examples of UAV applications.

We identified and collected 15 UAV image datasets for inclusion in our study. Our selection criteria focused on datasets that showcased a variety of use cases, including traffic systems (car identification, person identification, and surveillance systems), damage classification from disasters, and other object detection and segmentation tasks. Each dataset was thoroughly reviewed and analyzed to understand its characteristics, intended use, and underlying methodologies.

Our analysis involved a detailed examination of the datasets, resulting in the comprehensive report included in this paper. This report outlines the behavior, agenda, and applications of each dataset, providing insights into their respective fields of use. By presenting these findings, we aim to highlight the versatility and potential of UAV datasets in advancing various research domains. Figure 1 depicts the sequential process of our work.

3.1 Search Terms

We got the datasets we surveyed in this paper mostly from the website, `https://paperswithcode.com/`. Before we found this website we used various search terms to search for the UAV dataset and came across the website through the search process. Example search strings:

- ("unmanned aerial vehicle" OR UAV OR drone OR Satellite) AND ("dataset" OR "image dataset" OR "dataset papers")
- (UAV OR "unmanned aerial vehicle") AND ("disaster dataset" OR "traffic surveillance")

These search strings and keywords facilitated a broad yet focused search, enabling us to gather a diverse set of UAV datasets that demonstrate their wide-ranging applications and research potential.

4 Data Diversity of UAV

The advent of Unmanned Aerial Vehicles (UAVs) has opened new frontiers in data collection and analysis, transforming numerous fields with their versatile applications. The datasets generated by UAVs are diverse, encompassing various data types and serving multiple purposes. This section provides an overview of the various uses of UAV datasets, examines their diversity, and explores the methods applied to utilize these datasets in different studies.

4.1 Overview of UAV Dataset Uses

UAV datasets are pivotal in numerous domains, including disaster management, surveillance, agriculture, environmental monitoring, and human behavior analysis. The unique aerial perspectives provided by UAVs enable the collection of high-resolution imagery and videos, which can be used for mapping, monitoring, and analyzing different environments and activities.

4.1.1 Disaster Management

UAV datasets are often used to figure out how much damage hurricanes, earthquakes, and floods have done. High-resolution images and videos captured by UAVs allow for precise mapping of affected areas and the identification of damaged infrastructure.

4.1.2 Surveillance

In urban and rural settings, UAV datasets support advanced surveillance activities. They facilitate the monitoring of traffic, detection of illegal activities, and overall urban planning by providing real-time, high-resolution aerial views.

Table 3: Summary of Experimented Methods and Results on Different Datasets

Dataset Name	**Experimental Methods in Base Dataset publication**	**Analysis on Results**
RescueNet[5]	PSPNet, DeepLabv3+, Attention UNet, Segmenter[51]	Attention UNet achieved the best performance among all evaluated models. PSPNet showed better performance compared to DeepLabv3+ by using pyramid pooling. DeepLabv3+ provided moderate results, improving on the loss of boundary information. Segmenter showed varying results depending on the backbone (ViT-Tiny vs. ViT-Small), with heavier backbones achieving better results.
UAV-Human[10]	Guided Transformer I3D Network, Video Transformers, Full Model (Author's novel method)	Night-vision and IR videos outperformed previous findings in low-light conditions, achieving 28.72% and 26.56% accuracy, respectively. However, depth sequences face noise issues, and fisheye distortion impacts performance. In ablation studies, using KL Divergence Constraint resulted in 21.68% accuracy, while employing guidance loss and Video Transformers yielded 21.49% accuracy without RGB stream guidance. Overall, the full model had the highest accuracy among fisheye-based methods.
AIDER[15]	Novel networks (ERNet, SCFCNet, SCNet, baseNet), VGG16, ResNet50, MobileNet	The VGG16 model had the highest accuracy at 91.9% but a low frame rate of 2, while consuming 59.3MB of memory. MobileNet had a high frame rate of 20 but lower accuracy at 88.5%. Custom networks like ERNet and SCFCNet had good accuracy at 90.1% and 87.7% with high frame rates of 53 and 76, making them suitable for real-time UAV applications.
AU-AIR[19]	YOLOv3-tiny, MobileNetv2-SSD Lite	YOLOv3-tiny achieved higher mAP (38.2%) and better FPS (22) compared to MobileNetv2-SSDLite (32.8% mAP and 19 FPS), highlighting its better performance for real-time object detection tasks using UAVs.
ERA[23]	VGG-16, DenseNet-121, NASNet-L, C3D (C3D†, C3D‡)[52]	DenseNet-121 achieved the highest overall accuracy (62.3%) among the models, followed by NASNet-L (60.2%) and VGG-16 (51.9%). The C3D models had the lowest accuracy (around 30%).
UAVid[25]	FCN-8s[53], Dilation Net, U-Net[54], MS-Dilation Net	MS-Dilation Net achieved the highest mean IoU score of 57.3% with pre-training and feature space optimization, demonstrating the best performance among the models evaluated.
VRAI[25]	MGN, RAM, RNN-HA, Ensemble methods (e.g. ID Classification Loss, Triplet + ID Loss), Novel methods (Multi-task, Multi-task + Discriminative Parts)	The multi-task model with discriminative parts achieved the highest mAP (78.63%) and CMC-1 (80.30%). The models using Triplet + ID Loss also showed high performance, particularly with Resnet-101 and Resnet-152 backbones.
FOR-instance[27]	None, as the paper is solely focused on constructing the dataset and explaining how to utilize it for model.	N/A

Table 4: Summary of Experimented Methods and Results on Different Datasets

Dataset Name	**Experimented Methods on Dataset**	**Analysis on results**
VERI-Wild[27]	GoogLeNet[55], Triplet[56], Softmax[57], CCL[58], HDC[59], Unlabeled GAN[60][61], EN (Embedding Network with Triplet and Softmax Loss), FDA-Net $\ominus$ Att, FDA-Net	FDA-Net consistently outperforms the other models across different settings, achieving the highest mAP (35.11%) and match rate (R=1 of 64.03% for small dataset). The proposed FDA-Net model demonstrates its effectiveness in vehicle re-identification tasks.
UAV-Assistant (UAVA)[28]	Singleshotpose[62], Direct, IoU based experimental methods (e.g. I0.1, I0.2, I0.1-0.4, G0.1, S0.1, S0.2), Generalized IoU based method (Gauss0.1)[63]	Gauss0.1 showed the best overall performance, particularly in the 6D Pose-5 and 6D Pose-10 metrics. Metrics such as NPE, OE, and CPE were used, with lower values indicating better performance and higher values for Acc5 and Acc10 indicating better performance.
KITE[29]	Baseline Systems (Unadapted System, Domain-Specific System), Domain Adaptation (Text-Only Adaptation, Rescoring), Multi-Modal Experiments (Text and Visual Information)	Domain adaptation and multi-modal approaches significantly improved the performance of speech recognition systems for UAV control. The Unadapted System had a WER of 56.2%, while the Domain-Specific System achieved 11.7%.
UAV-Gesture[30]	Pose-based CNN (P-CNN)	P-CNN achieved an overall accuracy of 91.9% for gesture recognition. The dataset included 119 video clips, 37,151 annotated frames, and 13 gestures, providing a robust resource for gesture and action recognition research.
DarkTrack2021[31]	Novel ensembled method: SCT	The full implementation of SCT (Spatial-Channel Transformer) with all components enabled showed the highest improvement in tracking performance, with success rate and precision gains of 13.3% and 15.4%, respectively.
UAVDark135[32]	ADTrack, State of the art trackers (e.g. AutoTrack, SiamFC++, ARCF-HC, SiamRPN++)	ADTrack outperformed all other models in both bright and dark conditions, showing superior performance with the highest DP and AUC scores on the UAVDark135 dataset.
BioDrone[33]	KeepTrack, UAV-KT, Generic SOT Trackers	UAV-KT, designed for flapping-wing UAVs, showed a 5% improvement over KeepTrack in precision, normalized precision, and success scores. Generic SOT Trackers were compared for robustness and performance across various conditions.

4.1.3 Agriculture

UAV datasets help in monitoring crop health, assessing irrigation needs, and detecting pest infestations. Multispectral and hyperspectral imaging from UAVs enable detailed analysis of vegetation indices and soil properties.

4.1.4 Environmental Monitoring

UAVs are used to monitor forest health, wildlife, and water bodies. They provide data for studying ecological changes, tracking animal movements, and assessing the impacts of climate change.

4.1.5 Human Behavior Analysis

UAV datasets contribute to analyzing human activities and behaviors in public spaces. They are used for action recognition, pose estimation, and crowd monitoring, offering valuable insights for security and urban planning.

4.2 Variability of UAV databases

The diversity of UAV datasets lies in their varied data types, capture conditions, and application contexts. This diversity ensures that UAVs can address a wide range of tasks, each requiring specific data characteristics.

4.2.1 Data Types

UAV datasets include RGB images, infrared images, depth maps, and multispectral and hyperspectral images[64]. To capture complex scenarios for human behavior analysis, the UAV-Human dataset, for example, combines RGB videos, depth maps, infrared sequences, and skeleton data.

4.2.2 Capture Conditions

A variety of conditions, such as different times of day, weather, light (low light or varied lumination), and flight altitudes, are encountered when gathering UAV datasets. This variety makes sure that models that were trained on these datasets are strong and work well in a variety of settings.

4.2.3 Application Contexts

UAV datasets are tailored for specific applications. For example, visualizing data, object annotations, and flight data are used to address specific problems that come up when monitoring traffic from the air. Furthermore, the application of high-resolution images of the damage taken after the disaster, which enable accurate assessment of the damage.

4.3 Methods Applied to the UAV Dataset

Various methods are applied to UAV datasets to extract valuable insights and solve specific problems. These methods include machine learning, computer vision techniques, and advanced data processing algorithms. In Table 3 and 4, an overview of the methods used and the analysis of results are given to gain a better understanding.

4.3.1 Machine Learning and Deep Learning

Deep learning models, such as convolutional neural networks (CNNs)[65], are widely used for tasks like object detection, segmentation, and classification. For example:

- The RescueNet dataset employs models like PSPNet, DeepLabv3+, and Attention UNet for semantic segmentation to assess disaster damage.
- The UAVid Dataset presents deep learning baseline methods like Multi-Scale-Dilation net. The ERA dataset establishes a benchmark for event recognition in aerial videos by utilizing pre-existing deep learning models like the VGG models (VGG-16, VGG19)[16], Inception-v3[66], the ResNet models (ResNet-50, ResNet-101, and ResNet-152)[17], MobileNet, the DenseNet models (DenseNet-121, DenseNet-169, DenseNet-201)[24], and NASNet-L[67].

In the domain of deep learning, ensemble methods play a crucial role. They not only assess model performance but also boost accuracy while keeping the model's equilibrium intact. Such as:

- In VRAI dataset, they utilized ensemble techniques such as Triplet Loss, Contrastive Loss, ID Classification Loss, and Triplet + ID Loss, and introduced multi-task and multi-task + discriminative parts. These ensemble methods performed better than the state-of-the-art methods in their claim.

4.3.2 Transfer Learning

Transfer learning is used to leverage pre-trained models on UAV datasets, allowing for quicker and more efficient training. Like,

- Pre-trained YOLOv3-Tiny and MobileNetv2-SSDLite models, for example, are used for real-time object detection in the AU-AIR[19] dataset.

4.3.3 Event Recognition

Unmanned Aerial Vehicles (UAVs) have proven to be highly proficient in the field of event recognition and have gained significant popularity in this domain. Like for example:

- The ERA dataset has been subjected to various methods for event recognition in aerial videos, including DenseNet-201 and Inception-v3. These methods have demonstrated notable accuracy in identifying dynamic events from UAV footage.
- The BioDrone dataset assesses single object tracking (SOT) models and investigates new optimization approaches for the cutting-edge KeepTrack method for robust vision, which is presented by flapping-wing unmanned aerial vehicles[33].

4.3.4 Multimodal Analysis

Combining data from multiple sensors enhances the analysis capabilities of UAV datasets. The multimodal approach of the UAV-Human dataset, which combines RGB, infrared, and depth data, makes a thorough analysis of human behavior possible.

4.3.5 Creative Algorithms

New algorithms are created to tackle particular problems in the analysis of data from unmanned aerial vehicles. For example:

- The UAV-Gesture[30] dataset employs advanced gesture recognition algorithms to enable UAV navigation and control based on human gestures.
- The UAVDark135[32] makes use of ADTrack, a tracker that adapts to varying lighting conditions and makes use of discriminative correlation filters. It also has anti-dark capabilities.
- To address the issue of fisheye video distortions, the authors of the UAV-Human[10] dataset suggest a fisheye-based action recognition method that uses flat RGB videos as guidance.
- To classify disaster events from an unmanned aerial vehicle (UAV), the authors of the AIDER[15] dataset have created a lightweight convolutional neural network (CNN) architecture that they have named ERNet.
- VERI-Wild[27] introduces FDA-Net, a novel method for vehicle identification. It includes an embedding discriminator and a feature distance adversary network to enhance the model's capacity to differentiate between various automobiles.

4.3.6 Managing Diverse Conditions

Various environmental conditions, such as different lighting, weather, and occlusions, present challenges that are often addressed by methodologies. Like, DarkTrack2021 used the low-light enhancer-based method SCT to handle performance in low-light conditions.

The diversity of UAV datasets is a cornerstone of their utility, enabling a wide array of applications across different fields. From disaster management to human behavior analysis, the rich variety of data types, capture conditions, and application contexts ensures that UAV datasets can meet the specific needs of each task. The application of advanced methods, including deep learning, transfer learning, and multimodal analysis, further enhances the value derived from these datasets, pushing the boundaries of what UAVs can achieve in research and practical applications.

5 The Potential of Computer Vision Research in UAV Datasets

Unmanned Aerial Vehicles (UAVs) have greatly expanded the fields of computer vision research. UAV datasets offer unique and flexible data that is used in a range of computer vision tasks, from recognizing actions to finding objects. This section explores how UAV datasets are advancing computer vision research, contributing to various tasks from action recognition to object detection, as illustrated in Figure 2, which highlights the diverse applications and the development of new methods centered around these datasets.

5.1 Leveraging UAV Datasets for Computer Vision Applications

Human behavior analysis, emergency response, tracking at night, surveillance, and many other uses can be done with UAV datasets in computer vision. These are some of the areas where UAV datasets are used, along with an example of how to describe a dataset based on the datasets we talked about in our research paper.

Table 5: Summary of Methods Employed on Uav Datasets and Their Benefits

Employed Method	**Name of the Dataset**	**Benefit from the Use of Method**
Attention UNet28, ViT-Tiny, ViT-Small	RescueNet[5]	Improved disaster response strategies, enhanced model performance in segmentation tasks through transfer learning
Fisheye-based action recognition approach, HigherHRNet, AlphaPose	UAV-Human[10]	Robust models for human behavior understanding
ERNet	AIDER[15]	High performance with minimal memory requirements, suitable for real-time aerial image classification
YOLOv3-Tiny, MobileNetv2-SSDLite	AU-AIR[19]	Real-time object detection on UAVs, bridging the gap between computer vision and robotics
DenseNet-201, I3D-Inception-v1, TRN-Inception-v3	ERA[23]	High performance in single-frame and video classification tasks
Multi-Scale-Dilation net, FSO, 3D CRF	UAVid[25]	Enhanced semantic segmentation performance in urban scenes, addressing large-scale variation and moving object recognition
Convolutional and connection layers, weight matrices, weighted pooling	VRAI[26]	Superior vehicle re-identification performance
Aggregating tree-wise F1 scores, weighting coefficients for averaging F1 scores	FOR-instance[34]	Improved methods for individual tree segmentation, crucial for understanding forest ecosystems
FDA-Net (Feature Distance Adversary Network)	VERI-Wild[27]	Enhanced discriminative capability in vehicle re-identification tasks
Smooth silhouette loss	UAV-Assistant (UAVA)[28]	Improved performance in 3D pose estimation tasks
Time delay neural network, domain adaptation techniques	KITE[29]	Enhanced UAV command recognition systems through visual context and domain adaptation
Pose-based Convolutional Neural Network (P-CNN)	UAV-Gesture[30]	High accuracy in gesture recognition for UAV control
Spatial-Channel Transformer, curve projection model	DarkTrack2021[31]	Improved nighttime UAV tracking accuracy by enhancing low-light images
Illumination adaptive, anti-dark capabilities, efficient image enhancer	UAVDark135[32]	Superior performance in all-day aerial object tracking, adaptability to different light conditions
KeepTrack-optimized UAV-KT	BioDrone[33]	Addresses challenges in tracking tiny targets with drastic appearance changes, providing a robust benchmark for vision research

5.1.1 Human Behavior Understanding and Gesture Recognition

The UAV-Human platform is essential for utilizing UAVs to study human behavior, including a range of conditions and perspectives for pose estimation and action recognition. This dataset contains multi-modal information, including skeleton, RGB, infrared, and night vision modalities. Essential for UAV control and gesture identification, UAV-Gesture contains 119 high-definition video clips with 13 gestures for command and navigation that are marked with body joints and gesture classes. Because this dataset was captured outside, it has more practical UAV control applications because of the variations in phase, orientation, and body shape.

5.1.2 Emergency Response and Disaster Management

RescueNet provides detailed pixel-level annotations and high-resolution images for 10 classes, including buildings, roads, pools, and trees. It is designed for post-disaster damage assessment using UAV imagery. It supports semantic segmentation using state-of-the-art models, enhancing natural disaster response and recovery strategies. AIDER focuses on classifying disaster events, utilizing images of traffic accidents, building collapses, fires, and floods to support real-time disaster management applications by training convolutional neural networks (CNNs).

5.1.3 Traffic Surveillance and Vehicle Re-Identification

In traffic surveillance, AU-AIR prioritizes real-time performance and offers annotations for a variety of object categories, including cars, buses, and pedestrians. It bridges the gap between computer vision and robotics by offering multi-modal sensor data for advanced research in data fusion applications. VRAI is the largest UAV-based vehicle re-identification dataset, containing over 137,613 images of 13,022 vehicles with annotations for unique IDs, color, vehicle type, attributes, and discriminative parts. It supports vehicle ReID tasks with diverse scenarios and advanced algorithms. VERI-Wild, which contains over 400,000 photos of 40,000 vehicles taken by 174 CCTV cameras in various urban settings, is essential for research on vehicle re-identification. It uses techniques like FDA-Net to improve ReID accuracy by addressing variations in backgrounds, illumination, occlusion, and viewpoints.

5.1.4 Event Recognition and Video Understanding

For training models in event recognition in UAV videos, ERA contains 2,864 labeled video snippets for 24 event classes and 1 normal class that were gathered from YouTube. This dataset captures dynamic events in various conditions, supporting temporal event localization and video retrieval tasks.

5.1.5 Nighttime tracking and low-light conditions

Including 110 annotated sequences with over 100,000 frames, DarkTrack2021 is crucial for improving UAV tracking at night. By employing spatial-channel transformers (SCT) and non-linear curve projection models, it improves the quality of low-light images and offers a thorough assessment framework. The UAVDark135 dataset and the ADTrack algorithm are designed for all-day aerial tracking. ADTrack performs well in low light and adjusts to various lighting conditions thanks to its discriminative correlation filter foundation. More than 125,000 frames, specially annotated for low-light tracking scenarios, are included in the UAVDark135 dataset.

5.1.6 Object Tracking and Robust Vision

With 600 videos and 304,209 manually labeled frames, BioDrone is a benchmark for single object tracking with bionic drones. It captures challenges such as camera shake and drastic appearance changes, supporting robust vision analyses and evaluations of various single object tracking algorithms.

5.1.7 Urban Scene Segmentation and Forestry Analysis

UAVid provides annotations for eight classes and 30 high-resolution video sequences in 4K resolution to address segmentation challenges in urban scenes. It uses models such as Multi-Scale-Dilation net to support tasks like population density analysis and traffic monitoring. FOR-instance provides UAV-based laser scanning data for tree instance segmentation and is intended for use in point cloud segmentation in forestry. It facilitates benchmarking and method development by supporting both instance and semantic segmentation.

5.1.8 Multimodal Data Synthesis and UAV Control

UAV-Assistant facilitates monocular pose estimation by introducing a multimodal dataset featuring exocentric and egocentric views. It enhances 3D pose estimation tasks with novel smooth silhouette loss function and differentiable

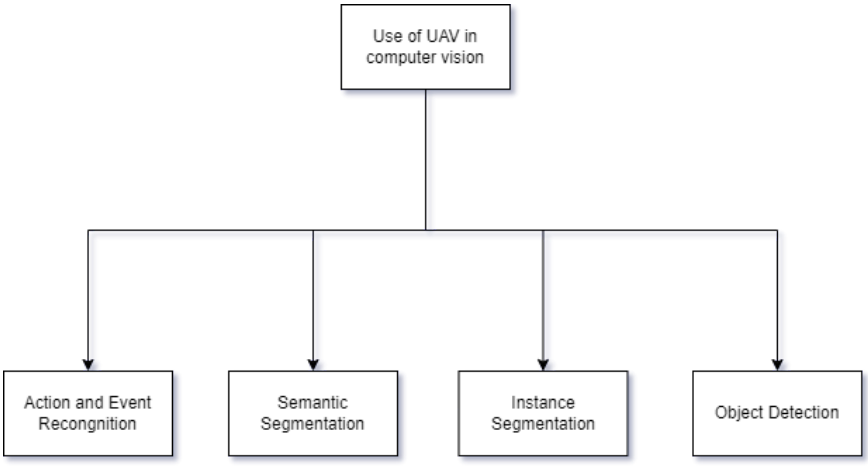


Figure 2: Diverse Applications of UAV Datasets in Computer Vision Research

rendering techniques. KITE incorporates spoken commands, audio, and images to enhance UAV control systems. It includes commands recorded by 16 speakers, supporting movement, camera-related, and scenario-specific commands with multi-modal approaches.

Together, these datasets improve a wide range of computer vision applications, including robust vision in difficult conditions, real-time traffic surveillance, emergency response, and human behavior analysis.

5.2 Development of Novel Methods Using UAV Datasets

UAV datasets have spurred the development of innovative methods in computer vision. As an example, the Guided Transformer I3D framework, which addresses distortions through unbounded transformations guided by flat RGB videos, was developed using the UAV-Human dataset. This framework enhances action recognition performance in fisheye videos. This approach is a prime example of how UAV datasets drive the creation of specialized algorithms to address particular difficulties brought about by aerial viewpoints.

The DarkTrack2021 benchmark introduces a Spatial-Channel Transformer (SCT) for enhancing low-light images in nighttime UAV tracking. Meanwhile, Bowen Li and team present the UAVDark135 dataset and the ADTrack algorithm for all-day aerial object tracking. ADTrack, equipped with adaptive illumination and anti-dark capabilities, outperforms other trackers in both well-lit and dark conditions. It processes over 30 frames per second on a single CPU, ensuring efficient tracking under various lighting conditions. The study emphasizes how crucial image illuminance data is and suggests a useful image enhancer to improve tracking performance in all-day situations.

For emergency response applications, the AIDER dataset has facilitated the development of ERNet, a lightweight CNN architecture optimized for embedded platforms. ERNet's architecture, which incorporates downsampling at an early stage and efficient convolutional layers, allows for real-time classification of aerial images on low-power devices. This showcases the practical use of UAV datasets in disaster management.

The VERI-Wild dataset introduces a novel approach called FDA-Net for vehicle reidentification. This method utilizes a unique type of network to generate difficult negative examples in the feature space. On the other hand, the VRAI dataset has developed a specialized vehicle ReID algorithm that leverages detailed annotation information to explicitly identify unique parts for each vehicle instance in object detection.

Ultimately, UAV datasets are essential in the field of computer vision research, providing distinct data that is invaluable for a diverse array of applications. They allow for the development of novel methods tailored to the specific challenges and opportunities presented by UAV technology, accelerating progress in areas such as human behavior analysis, emergency response, and nighttime tracking.

6 Constraints of UAVs

While Unmanned Aerial Vehicles (UAVs) have significantly advanced data collection and analysis in numerous fields, they are not without limitations, particularly concerning the datasets they generate. This section delves into the primary constraints associated with UAV datasets, emphasizing their impact on the field and suggesting areas for improvement.

6.1 Data Quality and Consistency

One of the most pressing limitations of UAV datasets is the inconsistency in data quality. Weather, time of day, and UAV stability are just a few variables that can affect the quality of data that UAVs collect. Such as, datasets collected during poor weather conditions or at night may need more visibility and increased noise, complicating subsequent analysis and model training. Even with advancements like low-light image enhancers and specialized algorithms for nighttime tracking, these solutions often need improvement and require further refinement to match the reliability of daytime data.

6.2 Limited Scope and Diversity

UAV datasets often need more diversity in terms of geographic locations, environmental conditions, and the variety of captured objects. Many existing datasets, such as AU-AIR and ERA, focus heavily on specific scenarios like urban traffic surveillance or disaster response, which limits their generalizability to other contexts. Additionally, datasets such as UAV-Human and UAVDark135 tend to feature limited subject diversity and controlled environments, which may not accurately represent real-world conditions. This lack of diversity can lead to models that perform well in specific conditions but struggle in untested environments.

6.3 Annotation Challenges

The process of annotating UAV datasets is often time-consuming and labor-intensive. High-resolution images and videos captured by UAVs require detailed, pixel-level annotations, which are essential for tasks like semantic segmentation and object detection. This is clearly seen in datasets such as RescueNet and FOR-Instance, where the annotation process is recognized as a major bottleneck. The intensive labor required for comprehensive annotation limits the availability of large, well-labeled datasets, which are crucial for training robust machine learning models.

6.4 Computational and Storage Demands

The high resolution and large volume of data generated by UAVs pose significant computational and storage challenges. Processing and analyzing large-scale UAV datasets demand substantial computational resources and advanced hardware, which may only be readily available to some researchers. For example, the dense and high-resolution images in datasets like UAVid and BioDrone require extensive processing power for effective utilization. Additionally, the storage of such vast amounts of data can be impractical for some institutions, hindering widespread access and collaboration.

6.5 Integration with Other Data Sources

Another limitation is the integration of UAV datasets with other data sources. While multimodal datasets that combine UAV data with other sensor inputs (such as satellite imagery, GPS data, and environmental sensors) provide richer insights, they also introduce complexity in data alignment and fusion. The AU-AIR dataset, which includes visual data along with GPS coordinates and IMU data, exemplifies the potential and challenges of such integration. Ensuring the synchronized and accurate fusion of data from multiple sources remains a technical hurdle that needs addressing.

6.6 Real-Time Data Processing

The ability to process and analyze UAV data in real-time is critical for applications like disaster response and surveillance. However, achieving real-time processing with high accuracy is challenging due to the aforementioned computational demands. Models such as those evaluated in the DarkTrack2021 and UAVDark135 datasets show promise but often require optimization to balance speed and accuracy effectively. Real-time processing also necessitates robust algorithms capable of handling dynamic environments and changing conditions without significant delays.

6.7 Ethical and Legal Considerations

Finally, the use of UAVs and their datasets is subject to various ethical and legal considerations. Issues such as privacy, data security, and regulatory compliance must be addressed to ensure responsible and lawful use of UAV technology.

These considerations can limit the scope of data collection and usage, particularly in populated areas or sensitive environments, thereby constraining the availability and applicability of UAV datasets.

Despite the transformative potential of UAV datasets across various disciplines, their limitations must be acknowledged and addressed to maximize their utility. Improving data quality, enhancing dataset diversity, streamlining annotation processes, and overcoming computational and storage challenges are essential steps. Additionally, integrating UAV data with other sources, advancing real-time processing capabilities, and adhering to ethical and legal standards will ensure that UAV datasets can be effectively leveraged for future research and applications. By tackling these limitations, the field can fully harness the power of UAV technology to drive innovation and deepen our understanding of complex, dynamic environments from an aerial perspective.

7 Prospects for Future UAV Research

Future studies on UAV datasets need to focus on a few crucial areas to improve their usefulness and cross-domain applicability as the field grows. The following suggestions highlight the crucial paths for creating UAV datasets and maximizing their potential for future innovations.

7.1 Enhancing Dataset Diversity and Representativeness

Further investigations ought to concentrate on generating more varied and representative UAV datasets. This involves capturing data in a wider range of environments, weather conditions, and geographic locations to ensure models trained on these datasets are robust and generalizable. To obtain comprehensive data for tasks like environmental monitoring, urban planning, and disaster response, datasets can be expanded to include a variety of urban, rural, and natural settings.

7.2 Incorporating Multimodal Data Integration

Integrating multiple data modalities, such as thermal, infrared, LiDAR[68], and hyperspectral[64] imagery, can significantly enrich UAV datasets. In the future, these data types should be combined to create multimodal datasets that provide a more comprehensive view of the scenes that were recorded. This integration can improve the accuracy of applications such as vegetation analysis, search and rescue operations, and wildlife monitoring.

7.3 Advancing Real-Time Data Processing and Transmission

For applications like emergency response and traffic monitoring that demand quick analysis and decision-making, developing techniques for real-time data processing and transmission is essential. Future research should focus on optimizing data compression, transmission protocols, and edge computing techniques to enable swift and efficient data handling directly on UAVs.

7.4 Improving Annotation Quality and Efficiency

High-quality annotations are vital for the effectiveness of UAV datasets in training machine learning models. Future studies should investigate automated and semi-automated annotation tools that leverage AI to reduce manual labor and improve annotation accuracy. Additionally, crowdsourcing and collaborative platforms can be utilized to gather diverse annotations, further enhancing dataset quality.

7.5 Addressing Ethical and Privacy Concerns

As UAVs become more prevalent, addressing ethical and privacy issues becomes increasingly important. Guidelines and frameworks for the ethical use of UAV data should be established by future research, especially for applications involving surveillance and monitoring. It is important to focus on creating methods that protect privacy and collect data in a way that respects regulations and earns the trust of the public.

7.6 Expanding Application-Specific Datasets

The creation of customized datasets for specific uses can effectively boost new ideas in certain areas. For instance, datasets focused on agricultural monitoring, wildlife tracking, or infrastructure inspection can provide domain-specific insights and improve the precision of related models. To address the specific needs of various industries, future research should give priority to developing such targeted datasets.

7.7 Enhancing Interoperability and Standardization

Standardizing data formats and annotation protocols across UAV datasets can make it easier for researchers and developers to use and make the datasets more interoperable. Future efforts should aim to establish common standards and benchmarks, enabling the seamless integration of datasets from various sources and promoting collaborative research efforts.

7.8 Utilizing Advanced Machine Learning Techniques

The application of cutting-edge machine learning techniques, such as deep learning and reinforcement learning, to UAV datasets holds immense potential for advancing UAV capabilities. Future research should explore innovative algorithms and models that can leverage the rich data provided by UAVs to achieve breakthroughs in areas like autonomous navigation, object detection, and environmental monitoring.

7.9 Leveraging Advanced Machine Learning Techniques

Longitudinal studies that collect UAV data over long periods of time can give us useful information about how things change over time in different settings. Future research should emphasize continuous data collection efforts to monitor changes in ecosystems, urban developments, and disaster-prone areas, enabling more informed and proactive decision-making.

7.10 Fostering Collaborative Research and Open Data Initiatives

Encouraging collaboration among researchers, institutions, and industries can accelerate advancements in UAV datasets. Open data initiatives that make UAV datasets public should be supported by future research. These initiatives will encourage innovation and allow a wider range of researchers to contribute to and use these resources.

By addressing these future research directions, the field of UAV datasets can continue to evolve, offering increasingly sophisticated tools and insights that drive progress across multiple domains. UAV datasets are still being improved and added to, which is very important for getting the most out of UAV technology and making room for new discoveries and uses.

8 Results and Discussion of Reviewed Papers

The datasets discussed in this section represent the application of the papers reviewed in this survey. Our analysis of the datasets revealed that KITE, RescueNet, and Biodrone are relatively new and have not been thoroughly investigated in the literature. While one of the datasets we reviewed, ERA, is not very recent, it still lacks the enough amount of study to fully emphasize its potential. The datasets included in our review were selected based on the number of citations their associated papers have received, emphasizing those with higher citation counts. We delved into several papers that make compelling use of the datasets we evaluated. In our examination, we carefully reviewed the details of the analysis of results and experiments conducted by other researchers. These researchers utilized the datasets we assessed as benchmarks and applied various methods. We have included the best results for the methods applied to the datasets we reviewed in this section and in Table 6, 7 and 8.

8.1 AU-AIR

In their study, Jiahui et al.[69] selected AU-AIR as a benchmark dataset to create their proposed real-time object detection model, RSSD-TA-LATM-GID, specifically designed for small-scale object detection. The performance of their model surpassed that of YOLOv4[115] and YOLOv3[116]. The researchers employed the MobileNetv-SSDLite ensemble approach, which yielded the lowest mean average precision (mAP) score.

Walambe et al.[71] employed baseline models on the AU-AIR dataset as one of their evaluative benchmarks. The objective of the study was to demonstrate the attainability of different techniques and ensemble techniques in the detection of objects with varying scales. The baseline technique yielded the highest performance, with a mean average precision (mAP) score of 6.63%. This outcome was achieved by employing color-augmentation on the dataset. The

[1] "Plot" here refers to the forest types that were looked at in the FOR-instance dataset release.

[2] DP: Dynamic Precision

[3] NDP: Normalized Dynamic Precision

[4] Top1/Top5: The authors of [84], employed the metric of Top1/Top5 for measuring accuracy in single-label classification.

Table 6: Performance Metrics and Results for Different Datasets and Methods

Dataset Name	Reference	Methods	Performance
AU-AIR[19]	[69]	YOLOv3	mAP: 59.83; Speed(FPS): 29
		YOLOv4	mAP: 67.35; Speed(FPS): 24
		RSSD-TA-LSTM-GID	mAP: 71.68; Speed(FPS): 23
	[70]	res2net50	mAP: 88.93; Speed(FPS): 45.73
		rs2net101	mAP: 90.52; Speed(FPS): 7.21
		hourglass-104	mAP: 91.62; Speed(FPS): 7.19
	[71]	RetinaNet	Voting Strategy: Unanimous; mAP(%): 6.63
		YOLO + RetinaNet	Voting Strategy: Consensus; mAP(%): 3.69
		RetinaNet + SSD	Voting Strategy: Consensus; mAP(%): 4.03
	[72]	Faster R-CNN	mAP(%): 13.77
		SSD	mAP(%): 9.1
		YOLOv3	mAP(%): 13.33
		YOLOv4	mAP(%): 25.94
FOR-instance[34]	[73]	PointNet	mIoU: 35.65; micro F1: 52.56
		PointNet++	mIoU: 33.00; micro F1: 49.57
		Point Transformers	mIoU: 22.97; micro F1: 37.13
	[74]	HFC (on CULS plot[1])	Precision: 0.89; Recall: 0.8; F1 score: 0.84
		HFC (on NIBIO plot)	Precision: 0.89; Recall: 0.85; F1 score: 0.87
		HFC (on NIBIO2 plot)	Precision: 0.85; Recall: 0.85; F1 score: 0.85
		HFC (on SCION plot)	Precision: 0.95; Recall: 0.90; F1 score: 0.92
		HFC (on RMIT plot)	Precision: 0.89; Recall: 0.85; F1 score: 0.87
		HFC (on TUWEIN plot)	Precision: 0.84; Recall: 0.80; F1 score: 0.82
UAV-Assistant[28]	[75]	BPnP[76]	ACC2: 95.2; ACC5: 98.36 ACC2: 55.31; ACC5: 85.34
		HigherHRNet[77]	ACC2: 89.92; ACC5: 97.75
		HRNet[78]	ACC2: 90.75; ACC5: 98.04

Table 7: Performance Metrics and Results for Different Datasets and Methods

Dataset Name	Reference	Methods	Performance
AIDER[15]	[79]	EmergencyNet	memory(MB): 0.368; F1 Score(%): 95.7
		VGG16	memory(mB): 59.39; F1 Score(%): 96.4
		ResNet50	memory(MB): 96.4; F1 Score(%): 96.1
	[80]	AISCC-DE2MS	MSE: 0.042; PSNR: 61.898
		Genetic Algorithm	MSE: 0.06; PSNR: 60.349
		Cat Swarm Algorithm	MSE: 0.12; PSNR: 57.339
		Artificial Bee Colony Algorithm	MSE: 0.165; PSNR: 55.956
DarkTrack2021[31]	[81]	SAM-DA-Track	AUC: 0.451; Precision (normalized): 0.524; Precision: 0.593
		UDAT	AUC: 0.421; Precision (normalized): 0.499; Precision: 0.570
		SiamBAN	AUC: 0.422; Precision (normalized): 0.491; Precision: 0.566
	[82]	SiamAPN	DP^2: 0.43; NDP^3: 0.389; AUC: 0.446
		SiamAPN++	DP: 0.494; NDP: 0.446; AUC: 0.375
UAV-Human[10]	[83]	Proposed Novel Method	Precision: 0.49; Recall: 0.49; F1 Score: 0.48
	[84]	CLIP[85]	Top1/$Top5^4$ (Filtering ratio 90%): 1.79 / 7.05
		ViFi CLIP[86]	Top1/Top5 (Filtering ratio 90%): 4.67 / 15.18
	[87]	2s-MS&TA-HGCN-FC (Novel method)	CSv1: 44.33; CSv2: 70.69
		4s-MS&TA-HGCN-FC (Novel method)	CSv1: 45.72; CSv2: 71.84
		FR-AGCN[88]	CSv1: 43.98; CSv2: 69.5
UAVDark135[32]	[89]	DCPT	Success Rate: 0.577; Precision: 0.703; Normalized Precision: 0.701
		DIMP50-SCT	Success Rate: 0.562; Precision: 0.717; Normalized Precision: 0.71
		DIMP18[90]	Success Rate: 0.542; Precision: 0.702; Normalized Precision: 0.69
	[91]	DL+SiamAPN	Success Rate: 0.389; Precision: 0.516
		SiamAPN[92]	Success Rate: 0.3; Precision: 0.424
		DL+DIMP50	Success Rate: 0.544; Precision: 0.7
		DIMP50[93]	Success Rate: 0.526; Precision: 0.672
VRAI[26]	[94]	Proposed Novel Method	mAP: 0.828; R-1 Accuracy: 0.844
		TransReID[95]	mAP: 0.814; R-1 Accuracy: 0.826
	[95]	RotTrans	mAP: 0.848; R-1 Accuracy: 0.838
		TransReID	mAP: 0.786; R-1 Accuracy: 0.803

Table 8: Performance Metrics and Results for Different Datasets and Methods

Dataset Name	Reference	Methods	Performance
UAV-Gesture[30]	[96]	Novel Multifeature+CNN method	Accuracy: 0.95
		P-CNN[97]	Accuracy: 0.91
		MLP_7j[98]	Accuracy: 0.94
	[98]	DD-Net_7j[99]	Accuracy: 0.915
		P-CNN	Accuracy: 0.919
		MLP_7j	Accuracy: 0.948
UAVid[25]	[100]	BANet	mIoU(%): 64.6
		MSD benchmark[25]	mIoU(%): 57.0
	[101]	A²-FPN	mIoU(%): 65.7
		MSD benchmark	mIoU(%): 57.0
	[102]	UNetFormer	mIoU(%): 67.8
		ABCNet	mIoU(%): 63.8
		BANet	mIoU(%): 64.6
		BoTNet	mIoU(%): 63.2
	[103]	MSD benchmark	mIoU(%): 57.0; FPS: 1.00
		BiSeNet[104]	mIoU(%): 61.5; FPS: 11.08
		CAN	mIoU(%): 63.5; FPS: 15.14
VERI-Wild[27]	[105]	FDA-Net[106]	mAP(small): 0.351; mAP(medium): 0.298; mAP(large): 0.228
		PVEN	mAP(small): 0.825; mAP(medium): 0.77; mAP(large): 0.697
	[107]	MLSL[108]	mAP(large): 0.366; R-1 accuracy (large): 0.775
		FastReID	mAP(large): 0.773; R-1 accuracy (large): 0.925
	[109]	GiT	mAP(T10000): 0.675; R-1 accuracy (T10000): 0.854
		PCRNet[110]	mAP(T10000): 0.671; R-1 accuracy (T10000): 0.85
	[111]	HPGN	mAP(T10000): 0.65; R-1 accuracy (T10000): 0.8268
		Triplet Embedding[112]	mAP(T10000): 0.516; R-1 accuracy (T10000): 0.699
	[113]	Baseline[114]	mAP(large): 0.65; R-1 accuracy (large): 0.95
		SAVER	mAP(large): 0.677; R-1 accuracy (large): 0.958

performance metrics for the ensemble methods YOLO+RetinaNet and RetinaNet+SSD were found to be 3.69% and 4.03%, respectively. The authors Saeed et al.[70] made modifications to the architecture of the CenterNet model by using other Convolutional Neural Networks (CNNs) as backbones, such as resnet18, hourglass-104, resnet101, and res2net101. Among all the CNNs as backbone. The findings are presented in Table 6.

Gupta and Verma in their paper [72] utilized the AU-AIR data as a reference point, employing a range of advanced models to achieve precise and automated detection and classification of road traffic. The YOLOv4 model achieved the highest mean average precision (mAP) score of 25.94% on the AU-AIR dataset. The Faster R-CNN and YOLOv3 models achieved the second and third highest maximum average precision (mAP) scores, with values of 13.77% and 13.33% respectively.

8.2 FOR-instance

Bountos et. al. extensively utilized the "FOR-Instance" dataset in their study, [73], while introducing their innovative approach FoMo-Net. The dataset was utilized to analyze point cloud representations obtained from LiDAR sensors in order to gain a deeper understanding of tree geometry. Existing baseline techniques such as PointNet, PointNet++, and Point Transformer were employed to accomplish these objectives on aerial modality. The corresponding findings are presented in Table 7. In a separate paper, Zhang et. al.[74] used the "FOR-instance" dataset to train their proposed HFC algorithm and compare its performance with other established approaches. The authors utilized several techniques and ensemble approaches (Xing2023, HFC+Xing2023, HFC+Mean Shift, HFC) on several forest types (CULS, NIBIO, NIBIO2, SCION, RMIT, TUWIEN) shown in the FOR-instance dataset. Among all the methods, HFC demonstrated superior performance. The optimal outcomes achieved by the HFC approach on various forest types represented in the FOR-instance dataset are presented in Table 7.

8.3 UAV-Assistant

Albanis et al. used the UAV-Assistant dataset as benchmark for their research, [75]. They conducted a comparative analysis of BPnP[76] and HigherHRNet's[77] 6DOF object pose estimation using several different criteria. Analysis revealed that loss functions play a crucial role in posture estimation. Specifically, the l_p loss function outperformed the l_h loss function, particularly in the case of the M2ED drone, resulting in improved accuracy metrics. HigherHRNet demonstrated greater performance compared to HRNet[78] on smaller objects such as the Tello drone, but not on the M2ED drone, indicating its potential superiority under smaller object classifications. Their analysis of qualitative heatmaps revealed that the l_p loss function performed better than the Gaussian-distributed l_h model in accurately locating keypoints. Table 7 displays the accuracy metrics (ACC2 and ACC5) obtained from the research conducted by Albanis and his colleagues. In the case of BPnP, we have included the accuracy for both M2ED and Tello drones respectively, as they achieved the highest accuracy outcomes. Regarding HRNet and HigherHRNet, they achieved the best accuracy specifically for M2ED.

8.4 AIDER

The AIDER dataset has been utilized as a benchmark by Alrayes et al. and the authors of AIDER in developing their innovative method, "EmergencyNet." In their paper, [79] various pre-trained models were applied to the AIDER dataset, with the best F1 accuracy achieved using VGG16 (96.4%) and ResNet50 (96.1%). However, the memory consumption for VGG16 and ResNet50 was quite high, at 59.39MB and 96.4MB respectively. However, EmergencyNet achieved 95.7% F1 accuracy with only 0.368MB of RAM. ResNet50 had nearly 24 million parameters, while VGG16 had 14.8 million. Alrayes et al. benchmarked their AISCC-DE2MS model with AIDER. They found that their algorithm outperformed the genetic, cat-swarm, and artificial bee colony algorithms. MSE and PSNR were utilized to evaluate. These methods were used to compare five photos to evaluate the model's performance. The best result from the five photos is shown in Table 6.

8.5 DarkTrack2021

Changhong Fu and his team utilized the DarkTrack2021 benchmark as a foundation for developing the Segment Anything Model (SMA) powered framework SAM-DA. Their research [81] focused on effectively addressing illumination variation and low ambient intensity. They conducted a comparative analysis between their model and various methods, particularly the Baseline tracker UDAT[117] method. Their novel approach outperformed the Baseline UDAT method, achieving substantial improvements of 7.1% in illumination variation and 7.8% in low ambient intensity. The authors evaluated 15 state-of-the-art trackers and found that SAM-DA demonstrated the most promising results. Additionally, Changhong Fu delved into Siamese Object Tracking in their another study [82], highlighting the significance of UAVs

in visual object tracking. They also leveraged the DarkTrack2021 datasets as a benchmark to assess model performance in low-illumination conditions, with detailed results and the applied models presented in Table 6.

8.6 UAV-Human

Azmat et al.[83] address UAV-captured data-based human action recognition (HAR) challenges and approaches in their UAV-Human dataset research. Azmat et al. evaluated their HAR system on 67,428 video sequences of 119 people in various contexts from the UAV-Human dataset. The approach has a mean accuracy of 48.60% across eight action classes, indicating that backdrops, occlusions, and camera motion hinder human movement recognition in this dataset. Lin et al.[84] studied text bag filtering techniques for model training, emphasizing data quality. Their ablation study indicated that text bag filtering ratio influences CLIP matching accuracy and zero-shot transfer performance. Filtering training data improved model generalization, especially in unsupervised learning. Huang et al.[87] evaluated the 4s-MS&TA-HGCN-FC skeleton-based action recognition model on the UAV-Human dataset. The model achieved 45.72% accuracy on the CSv1 benchmark and 71.84% on the CSv2 test, surpassing previous state-of-the-art techniques. They found that their technique can manage UAV-captured data's viewpoints, motion blurring, and resolution changes.

8.7 UAVDark135

Zhu et al.[89] and Ye et al.[91] used the UAVDark135 dataset to evaluate their strategies for increasing low-light tracking performance. The Darkness Clue-Prompted Tracking (DCPT) approach by Zhu et al. showed considerable gains, reaching a 57.51% success rate on UAVDark135. A 1.95% improvement over the base tracker demonstrates the effectiveness of including darkness clues. Additionally, DCPT's gated feature aggregation approach increased success score by 2.67%, making it a reliable nighttime UAV tracking system. Ye et al.'s DarkLighter(DL) approach improved tracking performance on the UAVDark135 dataset. DL improved SimpAPN[118][92] tracker AUC by over 29% and precision by 21%. It also worked well across tracking backbones, enhancing precision and success rates in light variation, quick motion, and low resolution circumstances. DL surpassed modern low-light enhancers like LIME by 1.68% in success rate and 1.45% in precision.

8.8 VRAI

VRAI was utilized to establish a vehicle re-identification baseline. Syeda Nyma Ferdous, Xin Li, and Siwei Lyu [94] tested their uncertainty-aware multitask learning framework on this dataset and achieved 84.47% Rank-1 accuracy and 82.86% mAP. This model's capacity to handle aerial image size and position fluctuations was greatly improved by multiscale feature representation and a Pyramid Vision Transformer (PVT) architecture. Shuoyi Chen, Mang Ye, and Bo Du[95] focused on vehicle ReID using VRAI. RotTrans, a rotation-invariant vision transformer, surpassed current innovative approaches by 3.5% in Rank-1 accuracy and 6.2% in mean average precision (mAP). This approach solved UAV-based vehicle ReID challenges that typical pedestrian ReID methods struggle with. The process was further complicated by the need to present results in a certain format for performance evaluation.

8.9 UAV-Gesture

Usman Azmat et al.[96] and Papaioannidis et al.[98] utilized the UAV-Gesture dataset to evaluate their recommendations for human action and gesture recognition. They used the UAV-Gesture collection of 119 high-definition RGB movies representing 13 unique motions used to control UAVs. The dataset is ideal for testing recognition systems due to its diversity of views and movement similarities. The Usman Azmat et al. method achieved 0.95 action recognition accuracy on the UAV-Gesture dataset. Mean precision, recall, and F1-score for the system were 0.96, 0.95, and 0.94. Several investigations supported by confusion matrices showed the system's ability to distinguish gestures. Papaioannidis et al. found that their gesture recognition method outperformed DD-Net[119] and P-CNN[120] by 3.5% in accuracy. The authors stressed the need of using 2D skeletal data from movies to boost recognition accuracy. Real-time performance makes their method suitable for embedded AI hardware in dynamic UAV situations.

8.10 UAVid

The UAVid dataset has been extensively utilized as a benchmark by several researchers in the development of innovative methods for semantic segmentation in urban environments. Wang et al.[100] introduced the Bilateral Awareness Network (BANet) and applied it to the UAVid dataset, achieving a notable mean Intersection-over-Union (mIoU) score of 64.6%. BANet's ability to accurately segment various classes within high-resolution urban scenes was demonstrated through both quantitative metrics and qualitative analysis, outperforming other state-of-the-art models like the MSD benchmark.

Similarly, Rui Li et al.[101] proposed the Attention Aggregation Feature Pyramid Network (A^2-FPN) and reported significant improvements on the UAVid dataset. A^2-FPN achieved the highest mIoU across five out of eight classes, surpassing BANet by 1% in overall performance. The model's effectiveness was particularly evident in its ability to correctly identify moving vehicles, a challenging task for many segmentation models.

Libo Wang et al.[102] introduced the UNetFormer, which further pushed the boundaries of semantic segmentation on the UAVid dataset. Achieving an impressive mIoU of 67.8%, the UNetFormer outperformed several advanced networks, including ABCNet[121] and hybrid Transformer-based models like BANet and BoTNet[122]. The UNetFormer demonstrated a strong ability to handle complex segmentation tasks, particularly in accurately identifying small objects like humans.

Lastly, Michael Ying Yang et al.[103] applied the Context Aggregation Network(CAN) to the UAVid dataset, achieving a mIoU score of 63.5% while maintaining a high processing speed of 15 frames per second (FPS). This model was noted for its ability to maintain consistency in both local and global scene semantics, making it a competitive choice for real-time applications in urban environments.

8.11 VERI-Wild

The VERI-Wild dataset has been extensively utilized as a benchmark by several researchers in the development of innovative methods for vehicle re-identification (ReID) in real-world scenarios. Meng et al.[105] introduced the Parsing-based View-aware Embedding Network (PVEN) and applied it to the VERI-Wild dataset, achieving significant improvements in mean Average Precision (mAP) across small, medium, and large test datasets, with increases of 47.4%, 47.2%, and 46.9%, respectively. PVEN's ability to perform view-aware feature alignment allowed it to consistently outperform state-of-the-art models, particularly in Cumulative Match Characteristic (CMC) metrics, where it showed a 32.7% improvement over FDA-Net at rank 1.

Similarly, Lingxiao He et al.[107] evaluated the FastReID toolbox on the VERI-Wild dataset, highlighting its effectiveness in accurately identifying vehicles across various conditions. FastReID achieved state-of-the-art performance, particularly in Rank-1 accuracy(R1-Accuracy) and mAP, showcasing its robustness in handling the complexities of vehicle ReID tasks in surveillance and traffic monitoring environments.

Fei Shen et al.[109] applied the GiT method on the VeRi-Wild dataset, securing top performance across all testing subsets, including Test3000(T3000), Test5000(T5000), and Test1000(T1000). The GiT method outperformed the second-place method, PCRNet, by 0.41% in Rank-1 identification rate and 0.45% in mAP on the Test1000 subset. The study emphasized the importance of leveraging both global and local features, as GiT demonstrated superior generalization across different datasets and conditions. In a separate study, Fei Shen et al.[111] developed the Hybrid Pyramidal Graph Network (HPGN) approach, which achieved the highest Rank-1 identification rate among the evaluated methods on the VERI-Wild dataset, so making more contributions to the advancing field of vehicle ReID. The findings highlighted the resilience of HPGN, especially in difficult circumstances such as fluctuating day and night situations, where alternative approaches exhibited a decrease in effectiveness.

Lastly, Khorramshahi et al.[113] presented a residual generation model that improved mAP by 2.0% and CMC1 by 1.0% compared to baseline models. The model's reliance on residual information, as indicated by a high alpha value (α = 0.94) which proved crucial in extracting robust features from the dataset. This self-supervised method further proved its adaptability and usefulness in vehicle ReID tasks by showcasing its efficacy on the VERI-Wild dataset.

9 Conclusion

In this survey paper, we looked at the current state of UAV datasets, highlighting their various applications, inherent challenges, and future directions. UAV datasets are essential in areas such as disaster management, surveillance, agriculture, environmental monitoring, and human behavior analysis. Advanced machine learning techniques have improved UAV capabilities, enabling more precise data collection and analysis. Despite their potential, UAV datasets face several challenges, including data quality, consistency, and the need for standardized annotation protocols. Ethical and privacy concerns necessitate strong frameworks to ensure responsible use. Future research should increase dataset diversity, integrate multimodal data, and improve real-time data processing. Improving annotation quality and promoting collaborative research and open data initiatives will increase dataset utility. To summarize, UAV datasets are at a critical stage of development, with significant opportunities for technological advancements. Addressing current challenges and focusing on future research directions will result in new discoveries, keeping UAV technology innovative and practical.

References

[1] Syed Agha Hassnain Mohsan, Muhammad Asghar Khan, Fazal Noor, Insaf Ullah, and Mohammed H Alsharif. Towards the unmanned aerial vehicles (uavs): A comprehensive review. *Drones*, 6(6):147, 2022.

[2] Kien Nguyen, Clinton Fookes, Sridha Sridharan, Yingli Tian, Feng Liu, Xiaoming Liu, and Arun Ross. The state of aerial surveillance: A survey. *arXiv preprint arXiv:2201.03080*, 2022.

[3] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2019.

[4] Shiyu Hu, Xin Zhao, Lianghua Huang, and Kaiqi Huang. Global instance tracking: Locating target more like humans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):576–592, 2022.

[5] Maryam Rahnemoonfar, Tashnim Chowdhury, and Robin Murphy. Rescuenet: a high resolution uav semantic segmentation dataset for natural disaster damage assessment. *Scientific data*, 10(1):913, 2023.

[6] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[7] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[9] Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021.

[10] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16266–16275, 2021.

[11] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020.

[12] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343, 2017.

[13] Veerachart Srisamosorn, Noriaki Kuwahara, Atsushi Yamashita, Taiki Ogata, Shouhei Shirafuji, and Jun Ota. Human position and head direction tracking in fisheye camera using randomized ferns and fisheye histograms of oriented gradients. *The Visual Computer*, 36(7):1443–1456, 2020.

[14] Konstantinos K Delibasis, Vassilis P Plagianakos, and Ilias Maglogiannis. Pose recognition in indoor environments using a fisheye camera and a parametric human model. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 2, pages 470–477. IEEE, 2014.

[15] Christos Kyrkou and Theocharis Theocharides. Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles. In *CVPR workshops*, pages 517–525, 2019.

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[17] 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.

[18] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[19] Ilker Bozcan and Erdal Kayacan. Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8504–8510. IEEE, 2020.

[20] Andrew Gilbert, Matthew Trumble, Charles Malleson, Adrian Hilton, and John Collomosse. Fusing visual and inertial sensors with semantics for 3d human pose estimation. *International Journal of Computer Vision*, 127:381–397, 2019.

[21] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[22] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[23] Lichao Mou, Yuansheng Hua, Pu Jin, and Xiao Xiang Zhu. Era: A data set and deep learning benchmark for event recognition in aerial videos [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 8(4):125–133, 2020.

[24] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[25] 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.

[26] Peng Wang, Bingliang Jiao, Lu Yang, Yifei Yang, Shizhou Zhang, Wei Wei, and Yanning Zhang. Vehicle re-identification in aerial imagery: Dataset and approach. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 460–469, 2019.

[27] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3235–3243, 2019.

[28] Georgios Albanis, Nikolaos Zioulis, Anastasios Dimou, Dimitrios Zarpalas, and Petros Daras. Dronepose: photorealistic uav-assistant dataset synthesis for 3d pose estimation via a smooth silhouette loss. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 663–681. Springer, 2020.

[29] Dan Oneata and Horia Cucu. Kite: Automatic speech recognition for unmanned aerial vehicles. *arXiv preprint arXiv:1907.01195*, 2019.

[30] Asanka G Perera, Yee Wei Law, and Javaan Chahl. Uav-gesture: A dataset for uav control and gesture recognition. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[31] Junjie Ye, Changhong Fu, Ziang Cao, Shan An, Guangze Zheng, and Bowen Li. Tracker meets night: A transformer enhancer for uav tracking. *IEEE Robotics and Automation Letters*, 7(2):3866–3873, 2022.

[32] Bowen Li, Changhong Fu, Fangqiang Ding, Junjie Ye, and Fuling Lin. All-day object tracking for unmanned aerial vehicle. *IEEE Transactions on Mobile Computing*, 22(8):4515–4529, 2022.

[33] Xin Zhao, Shiyu Hu, Yipei Wang, Jing Zhang, Yimin Hu, Rongshuai Liu, Haibin Ling, Yin Li, Renshu Li, Kun Liu, et al. Biodrone: A bionic drone-based single object tracking benchmark for robust vision. *International Journal of Computer Vision*, 132(5):1659–1684, 2024.

[34] Stefano Puliti, Grant Pearse, Peter Surovỳ, Luke Wallace, Markus Hollaus, Maciej Wielgosz, and Rasmus Astrup. For-instance: a uav laser scanning benchmark dataset for semantic and instance segmentation of individual trees. *arXiv preprint arXiv:2309.01279*, 2023.

[35] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015.

[36] Abhijit Kundu, Vibhav Vineet, and Vladlen Koltun. Feature space optimization for semantic video segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3168–3175, 2016.

[37] Ju Yong Chang and Kyoung Mu Lee. 2d–3d pose consistency-based conditional random fields for 3d human pose estimation. *Computer Vision and Image Understanding*, 169:52–61, 2018.

[38] Anh Nguyen and Bac Le. 3d point cloud segmentation: A survey. In *2013 6th IEEE conference on robotics, automation and mechatronics (RAM)*, pages 225–230. IEEE, 2013.

[39] Xiu-Shen Wei, Chen-Lin Zhang, Lingqiao Liu, Chunhua Shen, and Jianxin Wu. Coarse-to-fine: A rnn-based hierarchical attention model for vehicle re-identification. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part II 14*, pages 575–591. Springer, 2019.

[40] Xiaobin Liu, Shiliang Zhang, Qingming Huang, and Wen Gao. Ram: a region-aware deep model for vehicle re-identification. In *2018 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2018.

[41] Anh Nguyen and Bac Le. 3d point cloud segmentation: A survey. In *2013 6th IEEE conference on robotics, automation and mechatronics (RAM)*, pages 225–230. IEEE, 2013.

[42] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Interspeech*, pages 3214–3218, 2015.

[43] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[44] Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently scaling up crowdsourced video annotation: A set of best practices for high quality, economical video labeling. *International journal of computer vision*, 101:184–204, 2013.

[45] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3218–3226, 2015.

[46] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 445–461. Springer, 2016.

[47] Siyi Li and Dit-Yan Yeung. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[48] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2019.

[49] Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Learning target candidate association to keep track of what not to track. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13444–13454, 2021.

[50] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):583–596, 2014.

[51] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation, 2021.

[52] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[53] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2015.

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[55] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification, 2015.

[56] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[57] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. volume 9906, pages 869–884, 10 2016.

[58] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2167–2175, 2016.

[59] Yuhui Yuan, Kuiyuan Yang, and Chao Zhang. Hard-aware deeply cascaded embedding, 2017.

[60] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro, 2017.

[61] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020.

[62] Bugra Tekin, Sudipta N. Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction, 2018.

[63] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression, 2019.

[64] Dimensions ai: The most advanced scientific research database, n.d. Retrieved from `https://up42.com/blog/full-spectrum-multispectral-imagery-and-hyperspectral-imagery`.

[65] Keiron O'shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[66] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[67] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.

[68] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.

[69] Jiahui Yu, Hongwei Gao, Dalin Zhou, Jinguo Liu, Qing Gao, and Zhaojie Ju. Deep temporal model-based identity-aware hand detection for space human–robot interaction. *IEEE Transactions on Cybernetics*, 52(12):13738–13751, 2022.

[70] Zubair Saeed, Muhammad Haroon Yousaf, Rehan Ahmed, Sergio A Velastin, and Serestina Viriri. On-board small-scale object detection for unmanned aerial vehicles (uavs). *Drones*, 7(5):310, 2023.

[71] Rahee Walambe, Aboli Marathe, and Ketan Kotecha. Multiscale object detection from drone imagery using ensemble transfer learning. *Drones*, 5(3):66, 2021.

[72] Himanshu Gupta and Om Prakash Verma. Monitoring and surveillance of urban road traffic using low altitude drone images: a deep learning approach. *Multimedia Tools and Applications*, 81(14):19683–19703, 2022.

[73] Nikolaos Ioannis Bountos, Arthur Ouaknine, and David Rolnick. Fomo-bench: a multi-modal, multi-scale and multi-task forest monitoring benchmark for remote sensing foundation models, 2024.

[74] Cailian Zhang et al. Individual tree segmentation from uas lidar data based on hierarchical filtering and clustering. *International Journal of Digital Earth*, 17(1):2356124, 2024.

[75] Georgios Nikolaos Albanis, Nikolaos Zioulis, Anargyros Chatzitofis, Anastasios Dimou, Dimitrios Zarpalas, and Petros Daras. On end-to-end 6dof object pose estimation and robustness to object scale. In *ML Reproducibility Challenge 2020*, 2021.

[76] Bo Chen, Alvaro Parra, Jiewei Cao, Nan Li, and Tat-Jun Chin. End-to-end learnable geometric vision by backpropagating pnp optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8100–8109, 2020.

[77] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020.

[78] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[79] Christos Kyrkou and Theocharis Theocharides. Emergencynet: Efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:1687–1699, 2020.

[80] Fatma S Alrayes, Saud S Alotaibi, Khalid A Alissa, Mashael Maashi, Areej Alhogail, Najm Alotaibi, Heba Mohsen, and Abdelwahed Motwakel. Artificial intelligence-based secure communication and classification for drone-enabled emergency monitoring systems. *Drones*, 6(9):222, 2022.

[81] Changhong Fu, Liangliang Yao, Haobo Zuo, Guangze Zheng, and Jia Pan. Sam-da: Uav tracks anything at night with sam-powered domain adaptation, 2024.

[82] Changhong Fu, Kunhan Lu, Guangze Zheng, Junjie Ye, Ziang Cao, Bowen Li, and Geng Lu. Siamese object tracking for unmanned aerial vehicle: A review and comprehensive analysis, 2022.

[83] Usman Azmat, Saud S. Alotaibi, Naif Al Mudawi, Bayan Ibrahimm Alabduallah, Mohammed Alonazi, Ahmad Jalal, and Jeongmin Park. An elliptical modeling supported system for human action deep recognition over aerial surveillance. *IEEE Access*, 11:75671–75685, 2023.

[84] Wei Lin, Leonid Karlinsky, Nina Shvetsova, Horst Possegger, Mateusz Kozinski, Rameswar Panda, Rogerio Feris, Hilde Kuehne, and Horst Bischof. Match, expand and improve: Unsupervised finetuning for zero-shot action recognition with language knowledge, 2023.

[85] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[86] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners, 2023.

[87] Zengxi Huang, Yusong Qin, Xiaobing Lin, Tianlin Liu, Zhenhua Feng, and Yiguang Liu. Motion-driven spatial and temporal adaptive high-resolution graph convolutional networks for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(4):1868–1883, 2023.

[88] Zesheng Hu, Zihao Pan, Qiang Wang, Lei Yu, and Shumin Fei. Forward-reverse adaptive graph convolutional networks for skeleton-based action recognition. *Neurocomput.*, 492(C):624–636, jul 2022.

[89] Jiawen Zhu, Huayi Tang, Zhi-Qi Cheng, Jun-Yan He, Bin Luo, Shihao Qiu, Shengming Li, and Huchuan Lu. Dcpt: Darkness clue-prompted tracking in nighttime uavs, 2024.

[90] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6182–6191, 2019.

[91] Junjie Ye, Changhong Fu, Guangze Zheng, Ziang Cao, and Bowen Li. Darklighter: Light up the darkness for uav tracking. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3079–3085, 2021.

[92] Changhong Fu, Ziang Cao, Yiming Li, Junjie Ye, and Chen Feng. Siamese anchor proposal network for high-speed aerial tracking, 2021.

[93] Junjie Ye, Changhong Fu, Ziang Cao, Shan An, Guangze Zheng, and Bowen Li. Tracker meets night: A transformer enhancer for uav tracking. *IEEE Robotics and Automation Letters*, 7(2):3866–3873, 2022.

[94] Syeda Nyma Ferdous, Xin Li, and Siwei Lyu. Uncertainty aware multitask pyramid vision transformer for uav-based object re-identification. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2381–2385. IEEE, 2022.

[95] Shuoyi Chen, Mang Ye, and Bo Du. Rotation invariant transformer for recognizing object in uavs. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2565–2574, 2022.

[96] Usman Azmat, Saud S. Alotaibi, Maha Abdelhaq, Nawal Alsufyani, Mohammad Shorfuzzaman, Ahmad Jalal, and Jeongmin Park. Aerial insights: Deep learning-based human action recognition in drone imagery. *IEEE Access*, 11:83946–83961, 2023.

[97] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3218–3226, 2015.

[98] Christos Papaioannidis, Dimitrios Makrygiannis, Ioannis Mademlis, and Ioannis Pitas. Learning fast and robust gesture recognition. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 761–765, 2021.

[99] Fan Yang, Yang Wu, Sakriani Sakti, and Satoshi Nakamura. Make skeleton-based action recognition model smaller, faster and better. In *Proceedings of the 1st ACM International Conference on Multimedia in Asia*, pages 1–6, 2019.

[100] Libo Wang, Rui Li, Dongzhi Wang, Chenxi Duan, Teng Wang, and Xiaoliang Meng. Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images, 2022.

[101] Rui Li, Chenxi Duan Libo Wang, Ce Zhang, and Shunyi Zheng. A2-fpn for semantic segmentation of fine-resolution remotely sensed images. *International Journal of Remote Sensing*, 43(3):1131–1155, 2022.

[102] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022.

[103] Michael Ying Yang, Saumya Kumaar, Ye Lyu, and Francesco Nex. Real-time semantic segmentation with context aggregation network. *ISPRS journal of photogrammetry and remote sensing*, 178:124–134, 2021.

[104] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[105] Dechao Meng, Liang Li, Xuejing Liu, Yadong Li, Shijie Yang, Zheng-Jun Zha, Xingyu Gao, Shuhui Wang, and Qingming Huang. Parsing-based view-aware embedding network for vehicle re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7103–7112, 2020.

[106] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3235–3243, 2019.

[107] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9664–9667, 2023.

[108] Saghir Alfasly, Yongjian Hu, Haoliang Li, Tiancai Liang, Xiaofeng Jin, Beibei Liu, and Qingli Zhao. Multi-label-based similarity learning for vehicle re-identification. *IEEE Access*, 7:162605–162616, 2019.

[109] Fei Shen, Yi Xie, Jianqing Zhu, Xiaobin Zhu, and Huanqiang Zeng. Git: Graph interactive transformer for vehicle re-identification. *IEEE Transactions on Image Processing*, 32:1039–1051, 2023.

[110] Xinchen Liu, Wu Liu, Jinkai Zheng, Chenggang Yan, and Tao Mei. Beyond the parts: Learning multi-view cross-part correlation for vehicle re-identification. In *Proceedings of the 28th ACM international conference on multimedia*, pages 907–915, 2020.

[111] Fei Shen, Jianqing Zhu, Xiaobin Zhu, Yi Xie, and Jingchang Huang. Exploring spatial significance via hybrid pyramidal graph network for vehicle re-identification. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):8793–8804, 2021.

[112] Ratnesh Kuma, Edwin Weill, Farzin Aghdasi, and Parthasarathy Sriram. Vehicle re-identification: an efficient baseline using triplet embedding. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2019.

[113] Pirazh Khorramshahi, Neehar Peri, Jun-cheng Chen, and Rama Chellappa. The devil is in the details: Self-supervised attention for vehicle re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 369–386. Springer, 2020.

[114] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[115] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection, 2020.

[116] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement, 2018.

[117] Junjie Ye, Changhong Fu, Guangze Zheng, Danda Pani Paudel, and Guang Chen. Unsupervised domain adaptation for nighttime aerial tracking, 2022.

[118] Ziang Cao, Changhong Fu, Junjie Ye, Bowen Li, and Yiming Li. Siamapn++: Siamese attentional aggregation network for real-time uav tracking. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3086–3092, 2021.

[119] Fan Yang, Sakriani Sakti, Yang Wu, and Satoshi Nakamura. Make skeleton-based action recognition model smaller, faster and better, 2020.

[120] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. P-cnn: Pose-based cnn features for action recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3218–3226, 2015.

[121] Rui Li, Shunyi Zheng, Ce Zhang, Chenxi Duan, Libo Wang, and Peter M Atkinson. Abcnet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. *ISPRS journal of photogrammetry and remote sensing*, 181:84–98, 2021.

[122] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition, 2021.

A Appendix

The following images were captured from the papers in which they were presented as a new dataset or from the dataset repositories referenced in their paper where they were made available as public dataset repositories.

A.1 AIDER

Figure 3: Aerial Image Dataset for Applications in Emergency Response (AIDER): A selection of pictures from the Augmented Database

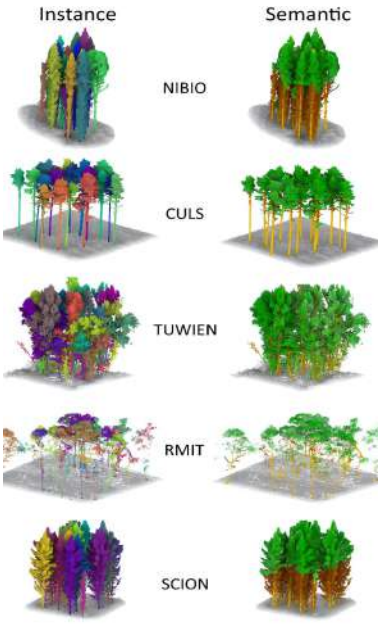
A.2 BioDrone

Figure 4: Illustrations of the flapping-wing UAV used for data collection and the representative data of BioDrone. Different flight attitudes for various scenes under three lighting conditions are included in the data acquisition process, ensuring that BioDrone can fully reflect the robust visual challenges of the flapping-wing UAVs.

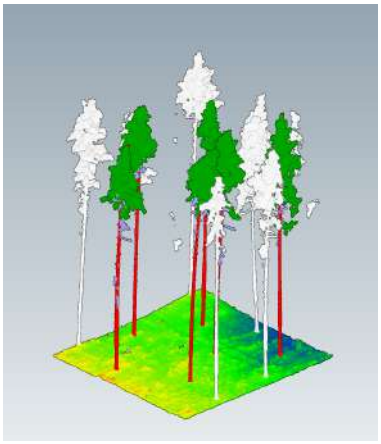
A.3 ERA

Figure 5: Overview of the ERA dataset. Overall, they have collected 2,864 labeled video snippets for 24 event classes and 1 normal class: post-earthquake, flood, fire, landslide, mudslide, traffic collision, traffic congestion, harvesting, ploughing, constructing, police chase, conflict, baseball, basketball, boating, cycling, running, soccer, swimming, car racing, party, concert, parade/protest, religious activity, and non-event. For each class, we show the first (left) and last (right) frames of a video. Best viewed zoomed in color.

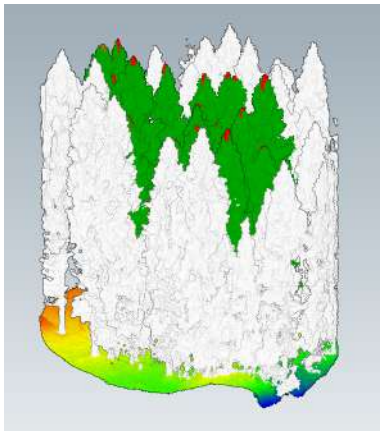
A.4 FOR-instance



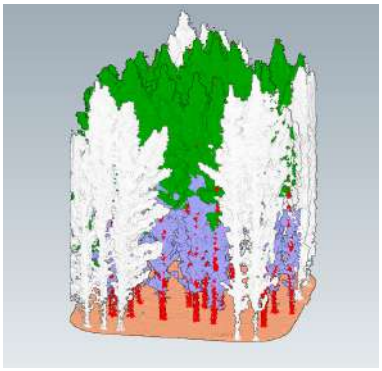
(a) Image 1



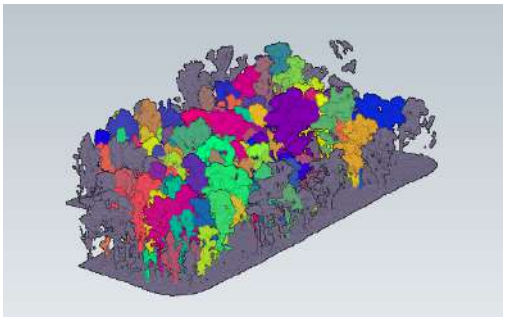
(b) Image 2



(c) Image 3



(d) Image 4



(e) Image 5

Figure 6: Samples of the various FOR-instance data collections' instance and semantic annotations.

A.5 UAVDark135

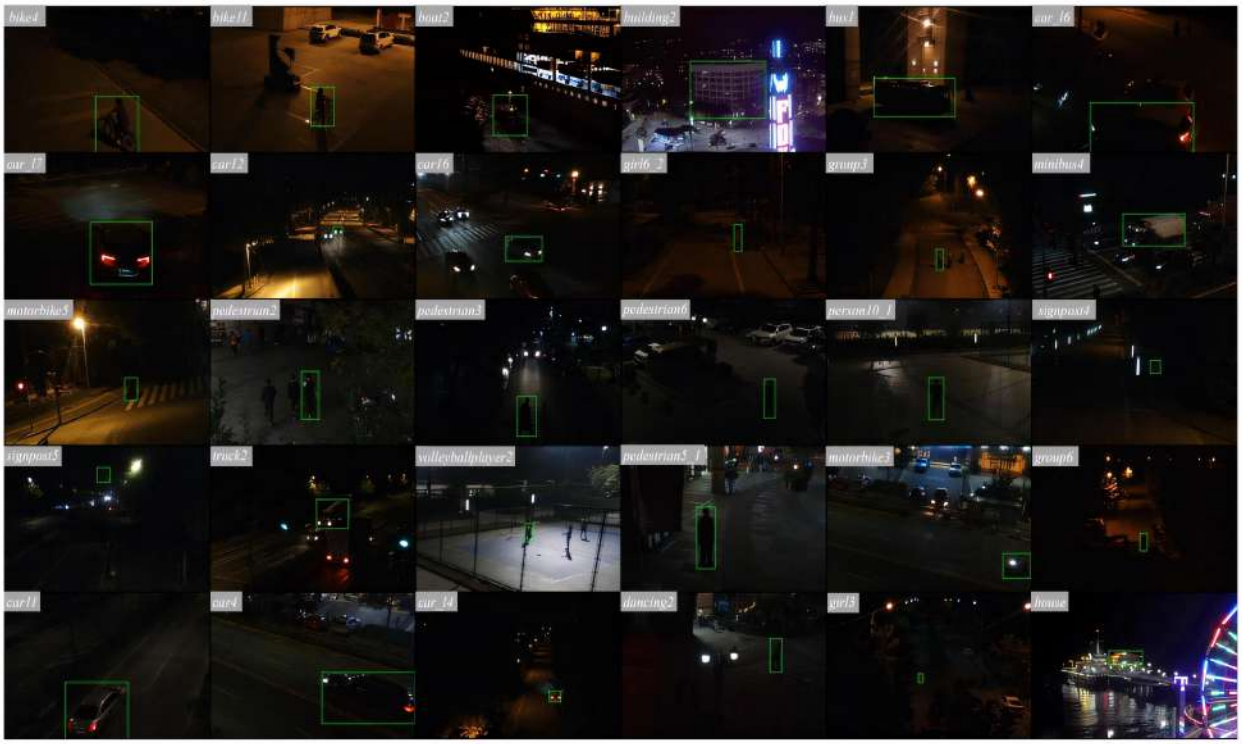


Figure 7: The first frames of representative scenes in newly constructed UAVDark135. Here, target ground-truths are marked out by green boxes and sequence names are located at the top left corner of the images. Dark special challenges like objects' unreliable color feature and objects' merging into the dark can be seen clearly.

A.6 UAV-Human

Figure 8: Examples of action videos in UAV-Human dataset. The first and second rows show two video sequences of significant camera motions and view variations, caused by continuously varying flight attitudes, speeds and heights. The last three rows display action samples of the dataset, showing the diversities, e.g., distinct views, various capture sites, weathers, scales, and motion blur.

A.7 UAVid

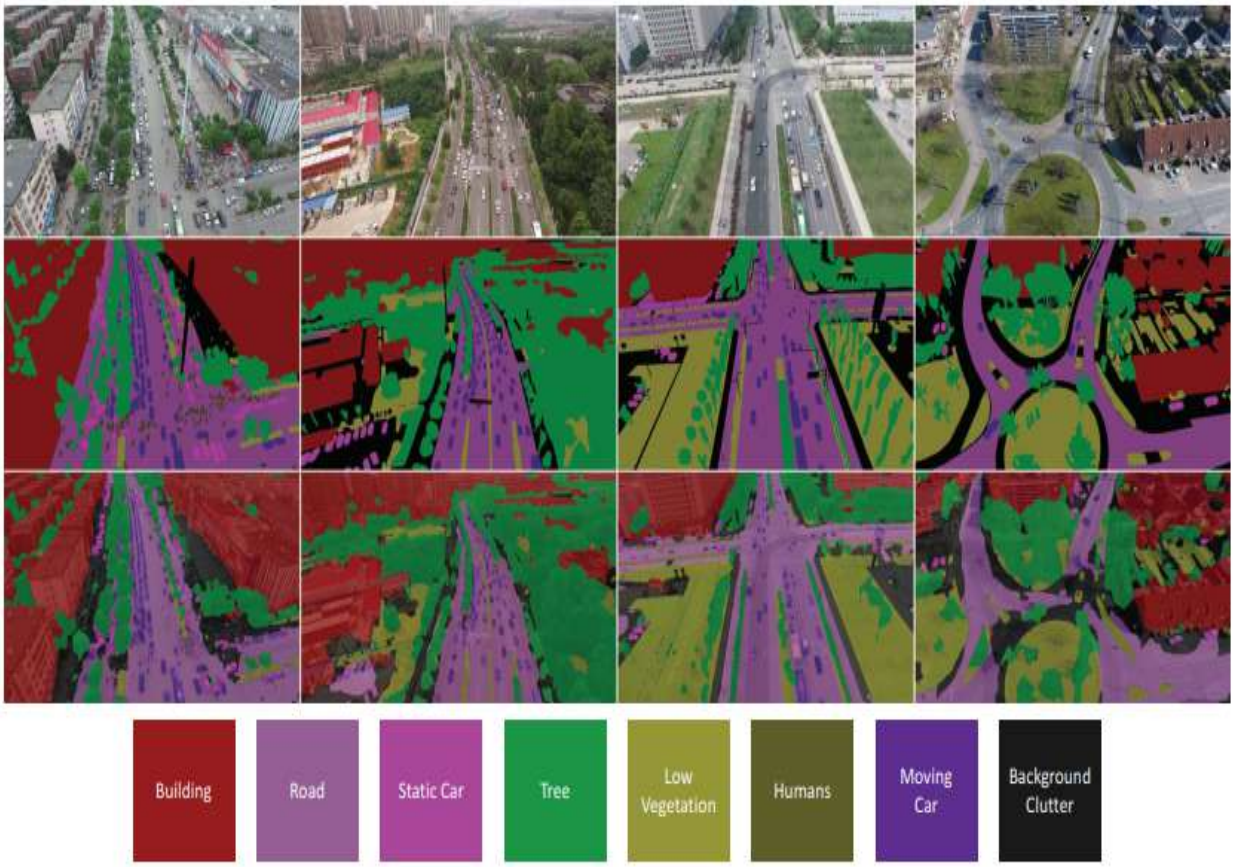


Figure 9: Example images and labels from UAVid dataset. First row shows the images captured by UAV. Second row shows the corresponding ground truth labels. Third row shows the prediction results of MS-Dilation net+PRT+FSO model. The last row shows the labels.

A.8 DarkTrack2021

Figure 10: Initial frames of specific sequences from the DarkTrack2021 archive. Objects being tracked are indicated by green boxes, and sequence names are shown in the top left corner of the photos.

A.9 VRAI

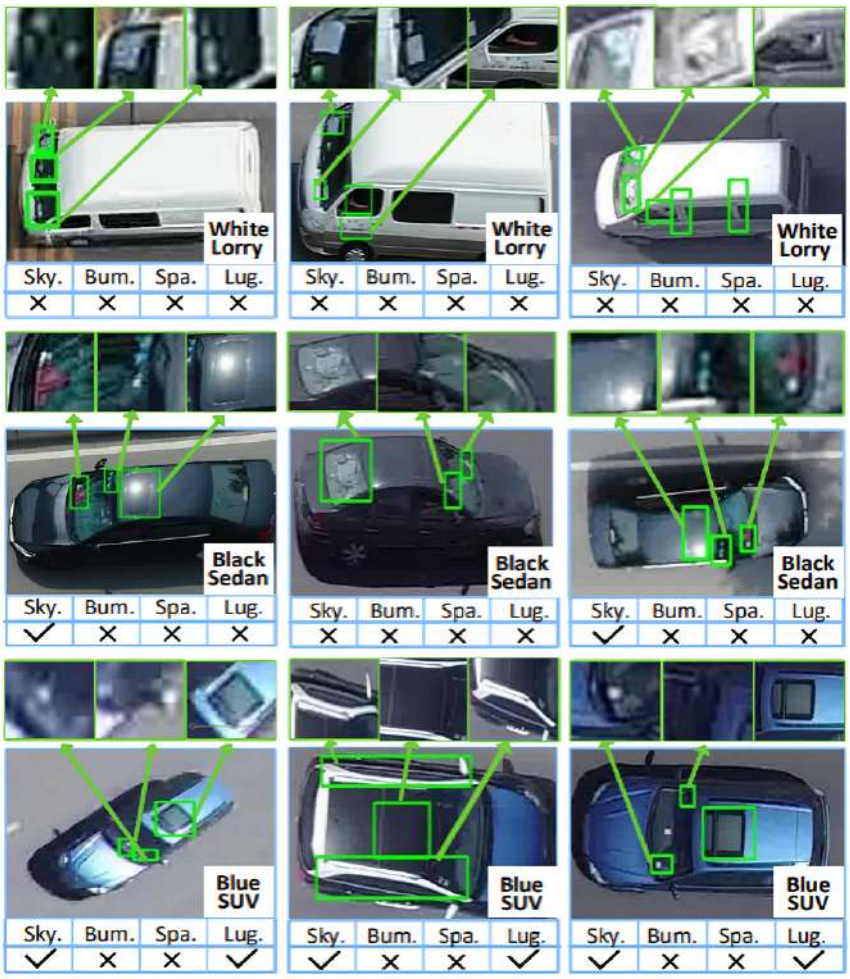


Figure 11: Overview of our gathered dataset for Unmanned Aerial Vehicle (UAV)-based vehicle ReID. In order to facilitate thorough investigation, the authors have included a comprehensive range of information in the dataset, such as color, vehicle type, Skylight (Sky.), Bumper (Bum.), Spare tire (Spa.), Luggage rack (Lug.), and distinguishing components.

A.10 VERI-Wild

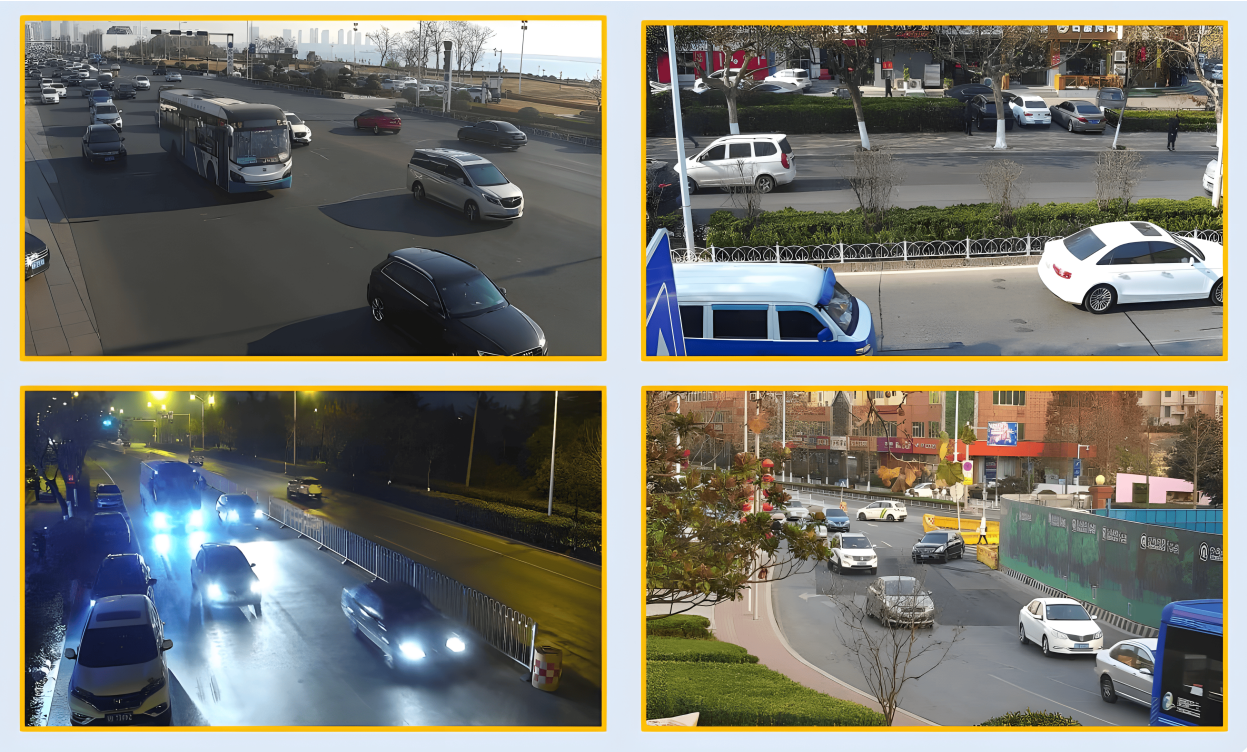


Figure 12: Exemplary photos extracted from the dataset. The dataset is obtained from a comprehensive real video surveillance system including 174 cameras strategically placed around an urban area spanning over 200 square kilometers.

A.11 RescueNet

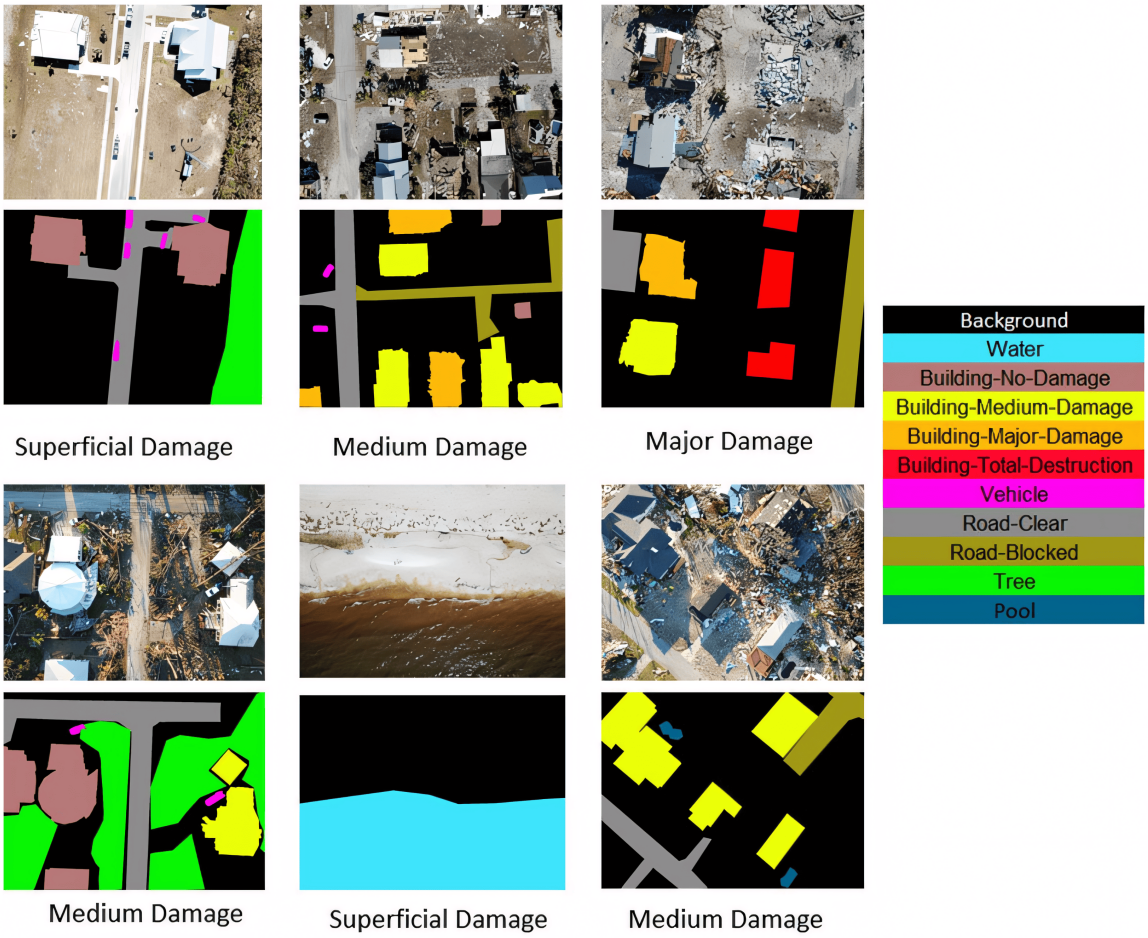


Figure 13: Graphical representation of complex scenes from the RescueNet dataset. The first and third rows display the original photos, while the lower rows provide the associated annotations for both semantic segmentation and image classification functions. Displayed on the right are the 10 classes, each represented by their segmentation color.

A.12 UAV-Assistant

Figure 14: This diagram illustrates the many modalities present in the UAV-Assistant dataset, which consists of a randomly chosen collection of images. The uppermost row displays color photos, the second row displays depth, the third row displays the normal map, and the last row displays flight silhouettes of the drone.

A.13 AU-AIR

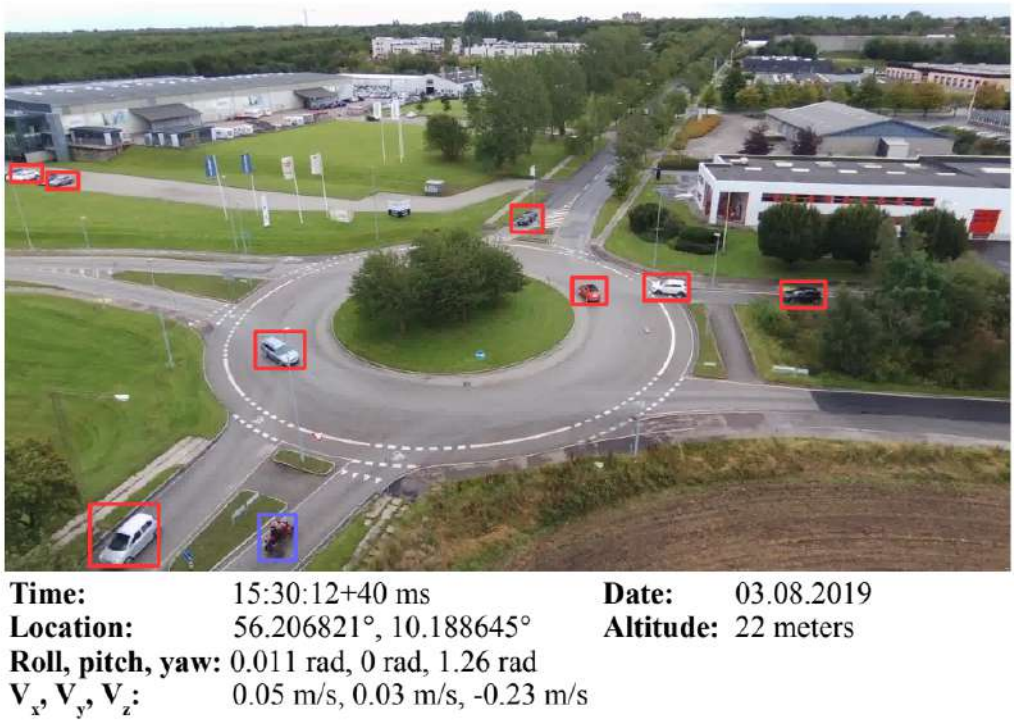


Figure 15: The AU-AIR dataset includes extracted frames that are annotated with object information, time stamp, current location, altitude, velocity of the UAV, and rotation data observed from the IMU sensor. This figure presents an exemplar of it.

A.14 UAV-Gesture

Figure 16: This diagram displays thirteen explicitly chosen gestures, each accompanied by a single picked image. Directions of hand movement are shown by the arrows. The amber color marks serve as approximate indicators of the initial and final locations of the palm for ONE iteration. Neither the Hover nor Land gestures are dynamic gestures.

A.15 Kite

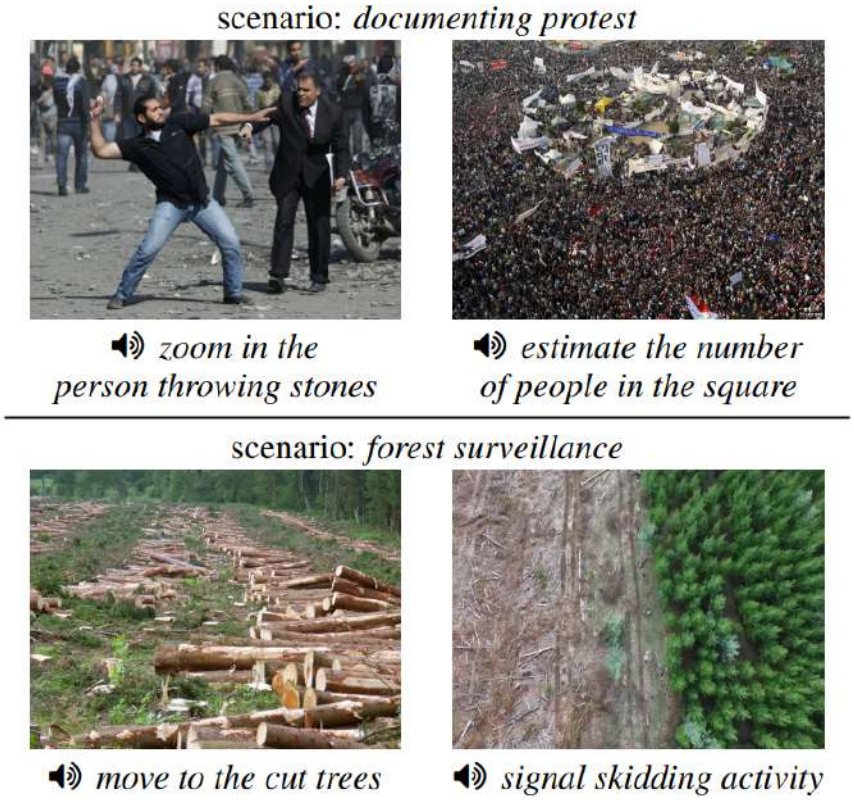


Figure 17: Exemplary commands and visual representations derived from the KITE dataset.