Click&Describe: Multimodal Grounding and Tracking for Aerial Objects - Supplementary Material

Rupanjali Kukal Jay Patravali Fuxun Yu Simranjit Singh Nikolaos Karianakis Rishi Madhok
Microsoft
{rkukal, jaypatravali, fuxunyu, simsingh, nikolaos.karianakis, rishi.madhok}@microsoft.com

Please refer to the accompanying webpage (site.html) and videos (in folders 'results' and 'videos') for visual results. It should be noted that, due to size limitations, certain videos have been compressed and/or cropped to a lower quality. **Please note, dataset URL and webpage will be made public on acceptance.** We declare that we bear all responsibility in case of violation of rights, etc., and confirmation of the data license.

1. Additional Ablation

Instead of providing precise bounding box annotations at inference time, using a click input enables the user to provide approximate target information in a more natural and user-friendly way. This significantly reduces the time required to provide input at inference, by approximately 43% and eventually, human effort—especially in real-world applications where drone pilots often rely on small interface screens. We compare the time it takes for 5 users to annotate target objects in sample aerial images via click and bounding box. This experiment is conducted on screens of 3 different sizes: a monitor, laptop screen, and a phone screen (to mimic smaller screen sizes accessible to pilots). We select 100 aerial images covering different resolutions for this experiment and the results are shown in Tab. 1.

Description of Target	One Click	Two Clicks	% Difference
24-inch Monitor			
Small	1.45	2.72	47%
Medium	1.14	1.93	41%
Large	0.88	1.53	42%
Dark	1.61	2.67	40%
13-inch Laptop			
Small	1.29	2.32	44%
Medium	0.84	1.45	42%
Large	0.95	1.33	29%
Dark	2.31	4.11	44%
iPhone 14 Pro			
Small	2.51	4.98	50%
Medium	2.11	3.89	46%
Large	1.45	2.58	44%
Dark	1.78	3.87	54%

Table 1. We find that annotating precise bounding boxes takes 43% more time as compared to annotating with just an approximate click. The absolute time difference will scale up as the number of sequences to be annotated increases.

2. Additional Dataset Samples

We show additional zoomed-in samples from the different aerial datasets with their natural language description and attributes in Fig. 1.

3. Additional Visual Results

We show additional results of CLaVi on sequences addressing the challenges specific to aerial datasets in Fig. 2. For the first sequence (row 1), the target object is placed in a dense urban environment with similar-looking vehicles and is initially small in size (column 1). As the video progresses, we observe some partial occlusion caused by the pole (column 3). Likewise, in the second sequence (row 2), the target object is small and surrounded by similar-looking objects. The camera position of the aerial vehicle shifts from an angle to the overhead of the vehicle. In the fifth sequence (row 5), a basketball player wearing red clothes is being tracked. Due to occlusions caused by the rapid movement of the target, our prediction drifts to another player wearing black clothes (column 2). However, our model can track the original player, showing the effectiveness of the memory modules in preventing motion drift. We also observe motion blur in the third sequence (row 3). We highlight in the last sequence in the figure (row 6), an example of nighttime target tracking with occlusion via lack of illumination or by other obstacles. All the above examples show that our method is robust to changes in target size, dense scenarios with similar-looking targets, motion blur, and other obstacles inherent to aerial data.

4. Attributes Definitions

In Tab. 2, we provide a detailed definition of the attributes in AerTrack-460 benchmark.

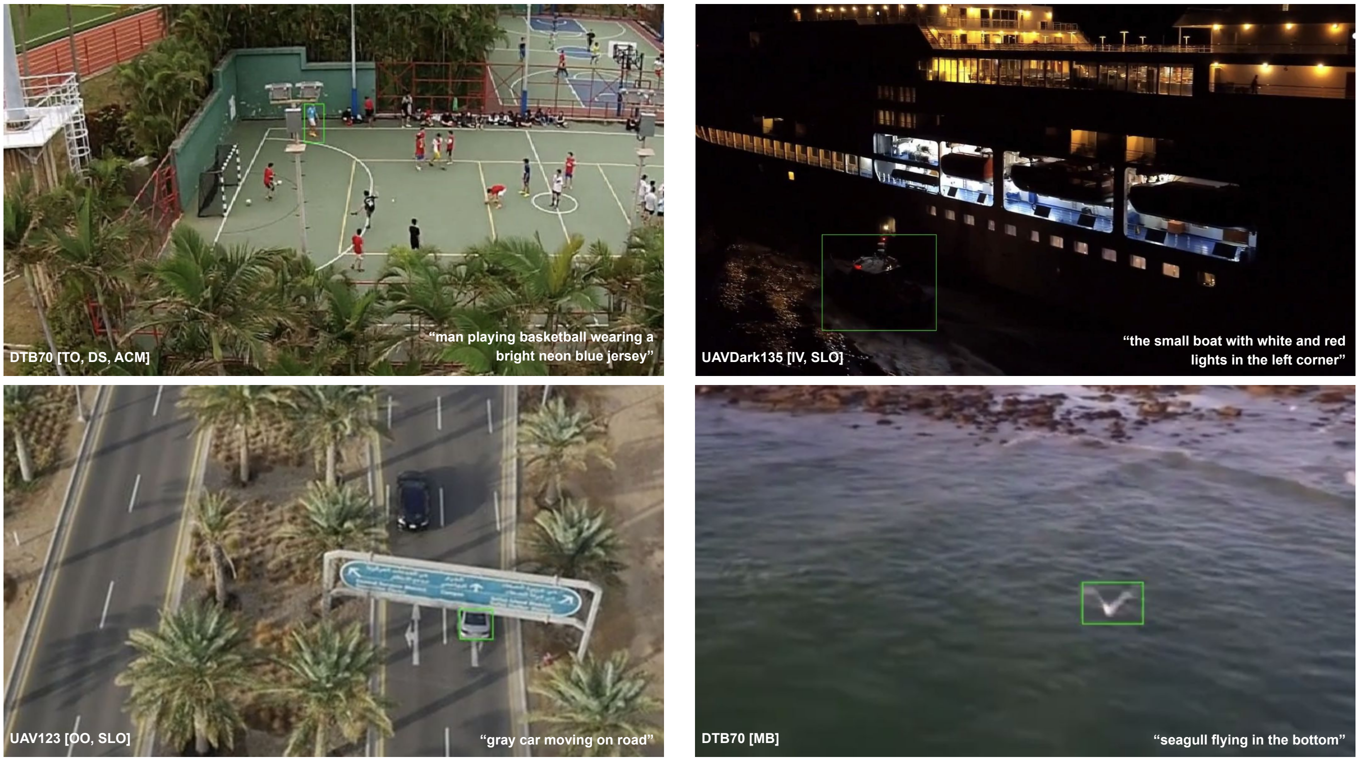

DTB70 [TO, DS, ACM]
“man playing basketball wearing a bright neon blue jersey”
UAVDark135 [IV, SLO]
“the small boat with white and red lights in the left corner”
UAV123 [OO, SLO]
“gray car moving on road”
DTB70 [MB]
“seagull flying in the bottom”

Figure 1. **AerTrack-460.** Zoomed-in snippets with parent dataset [attributes] and natural language annotation for UAVDT [1], UAV123 [5], UAVDark135 [3] and DTB70 [4]

Attribute	Definition
OO	The target is partially or fully occluded
ACM	Abrupt motion of the camera
MB	Region around target is blurred due to target or camera motion
SV	Variation in scale
VC	The change in viewpoint causing significant changes in the appearance of the target (eg: 360 degree rotation of the target object)
TO	The target is small and unidentifiable in atleast over 10 frames
SLO	Similar semantics of the target with other background objects
DS	The target is densely packed around background clutter
IV	Illumination around target region changes

Table 2. **AerTrack-460 attributes definition**

5. Datasheet for AerTrack-460

Following below, we provide a datasheet [2] describing the collection of data which form AerTrack-460.

5.1. Motivation

1. **For what purpose what was the dataset created?** (Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.)

   This dataset was created to provide Natural Language annotations for aerial datasets.

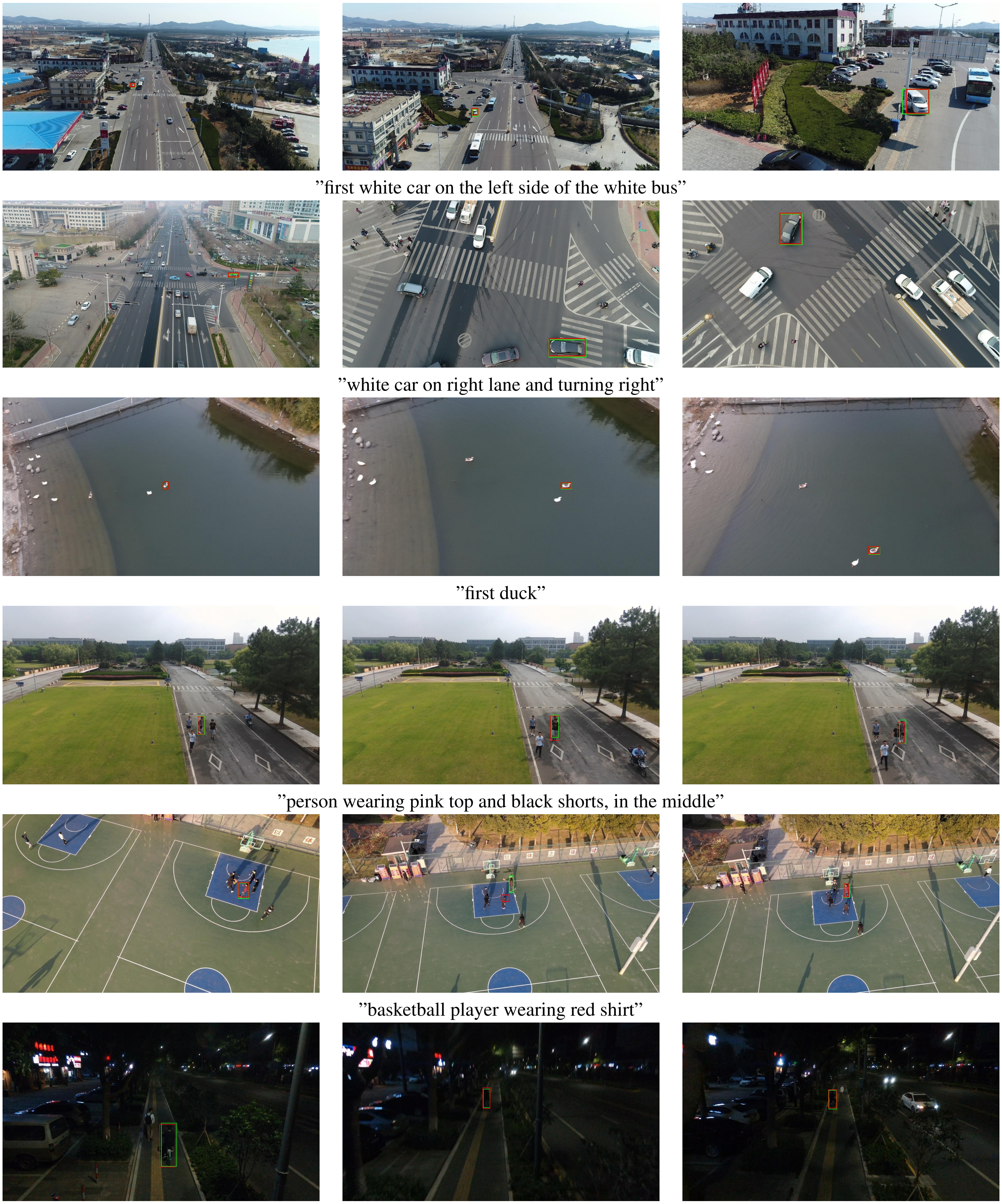
2. **Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

   Members from the Mixed Reality team at Microsoft created the dataset.

3. **Who funded the creation of the dataset?** (If there is an associated grant, please provide the name of the grantor and the grant name and number.)

   The work behind dataset creation was funded by Microsoft.

4. **Any other comments?**



”first white car on the left side of the white bus”

”white car on right lane and turning right”

”first duck”

”person wearing pink top and black shorts, in the middle”

”basketball player wearing red shirt”

”the man riding a bike on the right side of the lane”

Figure 2. **Qualitative Results.** We compare bounding box predictions from JointNLT (NL) (green) and CLaVi (red). Predictions best visible when zoomed in. Each row represents a single test scene (progressing temporally) from the AerTrack-460 test set

None.

5.2. Composition

1. **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** (Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.)

   Each instance comprises of a video sequence along with a textual annotation which is an English sentence describing the target object by its color, action, and surroundings in the first frame.. Further, we also provide the attribute of each video sequence.

2. **How many instances are there in total (of each type, if appropriate)?**

   The dataset consists of 461 video instances comprising of language annotation and the video attributes.

3. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** (If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).]

   The dataset contains all possible instances.

4. **What data does each instance consist of?** (”Raw” data (e.g., unprocessed text or images)or features? In either case, please provide a description.)

   Each instance consists of a video sequence, video attributes, bounding box annotation for every frame, and natural language annotation for the entire sequence.

5. **Is there a label or target associated with each instance? If so, please provide a description.**

   In every sequence, the target is the ground truth bounding box of the object being tracked in the video sequence.

6. **Is any information missing from individual instances?** (If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.)

   No.

7. **Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)?** (If so, please describe how these relationships are made explicit.)

   N/A

8. **Are there recommended data splits (e.g., training, development/validation, testing)?** (If so, please provide a description of these splits, explaining the rationale behind them.)

   The data will be divided into training and testing split in the download link.

9. **Are there any errors, sources of noise, or redundancies in the dataset? (If so, please provide a description.)**

   There might be some redundancies in language annotations. We did our best to minimize these, but some certainly remain.

10. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** (If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.)

    The dataset combines data from 5 data sources, links to whom will be mentioned on the github page.

11. **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals’ non-public communications)?** (If so, please provide a description.)

    To the best of our knowledge, no.

12. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** (If so, please describe why.)

    No.

13. **Does the dataset relate to people?** (If not, you may skip the remaining questions in this section.)

    The dataset has instances where some target objects are people.

14. **Does the dataset identify any subpopulations (e.g., by age, gender)?** (If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.)

    This is not explicitly identified in any of the datasets that are in AerTrack-460.

15. **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** (If so, please describe how.)

    Yes; in some examples the face of the person is visible.

16. **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** (If so, please provide a description.)

    No.

17. **Any other comments?**

    No.

5.3. Collection Process

1. **How was the data associated with each instance acquired?** (Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.)

   The data collection process involves 3 parts.

   1. Aerial sequence collection: We curate a list of diverse publicly available aerial datasets, which we download from their respective sources.

   2. Sequence attribute annotation: Given the video sequence, we assign the various aerial attributes, which are defined in detail in 4

   3. Language annotation: Finally we annotate the target in first frame of video sequence using natural language. To guarantee high-quality annotation, each video is processed in two parts: annotation and validation. We first annotate the target using natural language. Then, the annotation results are validated by other members of the team. If an annotation result is not unanimously agreed by all the authors, the original label is revised.

2. **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** (How were these mechanisms or procedures validated?)

   Annotations were done using a simple python script which will be made available soon.

3. **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

   See answer to question #2 in Composition.

4. **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

   All collection and annotation was done by the authors.

5. **Over what timeframe was the data collected?** (Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.)

   The dataset was collected in October 2023.

6. **Were any ethical review processes conducted (e.g., by an institutional review board)?** (If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.)

   No ethical review process was conducted with respect to the collection of the data. The annotations went through a manual ethical review by the team.

7. **Does the dataset relate to people?** (If not, you may skip the remaining questions in this section.)

   Yes; there are some instances where the target object is person.

8. **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

   We obtained the data via a third party, i.e. downloaded publicly available datasets.

9. **Were the individuals in question notified about the data collection?** (If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.)

The data is curated from publicly available datasets. Please refer to original dataset for more information.

10. **Did the individuals in question consent to the collection and use of their data?** (If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.)

    N/A

11. **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** (If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).)

    N/A

12. **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted?** (If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.)

    No.

13. **Any other comments?**

    No.

5.4. Preprocessing/Cleaning/ Labeling

1. **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** (If so, please provide a description. If not, you may skip the remainder of the questions in this section.)

   Yes; we provide language annotations and attributes per sequence.

2. **Was the ”raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** (If so, please provide a link or other access point to the ”raw” data.)

   Yes. The original dataset without the language annotations can be found in the dataset.

3. **Is the software used to preprocess/clean/label the instances available?** (If so, please provide a link or other access point.)

   The script to label data is straight-forward and easily reproducible. We will release the scripts with the dataset.

4. **Any other comments?**

   None.

5.5. Uses

1. **Has the dataset been used for any tasks already?** (If so, please provide a description.)

   The dataset has been used for baselining of CLaVi: language-based joint grounding and tracking.

2. **Is there a repository that links to any or all papers or systems that use the dataset?** (If so, please provide a link or other access point.)

   No.

3. **What (other) tasks could the dataset be used for?**

   The dataset can be used for multi-modal computer vision and/or language tasks in the aerial domain.

4. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** (For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?)

   None, to the best of our knowledge.

5. **Are there tasks for which the dataset should not be used?** (If so, please provide a description.)

   None, to the best of our knowledge.

6. **Any other comments?**

   None.

5.6. Distribution

1. **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** (If so, please provide a description.)

   The dataset will be made public soon.

2. **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? (Does the dataset have a digital object identifier (DOI)?)**

   The dataset will be made available on GitHub.

3. **When will the dataset be distributed?**

   The dataset will be distributed soon.

4. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** (If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.)

   The dataset is licensed under a CC by 4.0 license.

5. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** (If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.)

   Not to our knowledge.

6. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** (If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.)

   None to our knowledge.

7. **Any other comments?**

   No.

5.7. Maintenance

1. **Who is supporting/hosting/maintaining the dataset?**

   Author Rupanjali Kukal is maintaining and hosting on Github.

2. **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

   Please reach out to any of the authors.

3. **Is there an erratum?** (If so, please provide a link or other access point.)

   Currently, no. As errors are encountered, future versions of the dataset may be released (but will be versioned). They will all be provided in the same github location.

4. **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances')?** (If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?)

   Refer to above answers.

5. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** (If so, please describe these limits and explain how they will be enforced.)

   No.

6. **Will older versions of the dataset continue to be supported/hosted/maintained?** (If so, please describe how. If not, please describe how its obsolescence will be communicated to users.)

   Yes; all data will be versioned.

7. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? (If so, please provide a description. Will these contributions be validated/verified?** If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.)

   Suggestions can be submitted on github or via email and more extensive augmentations may be accepted at the authors' discretion.

8. **Any other comments?**

   None.

References

[1] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. *CoRR*, abs/1804.00518, 2018. 2

[2] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *CoRR*, abs/1803.09010, 2018. 2

[3] Bowen Li, Changhong Fu, Fangqiang Ding, Junjie Ye, and Fuling Lin. All-day object tracking for unmanned aerial vehicle. *CoRR*, abs/2101.08446, 2021. 2

[4] Siyi Li and Dit-Yan Yeung. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017. 2

[5] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 445–461, Cham, 2016. Springer International Publishing. 2