NOMAD: A Natural, Occluded, Multi-scale Aerial Dataset, for Emergency Response Scenarios

Arturo Miguel Russell Bernal, Walter Scheirer, Jane Cleland-Huang
Computer Science and Engineering Department, University of Notre Dame, Indiana, USA
arussel8@nd.edu, walter.scheirer@nd.edu, janehuang@nd.edu

Abstract

With the increasing reliance on small Unmanned Aerial Systems (sUAS) for Emergency Response Scenarios, such as Search and Rescue, the integration of computer vision capabilities has become a key factor in mission success. Nevertheless, computer vision performance for detecting humans severely degrades when shifting from ground to aerial views. Several aerial datasets have been created to mitigate this problem, however, none of them has specifically addressed the issue of occlusion, a critical component in Emergency Response Scenarios. Natural, Occluded, Multi-scale Aerial Dataset (NOMAD) presents a benchmark for human detection under occluded aerial views, with five different aerial distances and rich imagery variance. NOMAD is composed of 100 different Actors, all performing sequences of walking, laying and hiding. It includes 42,825 frames, extracted from 5.4k resolution videos, and manually annotated with a bounding box and a label describing 10 different visibility levels, categorized according to the percentage of the human body visible inside the bounding box. This allows computer vision models to be evaluated on their detection performance across different ranges of occlusion. NOMAD is designed to improve the effectiveness of aerial search and rescue and to enhance collaboration between sUAS and humans, by providing a new benchmark dataset for human detection under occluded aerial views.

1. Introduction

Advances in technology, including improvements in edge computing and Artificial Intelligence (AI), have led to increased use of small Unmanned Aerial Systems (sUAS) across a broad range of applications [53, 54, 71], such as emergency response [6,26,54,71]. sUAS are empowered to perform Computer Vision (CV) tasks, such as aerial surveillance and autonomous person detection and tracking, where timely and efficient performance potentially can make the difference between life or death [21, 27, 32]. Higher levels of sUAS autonomy, supported by CV, increase collaboration between humans and sUAS, allowing emergency responders to focus attention on mission level goals [1, 20] while sUAS perform lower-level person detection tasks.

However, there are many open challenges in deploying CV on sUAS for emergency response [19, 52]. These challenges include the non-trivial, highly prevalent problem of occlusion, which occurs when targets of aerial search are partially hidden from view. For example, a drowning victim who is partially submerged in water, people buried in debris following an earthquake, hidden by smoke in a fire, or laying behind trees and rocks in search and rescue missions. Occlusion could also be intentional when a suspect is hiding from law-enforcement, caused by pose and image perspective, or introduced in far-distance aerial views due to glare, shades, blur, and low resolution. Prior CV research on occlusion has focused on generic object detection [48, 67], as well as on pedestrian detection [60], demonstrating how occlusion drastically affects model performance [60, 82, 85]. However, the occlusion problem is exacerbated even further when shifting from ground to aerial views [59], where additional challenges surrounding the incorporation of CV capable sUAS into emergency response scenarios include biased training datasets, coupled with real-life challenges such as vibration, wind and atmospheric turbulence, harsh weather and low visibility conditions, diverse scenery, and the need for generalization at different distances and resolutions. CV systems deployed for emergency response must be able to reliably handle person detection under all of these variable conditions.

We therefore address these challenges through presenting NOMAD (Natural Occluded Multi-scale Aerial Dataset), a benchmark dataset aimed at human detection under occluded aerial views, as summarized in Fig. 1. NOMAD is composed of 100 different actors, each performing sequences of walking, laying and hiding. It includes 42,825 frames, extracted from 5.4k resolution videos. Actors are manually annotated with a bounding box and a label describing 10 different visibility levels, categorized according to the percentage of the human body visible inside the bounding box, allowing the detection performance of CV models to be evaluated across 10 different ranges of

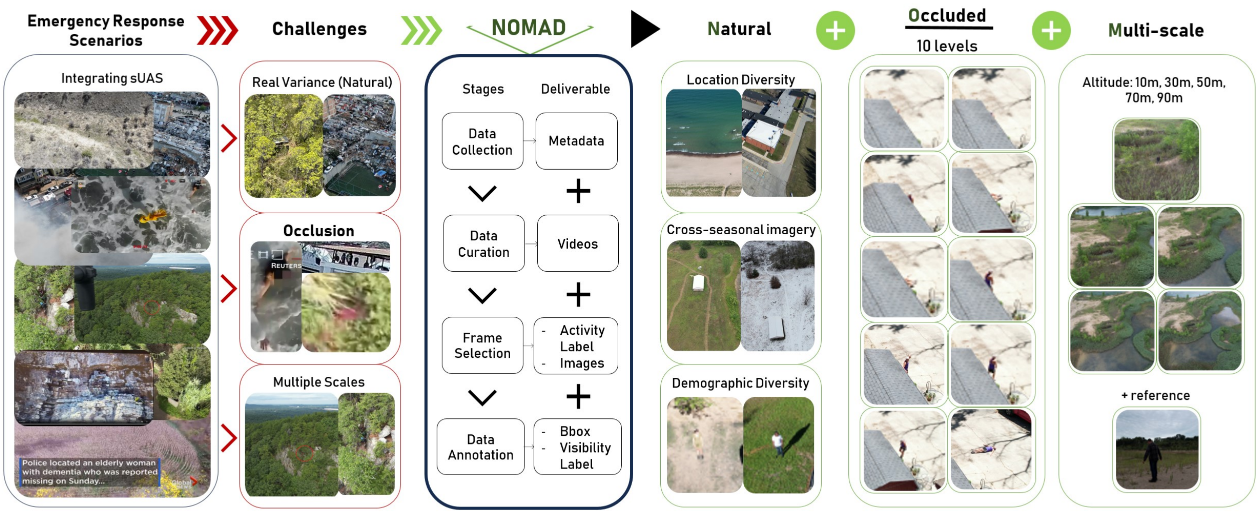


Figure 1. Development and characteristics of NOMAD. Integration of sUAS into emergency response scenarios have aided first responders and rescued victims [12, 15, 21, 27, 31, 32, 36, 42, 46] (first column). Nevertheless, multiple challenges inherent to these situations degrade CV performance and halt sUAS full integration, including the highly prevalent problem of occlusion (second column). We present NOMAD, Natural Occluded Multi-scale Aerial Dataset, providing the research community with emergency response related videos and selected frames, as well as rich metadata and annotations, including a visibility label (third column). Facing emergency response scenarios, key characteristics of our dataset are: *Natural*: diversity of filming locations, cross-seasonal imagery, including winter scenarios, and a demographic diversity on gender, age and race, ranging from 18 to 78 years old, and including White Caucasians, Latinos, African descent, Asians, South Asians, Middle Eastern and Pacific Islander; *Occluded*: 10 defined ranges of occlusion, with a visibility label assigned to every bounding box; *Multi-scale*: five different distances, ranging from 10m to 90m altitude, and a ground reference view for every actor.

occlusion. Figure 1 summarizes the key characteristics of our dataset, including: *Natural*: representing a variety of natural and man-made locations; cross-seasonal imagery, ranging from summer to winter scenarios; demographic variety on age and race, ranging from 18 to 78 years old, and including White Caucasians, Latinos, African descent, Asians, South Asians, Middle Eastern and Pacific Islander; *Occluded*: with routines created to include occlusion and a visibility label assigned to every bounding box annotated; *Multi-scale*: with five different distances, ranging from 10m to 90m altitude, and a ground reference view for every actor.

The remainder of this article is organized as follows. Section 2 presents related work. Section 3 describes the data collection process. Section 4 describes the data curation, key-frame selection and data annotation. Section 5 discusses NOMAD characteristics and its potential uses. Section 6 reports baseline results achieved using state-of-the-art CV detection models under different levels of occlusion, and Sec. 7 summarizes the contributions of the work.

2. Related Work

2.1. Mobile Robotics for Emergency Response

There are numerous challenges associated with integrating mobile robotics into emergency response missions [19, 52, 54, 58, 71, 72]. Researchers, focusing on ground mobile robots, have explored mapping of emergency scenes [61, 65, 75], improved communication networks [50], and specialized architectures [33, 41]. User studies have demonstrated the benefits of including aerial robots in emergency response [76], potentially working in collaboration with ground robots [17], to enhance surveying and mapping capabilities [64]. Other studies have explored the integration of additional sensors, such as ground penetrating radar [63], or cellphone tracking for missing person search [3]. Finally, several researchers have explored efficient collaborations between humans and sUAS at the intersection of software engineering and human computer interaction [1, 2, 14, 20].

2.2. Real-World Object Detection

There are numerous challenges related to utilizing aerial CV for real-time emergency response [48]. Real-time CV applications tend to leverage the latest versions of the YOLO family [39, 74], as well as their modifications [35, 45, 56, 57], while other methods explore attention for object detection [55] and multimodal techniques [4]. The most recent work has focused on incremental learning of unknown classes, in the modality known as Open World Object Detection [40, 51, 83, 87], as well as its variations [80, 84]. The challenges of object detection under occlusion have also been studied [16, 60, 67]. Finally, techniques incorporating human perception have been explored for object detection [62] and other machine learning tasks [8, 9, 24, 30, 34, 37, 69], demonstrating a plausible approach to handling occlusion [18, 29, 66, 85].

2.3. Aerial Datasets

While many datasets have been collected to aid aerial detection of humans in search and rescue (SAR), none of them have addressed the critical issue of occlusion. HERIDAL [10] comprises of approximately 500 labelled 4,000 by 3,000 pixel images suitable for object detection tasks. SARD [68] comprises 1,981 manually labeled images extracted from video frames of persons simulating search and rescue situations in roads, quarries, grass land, and forested areas, under diverse weather conditions. However, both datasets lack rich generalization characteristics and environmental diversity. The recently published WiSARD [11] dataset, comprises the richest set of images associated with wilderness SAR scenarios, with 33,786 labeled RGB images, 22,156 labeled thermal images, and a subset consisting of 15,453 temporally synchronized visual-thermal image pairs. In addition to the useful multimodal imagery, the dataset includes environmental diversity across seasons and times of the day and night. WiSARD represents the richest dataset for *blind search* in wilderness scenarios, that is, search for any person on an area rather than the search for an specific person; NOMAD provides richer demographic diversity, includes man-made scenarios, provides rich metadata of actors, controlled multi-scales, and provides a new benchmark for occlusion. It is the only dataset, to our knowledge, to systematically address the issue of occlusion.

The BIRDSAI, VisDrone and UAVDT datasets [7,28,86] incorporate occlusion labels into their annotations; however, they lack rich human metadata. BIRDSAI is a longwave thermal infrared dataset containing nighttime images of animals and humans in Southern Africa. While suitable for improving *blind search* of persons in emergency scenarios, it only provides two levels of occlusion and lacks person metadata. VisDrone consists of 288 video clips formed by 261,908 frames and 10,209 static images, and is captured by various drone-mounted cameras, covering diverse locations, environments, objects (pedestrian, vehicles, bicycles, etc.), and density. However, it provides only three levels of occlusion and also lacks person metadata. Finally, while UAVDT provides four levels of occlusion, it focuses purely on vehicles and not people.

BRIAR, MEVID, UAV-Human, P-DESTRE and PRAI-1581 [22, 25, 43, 47, 81] provide rich metadata and are well suited for person re-identification. BRIAR and MEVID datasets offer great diversity of camera views, with BRIAR providing long range imagery of up to 1000m. BRIAR, so far, includes more than 350,000 still images and over 1,300 hours of video footage of approximately 1,000 subjects; MEVID, is part of the very-large-scale MEVA person activities dataset [23] and comprises 158 unique people wearing 598 outfits collected from 33 camera views. UAV-HUMAN includes 67,428 annotated video sequences of 119 subjects for action recognition, 22,476 annotated frames for pose estimation, 41,290 annotated frames of 1,144 identities for person re-identification, and 22,263 annotated frames for attribute recognition. While these three datasets represent the most complete datasets for their given purposes, none of them reference occlusion and all lack representation of emergency response scenarios. Finally, PRAI-1581 provides 1,581 identities and P-DESTRE provides rich metadata for 269 different identities, however, filming distances are only up to 60m and 6.7m, respectively. Additional categorized aerial datasets can be found in [59].

Overall, NOMAD provides the demographic and environmental diversity needed to tackle the person detection task of emergency response scenarios from aerial views, while being the first dataset to include an occlusion metric for person detection, and to provide detailed metadata and controlled multi-scale, making it suitable for many other CV tasks as described in Sec. 5.

3. Data Collection Process

Our data collection process followed our IRB approved protocol 21-11-6913. In a preliminary pilot study, our data collection procedure included strict instructions regarding the percentage of the body that the actor should expose to the sUAS' camera at each step. However, we observed that these instructions were difficult to follow causing disconnected movements, and so we replaced the instructions with simpler ones that led to more natural behavior.

3.1. Recruitment

As per our IRB protocol, all participants were at least 18 years old. Further, as the recruitment process evolved, participants from already well represented demographic groups were excluded in order to achieve a balanced gender distribution, a variety of age ranges, and a rich race distribution.

3.2. Location Selection

Approval for use of premises was attained from owners and responsible agencies for all locations filmed in the dataset. The 12 locations included: 3 different Schools, 2 paintball courts, 1 forest park, 1 golf course, 1 lake shore, 1 quarry, 2 farms, and 1 AMA flying field. This resulted in diversity of locations, including both natural and man-made influenced, and provided a variety of different types of obstacles for occlusion purposes.

3.3. Filming Sessions

All filming sessions followed IRB protocol guidelines with participants being informed of the purpose of their performance, the activities to be completed, and consent forms being signed. All flights were conducted by a certified FAA Part 107 remote pilot, and all FAA protocols were followed, with air space reserved through LAANC systems such as

AirMap and DroneUp. Although efforts were made to isolate the selected locations during the filming sessions, unexpected persons appeared during a few of the filming sessions. In most cases we stalled the filming until the person exited the scene; however, in a few cases, these persons agreed to appear on the dataset, signing a consent form. From here on, we call actors the participants performing the designated routine, while non-actors are other participants who agreed to appear in the dataset but were otherwise not engaged in the study.

Once the study introduction was completed, each actor was assigned a unique *obstacle* at the filming location, and then given instructions for performing the standard routine with respect to their obstacle as follows:

- Starting Frame: With few exceptions, the first frame represented a view of the actor completely visible.
- Hiding: All actors were instructed to hide behind their obstacle two times, with small variations in their hiding trajectory. This step allowed us to obtain varying degrees of occluded aerial views.
- Laying: To provide a variety of poses, actors were asked to lay down when completely visible and when partially occluded by their obstacle.
- Walking: Finally, actors performed a small walking trajectory at the end of their routine.
- General instructions: actors were informed of the dataset's focus on emergency response scenarios, and were therefore asked to position themselves as if they were hiding, trying to be rescued, or in need of help.

For the water routines, where the primary occlusion source was the water itself, small but important variations in instructions were given to simulate various drowning scenarios. All actors were asked to repeat their routine five times, with the sUAS set at five progressively distant locations set to 10m, 30m, 50m, 70m, and 90m, with the distance measured, through the sUAS' feedback from the First Person View (FPV) screen, horizontally and vertically from the expected starting point of the actor. Figure 2 illustrates the sUAS position at a distance of 10m.

Additionally, a reference view of the actor was filmed, with the sUAS positioned a few meters in front of the actor, while the actor performed 360°rotations. The first rotation involved arms hanging down and the second with arms extended up, providing multiple views of the actor at ground level. Finally, true negatives were also filmed by asking the actor to locate himself/herself outside of the camera view; please note that in a few cases true negatives may still contain consented non-actor participants. At the conclusion of the session, actors were given a 20$USD prepaid card.

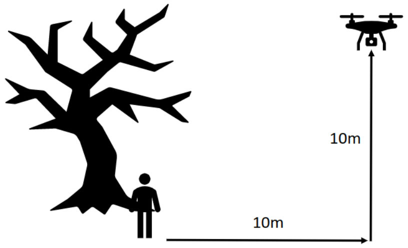


Figure 2. Filming process. Sample positioning of the sUAS at 10m horizontally and vertically from the actor's starting location.

4. Data Annotation Process

4.1. Data curation

Although efforts were made to avoid filming non-participants, during the revision of the films, unexpected persons were observed on a couple of videos. For videos where the non-participant was only visible at the beginning, at the end, or at non-keyframes, trimming the video was a direct solution, representing no impact to the quality of the data. Nevertheless, situations where found where trimming the portion of the video where the non-participant appeared on screen would represent a loss of information of the actor's performance; these situations were solved by blacking out the non-participant area of the frames.

4.2. Metadata

Metadata provided in this dataset can be divided into Environmental and Demographic categories, including outfit descriptions that could aid in re-identification tasks. Full list of metadata can be found in Tab. S1 in the Supplemental Material. Insight into selecting metadata factors was obtained through a previous series of semi-structured interviews with emergency responders, under IRB protocol 19-04-5269, to determine search terms used for describing missing persons. Clothing descriptions may include up to five words for salient figures. Hair length uses the same metric for males and females, with *bald* meaning absence of noticeable hair, *short* meaning ear-length, *medium* ranging from ear- to shoulder-length, and *long* meaning longer than shoulder-length. The Location descriptor School (Nature) aggregates filming sessions where the researcher should expect nature domination despite it being filmed at a school premises. The reported weather information was obtained from the nearest weather station to the filming location. Finally, the Exposure Value (EV) was separated from the Video descriptor; while the Video descriptor is a constant for all actors, the EV parameter was found to be different than 0 on a couple of films, indicating a change in the illumination, which is a relevant parameter for computer vision tasks [77]. Table S2 from the Supplemental Material displays the predefined lists of available colors for describing clothing and hair color. Clothing colors were selected to include the most common colors across the hue range on the HSV color space; colors for the Hair descriptor were selected based on emergency responders' classifiers.

4.3. Keyframe selection

Manually labeling every frame from the films would have been infeasible, therefore we selected 85 keyframes at each of the five distances for every actor. This resulted in 425 keyframes per actor, with the exception of Actor001, for whom 750 keyframes were selected. Keyframe selection was performed in accordance to the following guidelines: (1) the 85 selected frames tracked the actor across their entire routine, (2) each starting and ending frame of a trajectory was selected as a keyframe, (3) each change in direction of the actor's trajectory generated a new keyframe, (4) the set of keyframes included different poses (e.g., standing, sitting, laying down), and finally, (5) all keyframes had at least a part of the actor visible.

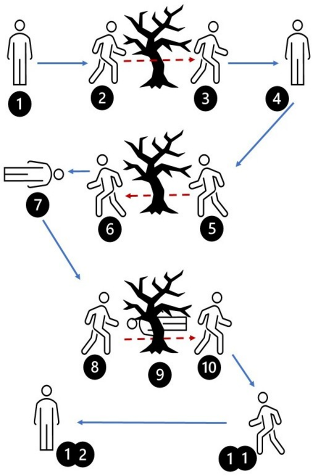
Finally, when the actor interacts with an obstacle, a custom sampling is performed to obtain views with different levels of occlusion as the actor moves behind and away from an obstacle. Figure 3a illustrates a sample routine with 12 keypoints selected following these guidelines. The red-dashed arrows indicate where sampling for occluded views would be performed. Activity labels were added to each keyframe as: Walking, Laying, Hiding, Hiding (Laying), Swimming, Drowning. Figure 3b shows sample labels for the 12 keypoints illustrated on Fig. 3a.

4.4. Annotations

All 42,825 selected frames were sent to Labelbox [44], a labelling company who employed expert annotators to add bounding boxes and visibility labels to all images.

4.4.1 Occlusion Label

Figure 4 displays how percentages were assigned to each body part of a person. The following is the procedure used to calculate the visibility label, that is, the amount of an actor that was visible at a particular instant: (1) Given an image, identify the body parts of a person that are visible. (2) Review the percentages of the identified body parts based on Fig. 4. (3) If less than half of a body part is visible, assign half of the percentage indicated in Fig. 4. (4) If more than half of the body part is visible, assign the full percentage as indicated in Fig. 4. (5) Add up the percentages obtained from each body part. (6) Assign the sum to one of the ten ranges of visibility, with upper bounds of 10 to 100. For example, selecting 10 means that the sum obtained from the percentages is greater than zero but less than or equal to 10, while selecting 20 means that the sum is greater than 10 and less than or equal to 20, and so on.



(a)

Keyframes	Activity
1 - 2	Walking
2 - 3	Hiding
3 - 5	Walking
5 - 6	Hiding
6 - 7	Walking
7	Laying
7 - 8	Walking
8 - 10	Hiding (L)
10 - 12	Walking

(b)

Figure 3. Keyframe selection process. (a) Sample routine with 12 keyframes selected. Sampling for occluded views is indicated by red-dashed arrows. (b) Sample Activity labels for the keyframes illustrated. Hiding (L) represents Hiding (Laying).

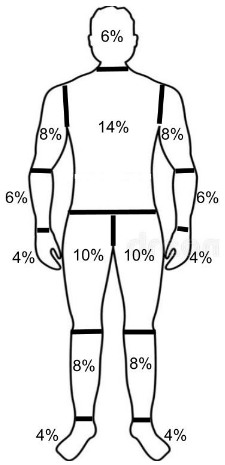


Figure 4. Visibility label calculation. Percentages assigned to each body part of a person.

Please note that shadows were not considered to be part of the human, and under normal circumstances, the actor's own clothing are not treated as a source of occlusion. Also note that although we are reporting a visibility metric, this is just the inverse of the occluded amount of the actor's body.

5. Dataset Characteristics

NOMAD provides 500 routine videos of 100 different actors, with each actor performing at five different distances, set as 10m, 30m, 50m, 70m, and 90m, including 42,825 frames manually annotated with a bounding box and visibility label. Videos' duration ranges from 30s to 180s depending on the actor's pace. It also provides one reference video per actor and 500 true negative short-duration videos, with each true negative video corresponding to one routine video. All videos are of 30fps, MP4-H265 coding, 5.4k video quality, and frames of 5472 by 3078 pixels.

5.1. Natural

Figure 5 shows the distribution of the 100 actors with respect to their filming locations. The variety of locations provides coverage of natural and man-made environments, and is aimed at training CV models for effectively supporting a wide range of emergency response scenarios.

To further increase robustness of our dataset in terms of environmental conditions, cross-seasonal imagery was collected, with temperatures ranging from 30F to 90F, wind speeds of 0MPH to 20MPH, morning, afternoon, and evening sessions, capturing hot sunny summer days, autumn colorful scenes, and winter's snowy conditions. Finally, we made every effort to mitigate potential demographic bias to support fair and equitable emergency response. Figure 6 presents the distributions of gender, age and race. The gender distribution shows a 50/50 male/female distribution, and although age distribution shows that the majority of the population was younger than 30 years old, actors across the range of 30 to 78 years old are still present in significant percentage. Finally, we show a comparison between our race distribution and the USA race distribution [13,38], showing an improvement with the purpose of generalizing CV models and mitigating potential biases [49,73]. While the USA federal census does not consider Latino/Hispanic as a race, and distribute their classification as an ethnicity distributed across races [13], we have incorporated it as a race in alignment with its recognition as a separate class by current computer vision models [70]. Lastly, although the non-rigid aspect of our routine creates uncertainty about specific levels of our visibility label, it allows the actor to provide data using more natural behaviour, adding fidelity to the actor's performance.

5.2. Occluded

NOMAD provides the data needed to face occluded persons' detection during high-pressure, life-or-death emergencies. It labels each bounding box with the degree of visibility on 10 levels, providing a representative number of frames at each level as shown in Fig. 7. The higher amount of frames at lower visibilities responds to the manual selection process as well as to the increasing annotation difficulty and additional sources of occlusion at further distances.

5.3. Multi-scale

Effective collaboration between sUAS and humans aims to exploit each of their individual strengths. sUAS have the ability to quickly scan large areas from greater altitudes, or to provide a focused close-up view of the target. NOMAD provides five different aerial distances, supporting both generalized models or models specialized for each distance. Table S3 from the Supplemental Material shows the expected minimum Ground Sampling Distance (GSD) for the five different distances, assuming that the actor is on the camera optical axis [5]. This is not always true as the actors moved to perform their routine and the camera gimbal position was often set to avoid potential areas of non-participants or areas outside the filming premises. This moved the actors away from the camera optical axis, increasing their GSD, and decreasing the number of pixels representing them, as well as creating a non-fixed pitch and adding real variance.

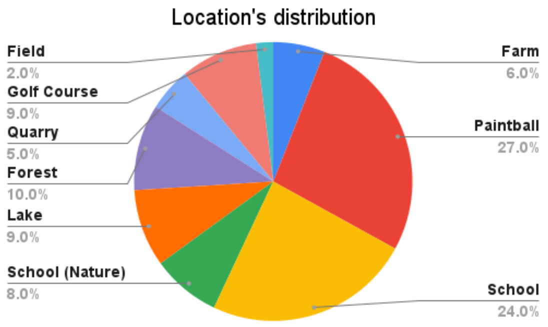


Figure 5. Distribution of the filming locations for the 100 actors.

5.4. Computer Vision Uses

The characteristics of NOMAD provide an environment to improve emergency response in four main areas of CV:

- Occlusion benchmark: the effort of NOMAD's ten levels of visibility aims to provide a new benchmark dataset to assess the research community's improvements on person detection under occlusion, a previously under-explored factor in aerial datasets.
- Person detection: search and rescue scenarios in remote areas tend to search for *any* person (i.e., blind search). The demographic and environmental diversity provided by NOMAD, as well as its multi-scale component, supported by the bounding boxes' annotations, can be leveraged to improve this general CV task.
- Person re-identification: additional to *blind search*, descriptions of the searched person translate the detection task to a CV re-identification problem, especially on crowded scenarios. NOMAD provides a rich metadata and a reference view of every actor to support re-identification tasks from aerial views.
- Person tracking: in many emergency response scenarios, the aim is to detect and then track. Due to the strategic selection of manually labelled keyframes, NOMAD allows the assessment of tracking techniques, following the actor's key movements and changes of direction throughout their full routine.

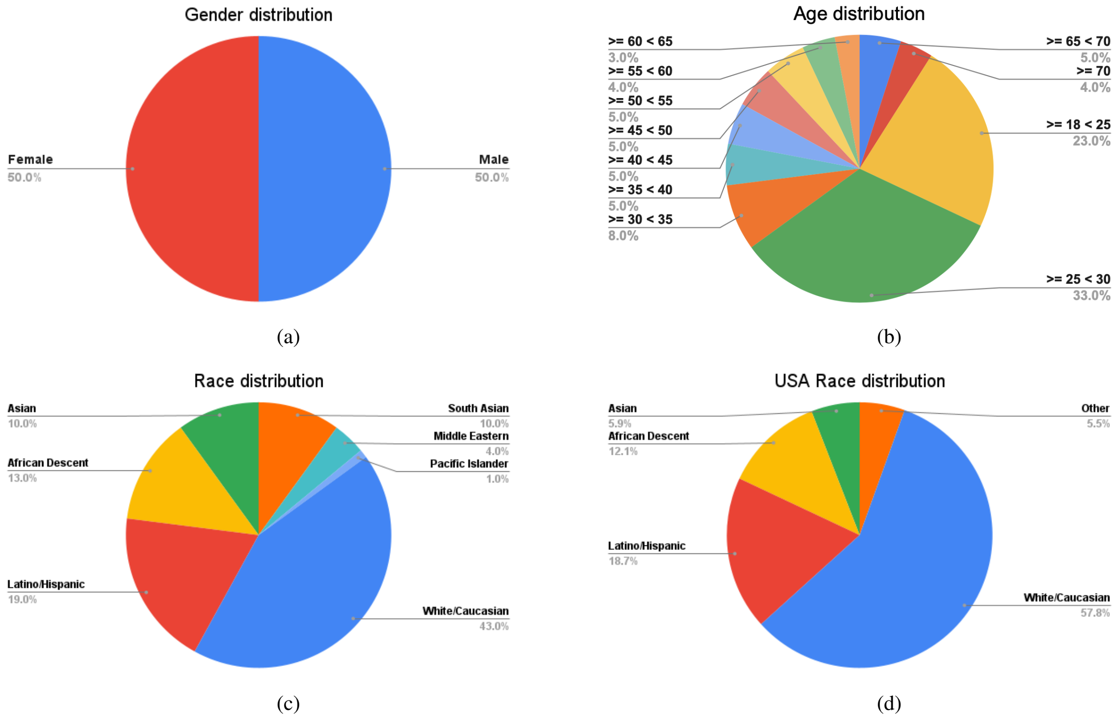


Figure 6. Distribution of the demographic descriptors of (a) Gender, (b) Age, and (c) Race, for our 100 actors. Our race distribution is compared to the (d) USA race distribution, improving generalization by mitigating possible biases.

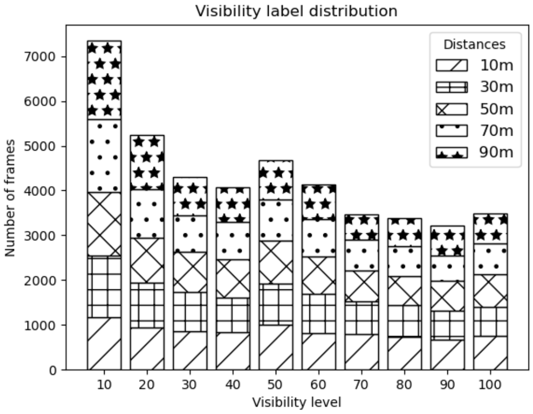
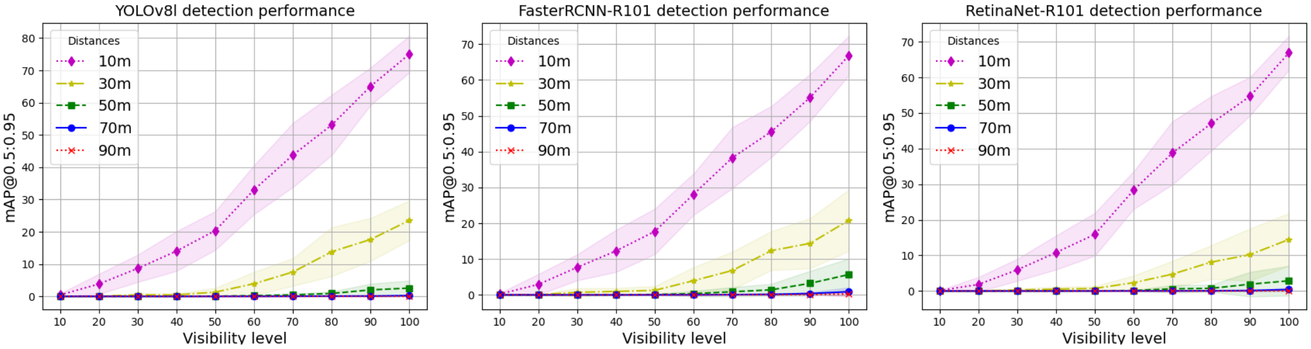


Figure 7. Distribution of the visibility label across the 42,825 manually annotated frames.

6. Computer Vision Models metrics

To demonstrate the use of NOMAD for benchmarking CV models at varying levels of occlusion, we compared the performance of three state-of-the-art CV models. Our first model was YOLOv8 from Ultralytics [39], representing the most recent upgrade to the YOLO family. YOLOv8 supports real-time detection with limited computational and memory resources, matching the requirements for sUAS-based aerial detection. Additionally, we selected a FasterRCNN and a RetinaNet model from the Detectron2 library [78]. The specific versions tested are YOLOv8l, FasterRCNN-R101-FPN, and RetinaNet-R101-FPN, with a reported mAP@0.5:0.95 of 52.9, 42.0 and 40.4 on the COCO benchmark, respectively. Available models of YOLOv8x and FasterRCNN-X101-FPN provide higher mAP, nevertheless, the latency of these models increases substantially compared to the gained mAP [39, 79].

For evaluation purposes, 10 folds of 10 actors each were randomly created, with a constant seed for reproducibility across models. From each fold, 50 tests were performed, one per each distance-visibility (5 distances, 10 visibilities), with the results of these 50 tests being averaged across the 10 folds. Figure 8 shows the averaged mAP@0.5:0.95 score and standard deviation of the CV models against different levels of occlusion and distances. We can observe that YOLOv8l performs better on the closest distance than the other models, a result supported by the higher initial mAP reported; nevertheless all models suffer from critical degradation as the distance increases. This behaviour is expected as their training data is focused on ground views rather than aerial ones. Finally, although this degradation is expected to be mitigated by fine-tuning the models with aerial data, the results emphasize the degradation problem that occlusion represents, for even though the models achieve decent scores at the closest distance with full visibility, the mAP values drastically drop as the occlusion increases. The usefulness of NOMAD to the research community can be justified from the previous baseline by three reasons: (1)



(a) (b) (c)

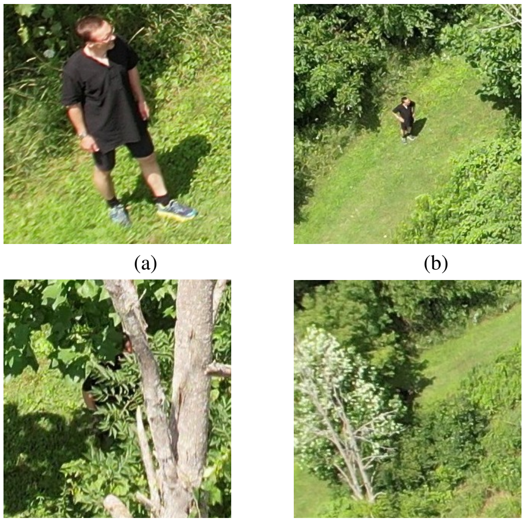
Figure 8. Performance across different levels of occlusion of (a) YOLOv8l's, (b) FasterRCNN-R101's, (c) RetinaNet-R101's pretrained weights when tested on NOMAD, with the task of person detection. Oclussion increases as the level of visibility decreases, therefore, mAP scores fall drastically as we increase in distance and occlusion. The higher performance of the models at the closest distance is expected as the data resembles ground view from COCO training data, nevertheless, mAP scores fall significantly as we increase in occlusion even for the closest distance, calling for improvement and robustness of models against occlusion in aerial views.

NOMAD is built with a real-world variance imagery (Natural), making it a fair benchmark towards emergency response scenarios; (2) occlusion on person detection can be assessed thanks to the granularity of our visibility label; (3) the multi-scale characteristic allows occlusion to be assessed also across different distances, key to the improvement of aerial detection on sUAS. To exemplify the difficulty of person detection on emergency response scenarios, Fig. 9 shows imagery from our tests with increasing difficulty, due to distance and occlusion.

7. Future Work and Conclusions

The structure and characteristics of NOMAD offer many opportunities for improvements in aerial human detection, recognition and tracking, especially the following:

- Detection under occlusion: NOMAD allows us to explore and understand the limits of detection under occluded views, with future work focusing on improving CV models' performance by exploiting human psychophysical metrics and temporal information.
- Person re-identification: Addressing emergency response scenarios, future work will focus on improving person re-identification through leveraging software architectures that support hybrid onboard/offboard solutions and integrate the human into the loop.
- Real-world deployment: We have found through our own experiences in deploying CV on sUAS that there are major additional challenges and some degradation in results. Using the models trained on NOMAD, we will deploy and evaluate occlusion-ready CV models on physical sUAS.



(a) (b) (c) (d)

Figure 9. Test samples. (a) Easy sample at 10m and 100 visibility. (b) Medium level difficulty sample (50m, 100 visibility). (c) Hard sample due to heavy occlusion (10m, 30 visibility). (d) Hard sample due to high distance and light occlusion (90m, 90 visibility).

In conclusion, as indicated by the results of our baseline evaluation, occlusion represents a non-trivial challenge that remains to be tackled. NOMAD's characteristics of Natural, Occluded, Multi-scale aerial views, provide a new benchmark dataset for tackling this challenge, and can serve as the next step in improving the accuracy of aerial search and detection for emergency response.

8. Acknowledgements

The work described in this paper was supported by the USA National Science Foundation under grant CNS-1931962. In addition, we thank all participants and annotators, as well as academic institutions, governmental agencies, private owners and company authorized supervisors for providing locations' accessibility. We also thank family and friends for their logistics support.

References

[1] Ankit Agrawal, Sophia J. Abraham, Benjamin Burger, Chichi Christine, Luke Fraser, John M. Hoeksema, Sarah Hwang, Elizabeth Travnik, Shreya Kumar, Walter J. Scheirer, Jane Cleland-Huang, Michael Vierhauser, Ryan Bauer, and Steve Cox. The next generation of human-drone partnerships: Co-designing an emergency response system. In Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik, editors, *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–13. ACM, 2020. 1, 2

[2] Md Nafee Al Islam, Muhammed Tawfiq Chowdhury, Ankit Agrawal, Michael Murphy, Raj Mehta, Daria Kudriavtseva, Jane Cleland-Huang, Michael Vierhauser, and Marsha Chechik. Configuring mission-specific behavior in a product line of collaborating small unmanned aerial systems. *Journal of Systems and Software*, 197:111543, 2023. 2

[3] Antonio Albanese, Vincenzo Sciancalepore, and Xavier Costa-Pérez. hierarquicallyo: An automated search-and-rescue drone-based solution for victims localization. *IEEE Transactions on Mobile Computing*, 21(9):3312–3325, 2022. 2

[4] Kirsnaragavan Arudpiragasam, Taraka Rama Krishna Kanth Kannuri, Klaus Schwarz, Michael Hartmann, and Reiner Creutzburg. Real-time pedestrian detection using radar-camera fusion and clustering. In *Multimodal Image Exploitation and Learning 2023*, volume 12526, pages 89–103. SPIE, 2023. 2

[5] Navaneeth Balamuralidhar, Sofia Tilon, and Francesco Nex. Multeye: Monitoring system for real-time vehicle detection, tracking and speed estimation from uav imagery on edge-computing platforms. *Remote Sensing*, 13(4), 2021. 6

[6] Arturo Miguel Russell Bernal and Jane Cleland-Huang. Hierarchically organized computer vision in support of multi-faceted search for missing persons. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–7. IEEE, 2023. 1

[7] Elizabeth Bondi, Raghav Jain, Palash Aggrawal, Saket Anand, Robert Hannaford, Ashish Kapoor, Jim Piavis, Shital Shah, Lucas Joppa, Bistra Dilkina, et al. Birdsai: A dataset for detection and tracking in aerial thermal infrared videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1747–1756, 2020. 3

[8] Aidan Boyd, Daniel Moreira, Andrey Kuehlkamp, Kevin Bowyer, and Adam Czajka. Human saliency-driven patch-based matching for interpretable post-mortem iris recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 701–710, January 2023. 2

[9] Aidan Boyd, Patrick Tinsley, Kevin W Bowyer, and Adam Czajka. Cyborg: Blending human saliency into the loss improves deep learning-based synthetic face detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6108–6117, 2023. 2

[10] Dunja Božić-Štulić, Željko Marušić, and Sven Gotovac. Deep learning approach in aerial imagery for supporting land search and rescue missions. *International Journal of Computer Vision*, 127(9):1256–1278, 2019. 3

[11] Daniel Broyles, Christopher R. Hayner, and Karen Leung. Wisard: A labeled visual and thermal image dataset for wilderness search and rescue. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9467–9474, 2022. 3

[12] Brunswick County Sheriff's Office, NC. Teamwork makes the dream work, 2022. https://www.facebook.com/brunswicksheriffNC/posts/355462306611443, Last accessed on 2023-08-27. 2

[13] Census Bureau. US census bureau quick facts, 2023. https://www.census.gov/quickfacts/, Last accessed on 2023-08-06. 6

[14] J. Casper and R.R. Murphy. Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 33(3):367–385, 2003. 2

[15] CBS NEWS. Italy police use drone to catch suspected arsonist in the act as wildfires char calabria, 2023. https://www.cbsnews.com/news/italy-fire-drone-arson-suspect-arrest-calabria-wildfires-europe-algeria/, Last accessed on 2023-08-27. 2

[16] Himanshu Chandel and Sonia Vatta. Occlusion detection and handling: a review. *International Journal of Computer Applications*, 120(10), 2015. 2

[17] Dimitrios Chatziparaschis, Michail G Lagoudakis, and Panagiotis Partsinevelos. Aerial and ground robot collaboration for autonomous mapping in search and rescue missions. *Drones*, 4(4):79, 2020. 2

[18] Xiaowu Chen, Qing Li, Dongyue Zhao, and Qinping Zhao. Occlusion cues for image scene layering. *Computer Vision and Image Understanding*, 117(1):42–55, 2013. 2

[19] Muhammed Tawfiq Chowdhury and Jane Cleland-Huang. Engineering challenges for ai-supported computer vision in small uncrewed aerial systems. In *2023 IEEE/ACM 2nd International Conference on AI Engineering – Software Engineering for AI (CAIN)*, pages 158–170, 2023. 1, 2

[20] Jane Cleland-Huang, Ankit Agrawal, Michael Vierhauser, Michael Murphy, and Mike Prieto. Extending MAPE-K to support human-machine teaming. In Bradley R. Schmerl, Martina Maggio, and Javier Cámara, editors, *International Symposium on Software Engineering for Adaptive and Self-Managing Systems, SEAMS 2022, Pittsburgh, PA, USA, May 22-24, 2022*, pages 120–131. ACM/IEEE, 2022. 1, 2

[21] CNN. Drone lifeguard saves 14-year-old from drowning, 2022. https://www.cnn.com/videos/us/2022/07/25/drone-rescues-teenager-drowning-spain-orig-cb-aw.cnn, Last accessed on 2023-08-27. 1, 2

[22] David Cornett, Joel Brogan, Nell Barber, Deniz Aykac, Seth Baird, Nicholas Burchfield, Carl Dukes, Andrew Duncan, Regina Ferrell, Jim Goddard, et al. Expanding accurate person recognition to new altitudes and ranges: The briar

dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 593–602, 2023. 3

[23] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. Meva: A large-scale multiview, multimodal video dataset for activity detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1060–1068, 2021. 3

[24] Adam Czajka, Daniel Moreira, Kevin Bowyer, and Patrick Flynn. Domain-specific human-inspired binarized statistical image features for iris recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 959–967, 2019. 2

[25] Daniel Davila, Dawei Du, Bryon Lewis, Christopher Funk, Joseph Van Pelt, Roderic Collins, Kellie Corona, Matt Brown, Scott McCloskey, Anthony Hoogs, et al. Mevid: Multi-view extended videos with identities for video person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1634–1643, 2023. 3

[26] DJI. Drone rescues around the world, 2023. https://enterprise.dji.com/drone-rescue-map/, Last accessed on 2023-08-04. 1

[27] Drones.R.Africa. Drone helps save hiker on table mountain, cape town, 2019. https://dronenews.africa/drone-helps-save-man-on-table-mountain-cape-town/, Last accessed on 2023-08-27. 1, 2

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

[29] Justin Dulay, Sonia Poltoratski, Till S Hartmann, Samuel E Anthony, and Walter J Scheirer. Guiding machine perception with psychophysics. *arXiv preprint arXiv:2207.02241*, 2022. 2

[30] Justin Dulay and Walter J Scheirer. Using human perception to regularize transfer learning. *arXiv preprint arXiv:2211.07885*, 2022. 2

[31] EL PAÍS. Gallery aerial images reveal scale of earthquake devastation in turkey and syria, 2023. https://english.elpais.com/international/2023-02-09/gallery-aerial-images-reveal-scale-of-earthquake-devastation-in-turkey-and-syria.html, Last accessed on 2023-08-27. 2

[32] EMS1. Drone helps N.C. first responders rescue missing teen, 2023. https://www.ems1.com/drones/articles/drone-helps-nc-first-responders-rescue-missing-teen-igX40AfVSNKreUJk/, Last accessed on 2023-08-27. 1, 2

[33] Han Fan, Victor Hernandez Bennetts, Erik Schaffernicht, and Achim J Lilienthal. Towards gas discrimination and mapping in emergency response scenarios using a mobile robot with an electronic nose. *Sensors*, 19(3):685, 2019. 2

[34] Ruth C Fong, Walter J Scheirer, and David D Cox. Using human brain activity to guide machine learning. *Scientific reports*, 8(1):5397, 2018. 2

[35] Prakhar Ganesh, Yao Chen, Yin Yang, Deming Chen, and Marianne Winslett. Yolo-ret: Towards high accuracy real-time object detection on edge gpus. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3267–3277, January 2022. 2

[36] Global News. Police drone locates missing 81-year-old woman in north carolina, 2017. https://globalnews.ca/news/3851197/police-drone-locates-missing-81-year-old-woman-in-north-carolina/, Last accessed on 2023-08-27. 2

[37] Samuel Grieggs, Bingyu Shen, Greta Rauch, Pei Li, Jiaqi Ma, David Chiang, Brian Price, and Walter J. Scheirer. Measuring human perception to improve handwritten document transcription. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6594–6601, 2022. 2

[38] Connie Hanzhang Jin, Ruth Talbot, and Hansi Lo Wang. What the new census data shows about race depends on how you look at it, 2021. https://www.npr.org/2021/08/13/1014710483/2020-census-data-us-race-ethnicity-diversity, Last accessed on 2023-08-06. 6

[39] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8, 2023. https://github.com/ultralytics/ultralytics, Last accessed on 2023-08-27. 2, 7

[40] K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5830–5840, June 2021. 2

[41] Tetsushi Kamegawa, Taichi Akiyama, Satoshi Sakai, Kento Fujii, Kazushi Une, Eitou Ou, Yuto Matsumura, Toru Kishutani, Eiji Nose, Yusuke Yoshizaki, et al. Development of a separable search-and-rescue robot composed of a mobile robot and a snake robot. *Advanced Robotics*, 34(2):132–139, 2020. 2

[42] KPRC 2 Click2Houston. Deputies use drone to find 3-year-old missing in washington brush, 2023. https://www.youtube.com/watch?v=tlwG-f_lRnM, Last accessed on 2023-08-27. 2

[43] SV Aruna Kumar, Ehsan Yaghoubi, Abhijit Das, BS Harish, and Hugo Proença. The p-destre: A fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices. *IEEE Transactions on Information Forensics and Security*, 16:1696–1708, 2020. 3

[44] Labelbox. Labelbox, 2023. [Online] Available: https://labelbox.com. 5

[45] Jeonghun Lee and Kwang-il Hwang. Yolo with adaptive frame control for real-time object detection applications. *Multimedia Tools and Applications*, 81(25):36375–36396, 2022. 2

[46] Lehighvalleylive. Ferry Street fire update: 10 homes damaged need to be demolished, cause still probed, 2023. https://www.lehighvalleylive.com/easton/2023/07/ferry-street-fire-update-10-

homes-damaged-need-to-be-demolished-cause-still-probed.html, Last accessed on 2023-08-27. 2

[47] 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. 3

[48] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International journal of computer vision*, 128:261–318, 2020. 1, 2

[49] Steve Lohr. Facial recognition is accurate, if you're a white guy, 2018. https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html, Last accessed on 2023-08-06. 6

[50] Georgios Loukas, Stelios Timotheou, and Erol Gelenbe. Robotic wireless network connection of civilians for emergency response operations. In *2008 23rd International Symposium on Computer and Information Sciences*, pages 1–6, 2008. 2

[51] Yuqing Ma, Hainan Li, Zhange Zhang, Jinyang Guo, Shanghang Zhang, Ruihao Gong, and Xianglong Liu. Annealing-based label-transfer learning for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11454–11463, June 2023. 2

[52] Thomas Manzini and Robin Murphy. Open problems in computer vision for wilderness sar and the search for patricia wu-murad. *arXiv preprint arXiv:2307.14527*, 2023. 1, 2

[53] MarketsAndMarkets. UAV market by point of sale, systems, platform, function, end use, application, type, mode of operation, mtow, range and region - global forecast to 2027, 2022. https://www.marketsandmarkets.com/Market-Reports/unmanned-aerial-vehicles-uav-market-662.html, Last accessed on 2023-08-04. 1

[54] Patrick McEnroe, Shen Wang, and Madhusanka Liyanage. A survey on the convergence of edge computing and AI for UAVs: Opportunities and challenges. *IEEE Internet of Things Journal*, 9(17):15435–15459, 2022. 1, 2

[55] Shuyu Miao, Shanshan Du, Rui Feng, Yuejie Zhang, Huayu Li, Tianbi Liu, Lin Zheng, and Weiguo Fan. Balanced single-shot object detection using cross-context attention-guided network. *Pattern recognition*, 122:108258, 2022. 2

[56] Chintakindi Balaram Murthy, Mohammad Farukh Hashmi, and Avinash G Keskar. Efficientlitedet: a real-time pedestrian and vehicle detection algorithm. *Machine Vision and Applications*, 33(3):47, 2022. 2

[57] Jamuna S Murthy, GM Siddesh, Wen-Cheng Lai, BD Parameshachari, Sujata N Patil, and KL Hemalatha. Object-detect: A real-time object detection framework for advanced driver assistant systems using yolov5. *Wireless Communications and Mobile Computing*, 2022, 2022. 2

[58] Keiji Nagatani, Seiga Kiribayashi, Yoshito Okada, Kazuki Otake, Kazuya Yoshida, Satoshi Tadokoro, Takeshi Nishimura, Tomoaki Yoshida, Eiji Koyanagi, Mineo Fukushima, et al. Emergency response to the nuclear accident at the fukushima daiichi nuclear power plants using mobile rescue robots. *Journal of Field Robotics*, 30(1):44–63, 2013. 2

[59] 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. 1, 3

[60] Chen Ning, Li Menglu, Yuan Hao, Su Xueping, and Li Yunhong. Survey of pedestrian detection with occlusion. *Complex & Intelligent Systems*, 7:577–587, 2021. 1, 2

[61] Farzad Niroui, Kaicheng Zhang, Zendai Kashino, and Goldie Nejat. Deep reinforcement learning robot for search and rescue applications: Exploration in unknown cluttered environments. *IEEE Robotics and Automation Letters*, 4(2):610–617, 2019. 2

[62] Chen Pan and Wei Qi Yan. Object detection based on saturation of visual perception. *Multimedia Tools and Applications*, 79:19925–19944, 2020. 2

[63] Maria Gaia Pensieri, Mauro Garau, and Pier Matteo Barone. Drones as an integral part of remote sensing technologies to help missing people. *Drones*, 4(2):15, 2020. 2

[64] Aveek Purohit, Zheng Sun, Frank Mokaya, and Pei Zhang. Sensorfly: Controlled-mobile sensing platform for indoor emergency response applications. In *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks*, pages 223–234, 2011. 2

[65] Frédéric Py, Giulia Robbiani, Giancarlo Marafioti, Yu Ozawa, Masahiro Watanabe, Kenichi Takahashi, and Satoshi Tadokoro. Smurf software architecture for low power mobile robots: experience in search and rescue operations. In *2022 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pages 264–269, 2022. 2

[66] Brandon RichardWebster, Samuel E. Anthony, and Walter J. Scheirer. Psyphy: A psychophysics driven evaluation framework for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2280–2286, 2019. 2

[67] Kaziwa Saleh, Sándor Szénási, and Zoltán Vámossy. Occlusion handling in generic object detection: A review. In *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 000477–000484, 2021. 1, 2

[68] Sasa Sambolek and Marina Ivasic-Kos. Search and rescue image dataset for person detection - SARD, 2021. IEEE Dataport, https://dx.doi.org/10.21227/ahxm-k331. 3

[69] Walter J. Scheirer, Samuel E. Anthony, Ken Nakayama, and David D. Cox. Perceptual annotation: Measuring human vision to improve computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1679–1686, 2014. 2

[70] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021. 6

[71] Hazim Shakhatreh, Ahmad H. Sawalmeh, Ala Al-Fuqaha, Zuochao Dou, Eyad Almaita, Issa Khalil, Noor Shamsiah Othman, Abdallah Khreishah, and Mohsen Guizani. Unmanned aerial vehicles (uavs): A survey on civil applications and key research challenges. *IEEE Access*, 7:48572–48634, 2019. 1, 2

[72] Niko Sünderhauf, Oliver Brock, Walter Scheirer, Raia Hadsell, Dieter Fox, Jürgen Leitner, Ben Upcroft, Pieter Abbeel, Wolfram Burgard, Michael Milford, et al. The limits and potentials of deep learning for robotics. *The International journal of robotics research*, 37(4-5):405–420, 2018. 2

[73] Kushal Vangara, Michael C King, Vitor Albiero, Kevin Bowyer, et al. Characterizing the variability in face recognition accuracy relative to race. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 6

[74] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7464–7475, June 2023. 2

[75] Hongling Wang, Chengjin Zhang, Yong Song, Bao Pang, and Guangyuan Zhang. Three-dimensional reconstruction based on visual slam of mobile robot in search and rescue disaster scenarios. *Robotica*, 38(2):350–373, 2020. 2

[76] Christian Wankmüller, Maximilian Kunovjanek, and Sebastian Mayrgündter. Drones in emergency response–evidence from cross-border, multi-disciplinary usability tests. *International Journal of Disaster Risk Reduction*, 65:102567, 2021. 2

[77] Haiyu Wu, Vítor Albiero, KS Krishnapriya, Michael C King, and Kevin W Bowyer. Face recognition accuracy across demographics: Shining a light into the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1041–1050, 2023. 4

[78] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 7

[79] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2 ModelZoo, 2019. https://github.com/facebookresearch/detectron2/blob/main/MODEL_ZOO.md, Last accessed on 2023-10-31. 7

[80] Zhiheng Wu, Yue Lu, Xingyu Chen, Zhengxing Wu, Liwen Kang, and Junzhi Yu. Uc-owod: Unknown-classified open world object detection. In *European Conference on Computer Vision*, pages 193–210. Springer, 2022. 2

[81] Shizhou Zhang, Qi Zhang, Yifei Yang, Xing Wei, Peng Wang, Bingliang Jiao, and Yanning Zhang. Person re-identification in aerial imagery. *IEEE Transactions on Multimedia*, 23:281–291, 2020. 3

[82] Zhishuai Zhang, Cihang Xie, Jianyu Wang, Lingxi Xie, and Alan L Yuille. Deepvoting: A robust and explainable deep network for semantic part detection under partial occlusion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1372–1380, 2018. 1

[83] Xiaowei Zhao, Xianglong Liu, Yifan Shen, Yixuan Qiao, Yuqing Ma, and Duorui Wang. Revisiting open world object detection. *arXiv preprint arXiv:2201.00471*, 2022. 2

[84] Jiyang Zheng, Weihao Li, Jie Hong, Lars Petersson, and Nick Barnes. Towards open-set object detection and discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3961–3970, June 2022. 2

[85] Hongru Zhu, Peng Tang, Jeongho Park, Soojin Park, and Alan Yuille. Robustness of object recognition under extreme occlusion in humans and computational models. *arXiv preprint arXiv:1905.04598*, 2019. 1, 2

[86] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2021. 3

[87] Orr Zohar, Kuan-Chieh Wang, and Serena Yeung. Prob: Probabilistic objectness for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11444–11453, June 2023. 2

A. Supplemental tables

Category	Descriptor	Values
Demographic	Age	Years integer.
	Weight	Pounds integer.
	Height	Inches integer.
	Gender	Male, Female
	Race	White Caucasian, Latino Hispanic, African Descent, Asian, South Asian, Middle Eastern, Pacific Islander.
	Clothes	Including no more than three dominant clothes, described by its relative location to the body (Upper, Lower, Both), its color (see color list), a pattern (Horizontal Stripes, Vertical Stripes, Both, Plain, Random), and a figure.
	Hat	Any form of head covering, described by color.
	Gloves	Described by color.
	Shoes	Described by color and type. Types are categorized in 6 classes, by their coverage and formality: Bare, Open Informal (*e.g.* sandals), Open Formal (*e.g.* heels), Close Informal (*e.g.* tennis), Close Formal (*e.g.* Oxford), and Boots.
	Facial Hair	Mustache, Beard.
	Hair	Described by color and length (Bald, Short, Medium, Long).
Environmental	Date	MM:DD format.
	Time	Integer, in 24hr format.
	Location	Lake, Forest, Field, Golf Course, Quarry, Farm, School, School (Nature), Paintball.
	Weather	Temperature in Fahrenheit, wind speed in MPH, and a word descriptor (*e.g.* ”Sunny”).
	Video	5.4k video resolution, 30 fps, 5472 by 3078 pixels’ frames, and focal length of 8 mm.
	EV	String of 6 values, in increasing order of altitude and starting from the reference.

Table 1. Metadata information attached to every actor. Demographic category includes all descriptors that could aid in identification tasks, including outfit description.

Descriptors	Colors
Clothes, Hat, Gloves, Shoes	Red, Orange, Yellow, Green, Cyan, Blue, Purple, Pink, White, Gray, Black, Brown.
Hair	Black, Blond, Sandy, Brown, Gray, White, Red, Green, Blue, Pink, Other.

Table 2. List of colors associated to different metadata descriptors.

Distance	GSD [mm/pixel]	Face area [pixels]
10	4.242	1981
30	12.729	220
50	21.213	79
70	29.697	40
90	38.184	24

Table 3. Ground Sampling Distance (GSD) values at camera optical axis position, as well as the amount of pixels to cover the area of a person’s face assuming a rectangle area of 155mm by 230mm.