



**HAL**  
open science

## **EyeTrackUAV2: a Large-Scale Binocular Eye-Tracking Dataset for UAV Videos**

Anne-Flore Perrin, Vassilios Krassanakis, Lu Zhang, Vincent Ricordel,  
Matthieu Perreira, Da Silva, Olivier Le Meur

► **To cite this version:**

Anne-Flore Perrin, Vassilios Krassanakis, Lu Zhang, Vincent Ricordel, Matthieu Perreira, et al.. EyeTrackUAV2: a Large-Scale Binocular Eye-Tracking Dataset for UAV Videos. Drones, 2020, Drones 2020, 4 (2). hal-02391832v2

**HAL Id: hal-02391832**

**<https://univ-rennes.hal.science/hal-02391832v2>**

Submitted on 17 Dec 2019 (v2), last revised 10 Jan 2020 (v3)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.







L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC0 - Public Domain Dedication 4.0 International License

Article

# EyeTrackUAV2: a Large-Scale Binocular Eye-Tracking Dataset for UAV Videos

Anne-Flore Perrin <sup>1</sup>, Vassilios Krassanakis <sup>2,4</sup>, Lu Zhang <sup>3</sup>, Vincent Ricordel <sup>2</sup>, Matthieu Perreira Da Silva <sup>2</sup>, and Olivier Le Meur <sup>1</sup>

<sup>1</sup> Univ Rennes, CNRS, IRISA, 263 Avenue Général Leclerc, 35000 Rennes, France; anne-flore.perrin@irisa.fr, olemeur@irisa.fr

<sup>2</sup> Polytech Nantes, Laboratoire des Sciences du Numérique de Nantes (LS2N), Université de Nantes, 44306 Nantes CEDEX 3, France; matthieu.perreiradasilva@univ-nantes.fr, Vincent.Ricordel@univ-nantes.fr

<sup>3</sup> Univ Rennes, INSA Rennes, CNRS, IETR - UMR 6164, 35000 Rennes, France; lu.ge@insa-rennes.fr

<sup>4</sup> University of West Attica, Department of Surveying and Geoinformatics Engineering, 28 Agiou Spyridonos Str., 12243, Aigaleo, Greece; krasvas@uniwa.gr

\* Correspondence: anne-flore.perrin@irisa.fr; Tel.: +33-299-84-25-73 (A-F.P.)

Version December 13, 2019 submitted to Drones

**Abstract:** The fast and tremendous evolution of the Unmanned Aerial Vehicle (UAV) imagery gives place to the multiplication of applications in various fields such as military and civilian surveillance, delivery services, and wildlife monitoring. Combining UAV imagery with study of dynamic salience further extends the number of future applications. Indeed, considerations of visual attention open the door to new avenues in a number of scientific fields such as compression, retargeting, and decision-making tools. To conduct saliency studies, we identified the need for new large-scale eye-tracking datasets for visual salience in UAV content. Therefore, we address this need by introducing the dataset *EyeTrackUAV2*. It consists of the collection of precise binocular gaze information (1000 Hz) over 43 videos (RGB, 30 fps, 1280x720 or 720x480). Thirty participants observed stimuli under both free viewing and task conditions. Fixations and saccades were then computed with the I-DT algorithm, while gaze density maps were calculated by filtering eye positions with a Gaussian kernel. An analysis of collected gaze positions provides recommendations for visual salience ground-truth generation. It also sheds light upon variations of saliency biases in UAV videos when opposed to conventional content, especially regarding the center bias.

**Keywords:** Dataset, Saliency, Unmanned Aerial Vehicles (UAV), Videos, Visual attention, eye tracking, surveillance.

## 1. Introduction

For a couple of decades now, we have witnessed the fast advances and growing use of UAVs for multiple critical applications. UAVs refer here to unmanned aerial vehicles, autonomous or monitored from remote sites. This imagery enables a broad range of uses, from making vacation movies to drone races for mainstream civilian applications. Tremendous professional services are developed, among others fire detection [1], wildlife counting [2], journalism [3], precision agriculture, and delivery services. But most applications are military, from aerial surveillance [4], drone-based warfare [5] to moving targets tracking [6], object, person, and anomaly detection [7–9].

The UAV imagery proposes a new representation of visual scenes that makes all these new applications possible. UAV vision is dominant and hegemonic [10]. The bird point of view modifies the perspective, size and features of objects [11]. It introduces a loss of pictorial depth cues [12] such as horizontal line [13]. Also, UAV high autonomy in conjunction with large-field of view camera permits to cover large areas in limited time duration. Besides, embedded sensors can be multi-modal

30 and can include RGB, thermal, Infra-Red (IR), or multi-spectral sensors. Multiplying imagery  
31 modalities allows overcoming possible weaknesses of RGB-only cameras [10]. For instance, occlusions  
32 may be compensated by thermal information, and the capture of IR is desired for low-luminance  
33 environments [14].

34 UAV scene depiction is rich, comprehensive, and promising, which explains its success. But  
35 challenges to come are even more compelling. Edney-Browne [10] wondered how the capacity of  
36 UAV capturing the external reality (visuality) is related to perceptual and cognitive vision in humans.  
37 Variations in UAV characteristics such as perspective view and object size may change viewers'  
38 behavior towards content. Consequently, new visual attention processes may be triggered for this  
39 specific imaging. This means that studying UAV imagery in light of human visual attention not  
40 only opens the door to plenty of applications but could also enable to gather further knowledge on  
41 perceptual vision and cognition.

42 In the context of UAV content, there are very few eye-tracking datasets. This is the reason why  
43 we propose and present in this paper a new large-scale eye-tracking dataset, freely downloadable  
44 from internet. This dataset aims to strengthen our knowledge on human perception and could play a  
45 fundamental role for designing new computational models of visual attention.

46 The paper is organized as follows. In Section 2, we first justify and elaborate on the need for  
47 large-scale eye-tracking databases for UAV videos. Then, we introduce the entire process of dataset  
48 creation in Section 3. It describes the content selection, the experiment set up, and the implementation  
49 of fixations, saccades, and gaze density maps. Section 4 presents an in-depth analysis of the dataset.  
50 The study is two-fold: it explores what ground truth should be used for saliency studies, and brings  
51 to light the fading of conventional visual biases UAV stimuli. Finally, conclusions are provided in  
52 Section 5.

## 53 2. Related Work

54 Visual attention occurs to filter and sort out visual clues. Indeed, it is impossible to process  
55 simultaneously all the information of our visual field. Particular consideration should be dedicated to  
56 identifying which attentional processes are involved as they are diverse and aim at specific behaviors.  
57 For instance, one must make the distinction between overt and covert attention [15]. The former refers  
58 to a direct focus where eyes and head point. The latter relates to the peripheral vision, where attention  
59 is directed without eye movements towards it. In practice, when an object of interest is detected in  
60 the area covered by the covert attention, one may make a saccade movement to direct the eyes from  
61 the overt area to this position. The context of visualization is also important. For instance, we make a  
62 distinction between two content exploration processes [16]: (1) A no constraint examination named  
63 free viewing. The observer is rather free from cognitive loads and is supposed to mainly use bottom-up  
64 or exogenous attention processes driven by external factors, e.g. content and environment stimuli.  
65 (2) A task-based visualization, such as surveillance for instance. Cognitive processes such as prior  
66 knowledge, willful plans, and current goals guide the viewer's attention. This is known as top-down  
67 or endogenous attention. A strict division is slightly inaccurate in that both top-down and bottom-up  
68 processes are triggered during a visual stimuli in a very intricate interaction [17].

69 In computer science, it is common to study bottom-up and top-down processes through the visual  
70 saliency. Visual saliency is a representation of visual attention in multimedia content as a probability  
71 distribution per pixels [18]. Saliency analyses rest on the relation of visual attention to eye movements,  
72 and these latter are obtained through gaze collection with eye-trackers [19]. Saliency predictions help  
73 to understand computational cognitive neuroscience as it reveals attentional behaviors and systematic  
74 viewing tendencies such as center bias [17]. Multiple applications derive from saliency predictions  
75 such as compression [20], content-aware re-targeting, object segmentation [21], and detection [22,23].

76 Recently, there has been a growing interest on one particular application, which combines visual  
77 saliency and UAV content. Information overload in the drone program and fatigue in military operators  
78 may have disastrous consequences for military applications [10]. New methods and approaches are

79 required to detect anomaly in UAV footages and to ease the decision-making. Among them, we believe  
80 that computational models of visual attention could be used to simulate operators' behaviors [24].  
81 Eventually, thanks to predictions, operators' workloads can be reduced by eliminating unnecessary  
82 footages segments. Other works support the use of salience to enhance the efficiency of target-detection  
83 task completion. For instance Brunyé et al. [25] studied the combination of salience (in terms of opacity  
84 with the environment) and biological motion (presence and speed) in textured backgrounds. They  
85 concluded that salience is very important for slowly moving objects, such as camouflaged entities.  
86 Meanwhile, fast biological movements are highly attention-grabbing, which diminishes the impact  
87 of static salience. Accordingly, it makes sense to develop dynamic saliency models tailored to UAV  
88 content.

89 However, we demonstrate in [26] that current saliency models lack efficiency in terms of prediction  
90 for UAV content. This applies to all types of prediction models: handcrafted features and architecture  
91 implementing deep learning to a lesser extent, whether they are static or dynamic schemes. Typical  
92 handcrafted and low-level features learnt on conventional imaging may not suit UAV content. Besides,  
93 in conventional imaging the center position is the best location to have access to most visual information  
94 of a content [27]. This fact leads to a well-known bias in visual attention named central bias. This  
95 effect may be associated with various causes. For instance, Tseng et al. [28] showed a contribution  
96 of photographer bias, viewing strategy, and to a lesser extent, motor, re-centering, and screen center  
97 biases to the center bias. They are briefly described below:

- 98 • The **photographer bias** often emphasizes objects in the content center through composition and  
99 artistic intent [28].
- 100 • Directly related to photographer bias, observers tend to learn the probability of finding salient  
101 objects at the content center. We refer to this behavior as a **viewing strategy**.
- 102 • With regards to the Human Visual System (HVS), the central orbital position, that is when  
103 looking straight ahead, is the most comfortable eye position [29], leading to a **recentering bias**.
- 104 • Additionally, there is a **motor bias**, in which one prefers making short saccades and horizontal  
105 displacements [30,31].
- 106 • Lastly, onscreen presentation of visual content pushes observers to stare at the center of the  
107 screen frame [27,32]. This experimental bias is named the **screen center bias**.

108 The central bias is so critical in the computational modelling of visual attention that saliency models  
109 include this bias as prior knowledge or use it as a baseline to which saliency predictions are being  
110 compared [33]. The center bias is often represented by a centered isotropic Gaussian stretched to the  
111 video frame aspect ratio [34,35]. The presence of this bias in UAV videos has already been questioned  
112 in our previous work [26]. We showed that saliency models that heavily rely on the center bias were  
113 less efficient on UAV videos than on conventional video sequences. Therefore, we believe that the  
114 central bias could be less significant in drone footage as a result of the lack of photographer bias or  
115 due to UAV content characteristics. It would be beneficial to evaluate qualitatively and quantitatively  
116 the center bias on a larger dataset of UAV videos to support our assumption.

117 While it is now rather easy to find eye tracking data on typical images [34,36–44] or videos [45–49],  
118 and that there are many UAV content datasets [7,50–61], it turns out to be extremely difficult to find  
119 eye-tracking data on UAV content. This is even truer when we consider dynamic salience, which refers  
120 to salience for video content. To the best of our knowledge, *EyeTrackUAV1* dataset released in 2018 [11]  
121 is the only public dataset available for studying the visual deployment over UAV video. There exist  
122 another dataset *AVS1K* [62]. However, *AVS1K* is, to the present day, not publicly available. We thus  
123 focus here on *EyeTrackUAV1*, with the awareness that all points below but the last apply to *AVS1K*.

124 *EyeTrackUAV1* consists in 19 sequences (1280x720 and 30 frame per second (fps)) extracted from  
125 the *UAV123* database [54]. The sequence selection relied on content characteristics, which are the  
126 diversity of environment, distance and angle to the scene, size of the principal object, and the presence  
127 of sky. Precise binocular gaze data (1000 Hz) of 14 observers were recorded under free viewing  
128 condition, for every content. Overall, the dataset comprises eye-tracking information on 26599 frames,

129 which represents 887 seconds of video. In spite of a number of merits, this dataset presents several  
130 limitations for saliency prediction applications. These limitations have been listed in [26]. We briefly  
131 summarize them below:

- 132 • UAV may embed multi-modal sensors during the capture of scenes. Besides conventional RGB  
133 cameras, to name but a few thermal, multi-spectral, and infrared cameras consist of typical UAV  
134 sensors. Unfortunately, *EyeTrackUAV1* lacks non-natural content, which is of great interest for  
135 the dynamic field of salience. As already mentioned, combining content from various imagery in  
136 datasets is advantageous for numerous reasons. It is necessary to continue efforts toward the  
137 inclusion of more non-natural content in databases.
- 138 • In general, the inclusion of more participants in the collection of human gaze is encouraged.  
139 Indeed, reducing variable errors by including more participants in the eye tracking experiment  
140 is beneficial. It is especially true in the case of videos as salience is sparse due to the short  
141 displaying duration of a single frame. With regards to evaluation analyses, some metrics  
142 measuring similarity between saliency maps consider fixation locations for saliency comparison  
143 (e.g. any variant of Area Under the Curve (AUC), Normalized Scanpath Saliency (NSS), and  
144 Information Gain (IG)). Having more fixation points is more convenient for the use of such  
145 metrics.
- 146 • *EyeTrackUAV1* contains eye-tracking information recorded during free-viewing sessions. That  
147 is, no specific task was assigned to observers. Several applications for UAV and conventional  
148 imaging could benefit from the analysis and reproduction of more top-down attention, related to  
149 a task at hand. More specifically, for UAV content, there is a need for specialized computational  
150 models for person or anomaly detection.
- 151 • Even though there are about 26599 frames in *EyeTrackUAV*, they come from "only" 19 videos.  
152 Consequently, this dataset just represents a snapshot of the reality. We aim to go further by  
153 introducing more UAV content.

154 To extend and complete the previous dataset and to tackle these limitations, we have created the  
155 *EyeTrackUAV2* dataset, introduced below.

### 156 3. *EyeTrackUAV2* dataset

157 This section introduces the new dataset *EyeTrackUAV2* aiming at tackling issues mentioned above.  
158 *EyeTrackUAV2* includes more video content than its predecessor *EyeTrackUAV1*. It involves more  
159 participants, and considers both free and task-based viewing. In the following subsections, we first  
160 elaborate on the selection of video content, followed by a description of the eye-tracking experiment.  
161 It includes the presentation of the eye-tracking apparatus, the experiment procedure and setup, and  
162 the characterization of population samples. Finally, we describe the generation of the human ground  
163 truth, i.e. algorithms for fixation and saccade detection as well as gaze density map computation.

#### 164 3.1. Content selection

165 Before collecting eye-tracking information, experimental stimuli were selected from multiple  
166 UAV video datasets. We paid specific attention to select videos suitable for both free and task-based  
167 viewing as experimental conditions. Also, the set of selected videos has to cover multiple UAV flight  
168 altitudes, main surrounding environments, main sizes of observed objects and angles between the  
169 aerial vehicle and the scene, as well as the presence or not of sky. We consider these characteristics  
170 favor the construction of a representative dataset of typical UAV videos, as suggested in [11].

171 We have examined the following UAV datasets: UCF's dataset<sup>1</sup>, VIRAT [50], MRP [51], the  
172 privacy-based mini-drones dataset [52], the aerial videos dataset described in [53], UAV123 [54],

---

<sup>1</sup> [http://crcv.ucf.edu/data/UCF\\_Aerial\\_Action.php](http://crcv.ucf.edu/data/UCF_Aerial_Action.php)

Dataset	Native resolution	Proportion of content seen per degree of visual angle (%)	Videos number	Frames number (30 fps)	Duration (sec)
VIRAT [50]	720 x 480	1,19	12	17851	595,03
UAV123 [54]	1280 x 720	0,44	22	20758	691,93
DTB70 [56]	1280 x 720	0,44	9	3632	121,07
Overall			<b>43</b>	<b>42241</b>	<b>1408,03 (23:28 min)</b>

**Table 1.** Stimuli original datasets.

	Number of frames				Duration (MM:SS)			
	VIRAT	UAV123	DTB70	Overall	VIRAT	UAV123	DTB70	Overall
Total	17851	20758	3632	42241	09:55	11:32	02:01	23:28
Average	1488	944	404	982	00:50	00:31	00:13	00:33
Standard Deviation	847	615	177	727	00:28	00:21	00:06	00:24
Minimum	120	199	218	120	00:04	00:07	00:07	00:04
Maximum	3178	2629	626	3178	01:46	01:28	00:21	01:46

**Table 2.** Basic statistics on selected videos.

173 DTB70 [56], Okutama-Action [57], VisDrone [63], CARPK [58], SEAGULL [59], DroneFace [60], and  
 174 the aerial video dataset described in [55]). A total of 43 videos (RGB, 30 fps, 1280x720 or 720x480)  
 175 have been selected from databases *VIRAT*, *UAV123* and *DTB70*. These three databases are exhibiting  
 176 different contents for various applications, which makes the final selection representative of the UAV  
 177 ecosystem. We present below the main characteristics of the three selected datasets:

- 178 • *UAV123* includes challenging UAV content annotated for object tracking. We restrict the content  
 179 selection to the first set, which includes 103 sequences (1280x720 and 30 fps) captured by an  
 180 off-the-shelf professional-grade UAV (DJI S1000) tracking various objects in a range of altitudes  
 181 comprised between 5-25 meters. Sequences include a large variety of environments (e.g. urban  
 182 landscapes, roads, and marina), objects (e.g. cars, boats, and persons) and activities (e.g. walking,  
 183 biking, and swimming) as well as present many challenges for object tracking (e.g. long- and  
 184 short-term occlusions, illumination variations, viewpoint change, background clutter, and camera  
 185 motion).
- 186 • Aerial videos in the *VIRAT* dataset were manually selected (for smooth camera motion and  
 187 good weather conditions) from rushes of a total amount of 4 hours in outdoor areas with broad  
 188 coverage of realistic scenarios for real-world surveillance. Content includes "single person",  
 189 "person and vehicle", and "person and facility" events, with changes in viewpoints, illumination,  
 190 and visibility. The dataset comes with annotations of moving object tracks and event examples  
 191 in sequences. The main advantage of *VIRAT* videos is its perfect fit for military applications. It  
 192 covers fundamental environment contexts (events), conditions (rather poor quality and weather  
 193 condition impairments), and imagery (RGB and IR). We decided to keep the original resolution  
 194 of videos (720x480) to prevent the introduction of unrelated artifacts.
- 195 • The 70 videos (RGB, 1280x720 and 30 fps) from *DTB70* dataset are manually annotated with  
 196 bounding boxes for tracked objects. Sequences were shot with a DJI Phantom 2 Vision+ drone or  
 197 were collected from YouTube to add diversity in environments and target types (mostly humans,  
 198 animals, and rigid objects). There is also a variety of camera movements (both translation and  
 199 rotation), short- and long-term occlusions, and target deformability.

200 Table 1 reports for each database the number of sequences selected, their native resolution, duration  
 201 and frame number. Table 2 presents basic statistics of the database in terms of number of frames and  
 202 duration.

### 203 3.2. Content Diversity

204 Figure 1 presents the diversity of selected UAV sequences by illustrating the first frame of every  
 205 content. Visual stimuli cover a variety of visual scenes in different environments (e.g. public and

206 military environments, roads, buildings, sports, and port areas, etc.) and different moving or fixed  
 207 objects (e.g. people, groups of people, cars, boats, bikes, motorbikes, etc.). Selected videos were  
 208 captured from various flight heights and different angles between the UAV and the ground (allowing  
 209 or not the presence of sky during their observation). Also, three sequences, extracted from the VIRAT  
 210 dataset, were captured by IR cameras. Additionally, we considered various video duration as the  
 211 length of the video may possibly impact the behavior of observers due to fatigue, resulting in a lack of  
 212 attention and more blinking artifacts [10,64].

213 To quantitatively show the diversity of selected videos, we have computed temporal and spatial  
 214 complexity [65], named TI ( $\in [0, +\infty)$ ) and SI ( $\in [0, +\infty)$ ), respectively. These features are commonly used  
 215 in image quality domain for describing the properties of selected images. They characterize the  
 216 maximum standard deviation of spatial and temporal discrepancies over the entire sequence. The  
 217 higher a measure is, the more complex is the content. TI and SI are reported per sequence in Table 3.  
 218 The range of temporal complexity in sequences is broad, displaying the variety of movements present  
 219 in sequences. Spatial measures are more homogeneous. Indeed, the spatial complexity is due to  
 220 the bird point of view of the sensor. The aircraft high up position offers access to a large amount of  
 221 information. Table 3 reports a number of information for all selected sequences.

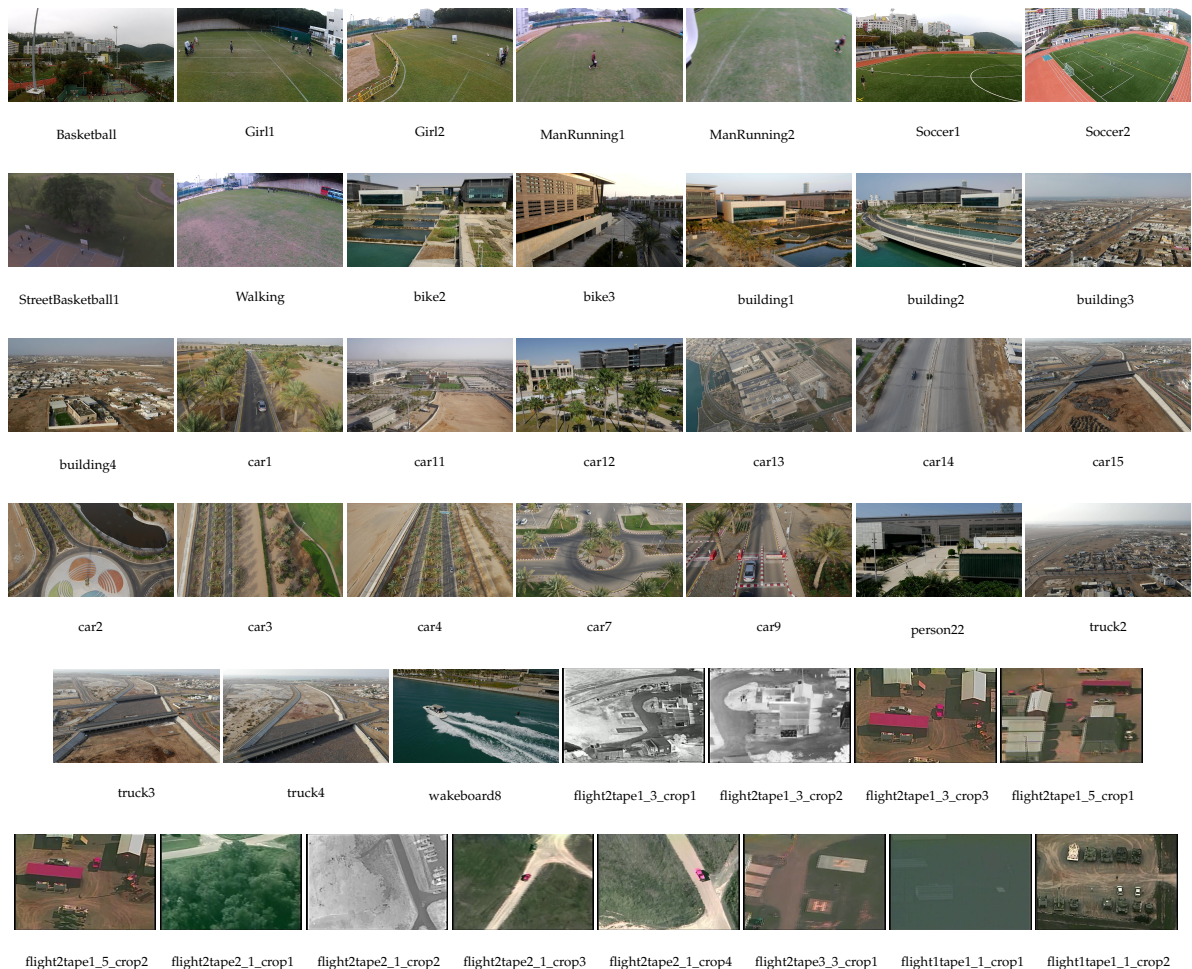


Figure 1. *EyeTrackUAV2* dataset: first frame of each sequence.

ID	Video	Dataset	Number of frames	Start frame	End frame	Duration (msec)	SI	TI	Altitude	Environment	Object size	Horizontal line (sea, sky)	Main angle	
1	09152008flight2tape1_3 (crop 1)	VIRAT	120	1	120	4000	0,455	32	High	Urban military - <b>IR</b>	Small	False	Oblique	
2	09152008flight2tape1_3 (crop 2)		367	137	503	12234	0,474	35	High	Urban military - <b>IR</b>	Small	False	Oblique	
3	09152008flight2tape1_3 (crop 3)		3178	4735	7912	105934	0,452	43	Intermediate	Urban military	Medium, Small	False	Oblique	
4	09152008flight2tape1_5 (crop 1)		972	218	1189	32400	0,467	37	Intermediate	Urban military	Medium, Small	False	Oblique	
5	09152008flight2tape1_5 (crop 2)		1715	4555	6269	57167	0,461	45	Intermediate	Urban military	Medium, Small	False	Oblique	
6	09152008flight2tape2_1 (crop 1)		1321	1	1321	44034	0,484	40	Intermediate, Low	Urban military	Medium, Big	False	Oblique	
7	09152008flight2tape2_1 (crop 2)		1754	2587	4340	58467	0,484	41	High	Roads rural - <b>IR</b>	Small	False	Oblique	
8	09152008flight2tape2_1 (crop 3)		951	4366	5316	31700	0,482	33	Intermediate	Urban military	Medium, Big	False	Oblique	
9	09152008flight2tape2_1 (crop 4)		1671	6482	8152	55700	0,452	32	High	Roads rural	Medium	False	Oblique, Vertical	
10	09152008flight2tape3_3 (crop 1)		2492	3067	5558	83067	0,474	42	Intermediate	Urban military	Small	False	Oblique	
11	09162008flight1tape1_1 (crop 1)		1894	1097	2990	63134	0,448	39	Low	Urban military, Roads rural	Medium, Small	False	Oblique	
12	09162008flight1tape1_1 (crop 2)		1416	4306	5721	47200	0,477	29	Intermediate, High	Urban military	Small	False	Oblique	
13	bike2	UAV123	553	1	553	18434	0,468	22	Intermediate	Urban, building	Small, Very small	True	Horizontal	
14	bike3		433	1	433	14434	0,462	19	Intermediate	Urban, building	Small	True	Horizontal	
15	building1		469	1	469	15634	0,454	12	Intermediate	Urban, building	Very Small	True	Horizontal	
16	building2		577	1	577	19234	0,471	37	Intermediate	Urban, building	Medium, Small	True	Horizontal	
17	building3		829	1	829	27634	0,451	27	High	Urban in desert	Small	True	Horizontal	
18	building4		787	1	787	26234	0,464	29	High, Intermediate	Urban in desert	None	True, False	Horizontal, Oblique	
19	car1		2629	1	2629	87634	0,471	59	Low, Intermediate	Road rural	Big, Medium	True	Oblique	
20	car11		337	1	337	11234	0,467	31	High	Suburban	Small	True, False	Horizontal, Oblique	
21	car12		499	1	499	16634	0,467	39	Low	Road urban, sea	Medium, Small	True	Horizontal	
22	car13		415	1	415	13834	0,461	26	High	Urban	Very very small	False	Oblique, Vertical	
23	car14		1327	1	1327	44234	0,471	25	Low	Road suburban	Medium	False	Oblique	
24	car15		469	1	469	15634	0,471	18	Intermediate	Road towards urban	Small, Very small	True	Oblique	
25	car2		1321	1	1321	44034	0,464	24	Intermediate	Road rural	Medium	False	Oblique, Vertical	
26	car3		1717	1	1717	57234	0,467	27	Intermediate	Road rural	Medium	False	Oblique, Vertical	
27	car4		1345	1	1345	44834	0,462	23	Intermediate, Low	Road rural	Big	False	Oblique, Vertical	
28	car7		1033	1	1033	34434	0,464	18	Intermediate	Road suburban	Medium	False	Oblique	
29	car9		1879	1	1879	62634	0,470	23	Intermediate, Low	Road suburban	Medium	False, True	Oblique, Horizontal	
30	person22		199	1	199	6634	0,456	31	Low	Urban sea	Medium, Big	True	Horizontal	
31	truck2		601	1	601	20034	0,453	24	High	Urban road	Small	True	Horizontal	
32	truck3		535	1	535	17834	0,472	18	Intermediate	Road towards urban	Small, Very small	True	Oblique	
33	truck4		1261	1	1261	42034	0,466	17	Intermediate	Road towards urban	Small	True	Oblique, Horizontal	
34	wakeboard8		1543	1	1543	51434	0,472	39	Low	Sea urban	Medium, Big	True, False	Oblique, Vertical, Horizontal	
35	Basketball		DTB70	427	1	427	14234	0,477	48	Intermediate	Field suburban	Medium	True	Oblique
36	Girl1			218	1	218	7267	0,481	31	Low	Field suburban	Big	True	Horizontal
37	Girl2			626	1	626	20867	0,482	30	Low	Field suburban	Big	True	Horizontal
38	ManRunning1			619	1	619	20634	0,483	23	Low	Field suburban	Big	True	Horizontal, Oblique
39	ManRunning2			260	1	260	8667	0,484	27	Low	Field suburban	Very big	False	Vertical, Oblique
40	Soccer1			613	1	613	20434	0,476	57	Low, Intermediate	Field suburban	Very big, Big	True	Horizontal
41	Soccer2			233	1	233	7767	0,475	24	High	Field suburban	Small	True	Oblique
42	StreetBasketball1			241	1	241	8034	0,379	37	Low	Field urban	Big	True, False	Oblique, Vertical
43	Walking			395	1	395	13167	0,476	31	Low	Field suburban	Big, Very big	True	Oblique
	Average			982			33 sec	0,47	31,27					
	Standard deviation			727			24 sec	0,02	10,32					
	Overall			42241			1408 sec							

**Table 3.** Stimuli ID and name, their original dataset, number of frames together with starting and ending frame number, duration and native resolution.



### 222 3.3. Experimental design

223 To record the gaze deployment of subjects while viewing UAV video sequences displayed  
224 onscreen, it is required to define an experimental methodology. All the details are presented below.

#### 225 3.3.1. Eye-tracking apparatus

226 A specific setup is designed to capture eye-tracking information on video stimuli. It includes a  
227 rendering monitor, an eye-tracking system, a control operating system, and a controlled laboratory test  
228 room. Figure 2 illustrates the experimental setup used during the collection of gaze information. We  
229 can observe the arrangement of all the systems described below.

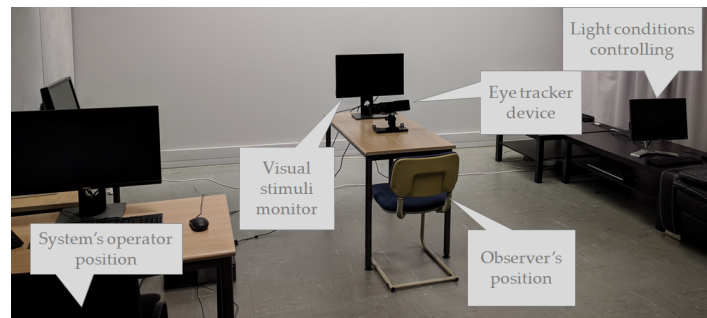


Figure 2. Experiment setup.

230 To run the experiment and collect gaze information, we used the EyeLink® 1000 Plus eye-tracking  
231 system<sup>2</sup>, in the head free-to-move *remote mode*, taking advantage of its embedded 25mm camera lens.  
232 The eye tracker principle is to detect and record the IR illuminator reflection rays on the observer's  
233 pupil [64]. This system enables the collection of highly precise gaze data at a temporal frequency of  
234 1000 Hz and a spatial accuracy between the visual angle range of 0.25 and 0.50 degree, according to  
235 the manufacturer. The eye tracker's camera was configured for each subject, without affecting the  
236 corresponding distance between them. This configuration guarantees to achieve an optimal detection  
237 of the observer's eyes and head sticker.

238 The experimental monitor which displayed stimuli was a 23.8 inches (52.70 x 29.65 cm) DELL  
239 P2417H computer monitor display<sup>3</sup> with full HD resolution (1920x1080) at 60 Hz and with a response  
240 time of 6 ms. As suggested by both the International Telecommunication Union (ITU)-Broadcasting  
241 service (Television) (BT).710 [66] and manufacturer, observers sited in distance of about 3H (1m ±  
242 10cm) from the monitor, where H corresponds to the stimuli display height so that observers have an  
243 assumed spatial visual angle acuity of one degree. Moreover, the eye tracker camera was placed 43 cm  
244 away from the experimental display, and thus about 67 cm from participants. Based on this setting,  
245 there are 64 pixels per degree of visual angle in each dimension, and the display resolution is about  
246 30x17 visual degrees.

247 Regarding software, the MPC-HC video player<sup>4</sup>, considered as one of the most lightweight  
248 open-source video players, rendered the experimental video stimuli. Also, we took advantage of the  
249 EyeLink toolbox [67] as it is part of the third version of Psychophysics Toolbox Psychtoolbox-3 (PTB-3)  
250<sup>5</sup> and added in-house communication processes<sup>6</sup> for sync between control and display systems. The  
251 control system consists of an additional computer, used by the experimenter to configure and control  
252 the eye-tracking system with an Ethernet connection.

<sup>2</sup> <https://www.sr-research.com/eyelink-1000-plus/>

<sup>3</sup> <https://www.dell.com/cd/business/p/dell-p2417h-monitor/pd>

<sup>4</sup> <https://mpc-hc.org/>

<sup>5</sup> <http://psychtoolbox.org/>

<sup>6</sup> LS2N, University of Nantes

253 Eventually, eye-tracking tests were performed in a room with controlled constant light conditions.  
254 The performed calibration set the constant ambient light conditions at approximately  $36.5 \text{ cd/m}^2$ , i.e.  
255 15% of the maximum stimuli monitor brightness -  $249 \text{ cm/m}^2$  - as recommended by the ITU-BT.500 [68],  
256 with the i1 Display Pro X-Rite® system.

### 257 3.3.2. Stimuli presentation

258 The random presentation of stimuli in their native resolution centered on the screen prevents  
259 ordering, resizing, and locating biases. Knowing that the monitor resolution is higher than that of  
260 selected sequences, video stimuli were padded with mid-grey. Additionally, to avoid possible biases in  
261 gaze allocation, a 2-second sequence of mid-gray frames was presented before playing a test sequence.  
262 Please note that the amount of original information contained in a degree of visual angle is not the  
263 same for VIRAT sequences than for other database content, as specified in Table 1.

264 Before starting the experiment, a training session is organized to get the subject familiar with the  
265 experiment design. It includes a calibration procedure and its validation followed by the visualization  
266 of one video. This UAV video is the sequence *car4* from the DTB70 dataset. To avoid any memory bias,  
267 this sequence is not included in test stimuli. Once subjects completed the training session, they could  
268 ask questions to experimenters before taking part into test sessions.

269 Regarding test sessions, they start with calibration and its validation. Follows then the  
270 visualization of 9 videos during which subjects do or do not perform a task. To ensure the optimal  
271 quality of the collected gaze data, each participant took part in five test sessions. Splitting the  
272 experiment into sessions decreases the tiredness and lack of attention in observers. Also, this design  
273 enables frequent calibration so that recordings do not suffer from the decrease of accuracy in gaze  
274 recordings with time [64].

275 With regards to calibration, the eye-tracking system is calibrated for each participant. The  
276 calibration reaches validation when the overall deviation of both eye positions is approximately below  
277 the fovea vision accuracy (e.g. a degree of visual angle [64,69]). The calibration procedure is repeated  
278 until validation.

279 The participation of an observer in the experiment lasts about 50 minutes. It includes test  
280 explanations, forms signing, and taking part in the training and the five test sessions. This duration is  
281 acceptable regarding the number of sessions and the fatigue in subjects.

### 282 3.3.3. Visual tasks to perform

283 *EyeTrackUAV2* aims to investigate two visual tasks. Indeed, we want to be able to witness visual  
284 attention processes triggered by top-down (or goal-directed) and bottom-up (or stimulus-driven)  
285 attention. Accordingly, we defined two visual tasks participants have to perform: the first condition is  
286 a Free Viewing (FV) task while the second relates to a surveillance-viewing Task (Task). The former  
287 task is rather common in eye-tracking tests [31,36,42,44,49,70]. Observers were simply asked to observe  
288 visual video stimuli without performing any task. For the surveillance-viewing task, participants  
289 were required to watch video stimuli and to push a specific button on a keyboard each time they  
290 observe a new - meaning not presented before - moving object (e.g. people, vehicle, bike, etc.) in  
291 the video. The purpose of this task is to simulate one of the basic surveillance procedures in which  
292 targets could be located anywhere when the visual search process was performed [71]. After reviewing  
293 typical surveillance systems' abilities [72], we have decided to define our task as object detection. The  
294 defined object detection task is compelling in that it encompasses target-specific training (repeated  
295 discrimination of targets and non-targets) and visual search scanning (targets potentially located  
296 anywhere) [71]. The surveillance-viewing task is especially interesting for a military context, in which  
297 operators have to detect anomaly in drone videos.

### 3.3.4. Population

Overall, 30 observers participated in each phase of the test. Tested population samples were different for these two viewing conditions. They were carefully selected to be as diverse as possible. For instance, they include people from more than 12 different countries, namely Algeria (3%), Brazil, Burundi, China, Colombia (10%), France (67%), Gabon, Guinea, South Arabia, Spain, Tunisia, and Ukraine. Additionally, we achieved gender and almost eye-dominance balance in both phases tests. Table 4 presents the detailed population characteristics for both tasks.

Each observer has been tested for visual acuity and color vision with Ishihara and Snellen tests [73,74]. Any failure to these tests motivated the dismissal of the person from the experiment. Before running the test, the experimenter provided subjects with written consent and information forms, together with oral instructions. This process made sure of the consent of participants and their understanding of the experiment process. It also ensures an anonymous data collection.

Sample statistics	FV	Task	Total
Participants	30	30	60
Female	16	16	32
Male	14	14	28
Average age	31,7	27,9	29,8
Std age	11,0	8,5	10,0
Min age	20	19	19
Max age	59	55	59
Left dominant eye	19	9	28
Right dominant eye	11	21	32
Participants with glasses	0	4	4

**Table 4.** Population characteristics.



**Figure 3.** Stimulus displayed in its native resolution, and padded with mid-gray to be centered. Colored information relates to Equation 1.

### 3.4. Post-processing of eye-tracking data

First, we transform collected raw signals into the pixel coordinate system of the original sequence. This conversion leads to what we refer to as binocular gaze data. Let us precise that the origin of coordinates is the top-left corner. Then, any gaze coordinates out of range are evicted, as they do not represent visual attention on stimuli. Once transformed and filtered, we extract fixation and saccade information and create gaze density maps from gaze data. The remainder of this section describes all post-processing functions.

#### 3.4.1. Raw data

At first, coordinates of the collected binocular gaze data were transformed into the pixel coordinate system of the visual stimulus. Additionally, we addressed the original resolution of sequences. Coordinates outside the boundaries of the original resolution of the stimulus were filtered out as they were not located in the video stimuli display area. The following formula presents how the collected coordinates are transformed for both eyes:

$$\begin{cases} x_S = \lfloor x_D - \frac{R_D^X - R_S^X}{2} \rfloor \\ y_S = \lfloor y_D - \frac{R_D^Y - R_S^Y}{2} \rfloor \end{cases} \quad (1)$$

where,  $(x_S, y_S)$  and  $(x_D, y_D)$  are the spatial coordinates on the stimulus and on the display, respectively. The operator  $\lfloor \cdot \rfloor$  allows to keep the coordinates if the coordinates are within the frame of the stimulus. Otherwise, the coordinate is discarded.  $(R_S^X, R_S^Y)$  and  $(R_D^X, R_D^Y)$  represent the stimulus resolution and the display resolution, respectively. For more clarity, Figure 3 displays the terms of the equation. Once

324 this remapping has been done for both eyes, the spatial binocular coordinates is simply given by the  
325 average of the spatial coordinates of left and right eyes.

326 During the surveillance-viewing task, subjects pushed a button when detecting an object (never  
327 seen before) in the content. Triggering this button action must be included in raw data. Consequently,  
328 we denote in raw data a button activation (respectively no detection reaction) with the Boolean value 1  
329 (respectively 0). Besides, for convenience, we have extracted the positions of the observer's dominant  
330 eyes and included them in raw gaze data.

### 331 3.4.2. Fixation and saccade event detection

332 To retrieve fixations from eye positions, we used the Dispersion-Threshold Identification (I-DT)  
333 [75] from the EyeMMV and LandRate toolboxes [76,77]. This algorithm performs "two-step" spatial  
334 and temporal thresholds. As exposed in [77,78], thanks to the very high precision of our eye-tracking  
335 equipment, we can combine the two-step spatial thresholds in one operation, as both thresholds have  
336 the same value. Ultimately, in our context, this algorithm conceptually implements a spatial noise  
337 removal filter and a temporal threshold indicating the minimum fixation duration. We have selected  
338 the minimum threshold values from the state of the art to ensure the performance of the fixation  
339 detection algorithm. Accordingly, spatial and temporal thresholds were selected to be equal to 0.7  
340 degree of visual angle and 80 ms [79], respectively. Finally, saccade events were calculated based on the  
341 computed fixations considering that a saccade corresponds to eye movements between two successive  
342 fixation points. When considering raw data of the dominant eye, I-DT exhibits a total number of  
343 fixations of 1 239 157 in FV and 1 269 433 in Task.

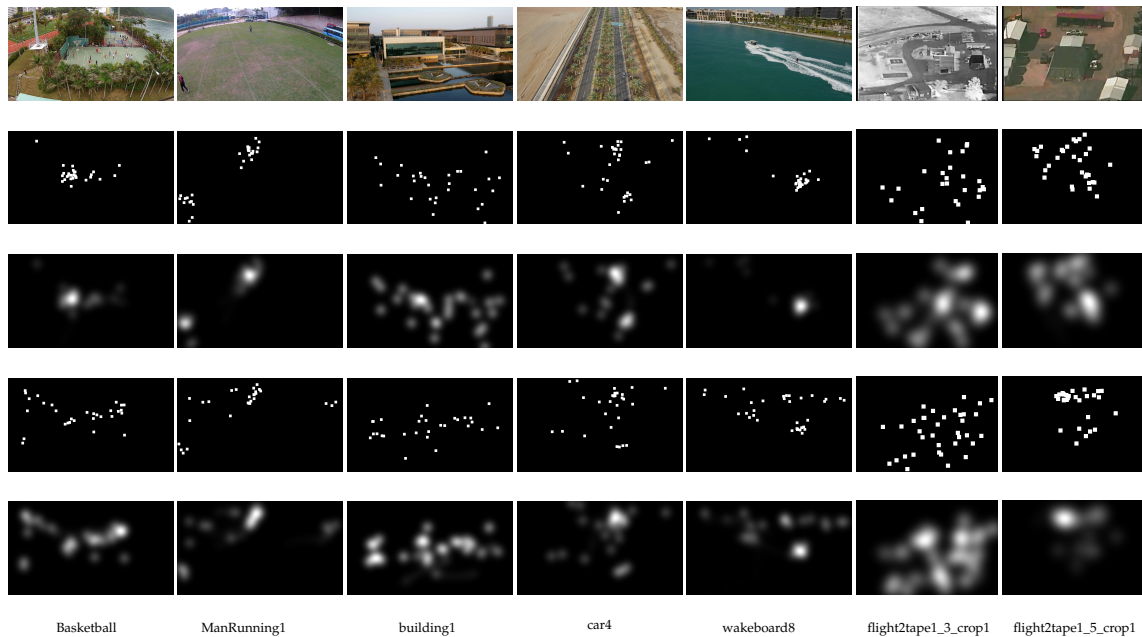
### 344 3.4.3. Human saliency maps

345 Saliency maps are a 2D topographic representation indicating the ability of an area to attract  
346 observers' attention. It is common to represent the salience of an image thanks to either its saliency map  
347 or by its colored representation, called heatmap. Saliency maps are usually computed by convolving  
348 the fixation map, gathering observers' fixations, with a Gaussian kernel representing the foveal part of  
349 our retina. More details can be found in [70]. In our context, we did not compute convolved fixation  
350 maps. We took benefit from the high frequency of acquisition of the eye-tracker system to compute  
351 saliency maps directly from raw gaze data (in pixel coordinates). For the sake of clarity, we from now  
352 will refer to the generated saliency maps as gaze density maps.

353 To represent salient regions of each frame, we followed the method described in [76]. We derived  
354 parameters from the experimental setup (e.g., a grid size of a pixel, a standard deviation of 0.5 degree  
355 of angle i.e.  $\sigma = 32$  pixels, and a kernel size of  $6\sigma$ ). For visualization purposes, gaze density maps  
356 were normalized between 0 and 255. Figure 4 presents gaze density maps obtained for both attention  
357 conditions in frame 100 of seven sequences. We have selected frame 100 to get free from the initial  
358 center-bias in video exploration occurring during the first seconds. These examples illustrate the  
359 sparsity of salience in videos in free viewing, while task-based attention usually presents more salient  
360 points, more dispersed in the content than FV, depending on the task and attention-grabbing objects.

### 361 3.5. *EyeTrackUAV2* in brief

362 We have created a dataset containing binocular gaze information collected during two viewing  
363 conditions ( free viewing and task) over 43 UAV videos (30 fps, 1280x720 and 720x480 - 42241 frames,  
364 1408 seconds) observed by 30 participants per condition, leading to 1 239 157 fixations in free-viewing  
365 and 1 269 433 in task-viewing for dominant eyes positions. Notably, selected UAV videos show  
366 diversity in rendered environments, movement and size of objects, aircraft flight heights and angles to  
367 the ground, duration, size, and quality. This dataset overcomes the limitations of *EyeTrackUAV1* in that  
368 it enables investigations of salience in more test sequences, on larger population samples, and for both  
369 free-viewing and task-based attention. Additionally, and even though they are still too few, three IR  
370 videos are part of visual stimuli.



**Figure 4.** Frame 100 of seven sequences of *EyeTrackUAV2* dataset, together with gaze density and fixation maps generated based on gaze data of dominant eye. Results are presented for both types of attention. The first row presents sequences hundredth frame, the second fixations for FV, the third gaze density maps for FV, the fourth fixations for task, and the fifth gaze density maps for Task.

371 Fixations, saccades, and gaze density maps were computed - for both eyes in additive and  
 372 averaged fashions (see Binocular and BothEyes scenarios described later) and for the dominant eye -  
 373 and are publicly available with original content and raw data on our FTP <sup>7</sup>. The code in MATLAB to  
 374 generate all ground truth information is also made available.

#### 375 4. Analyses

376 In this section, we characterize the proposed *EyeTrackUAV2* database. On one hand, we compare  
 377 saliency between six ground truth generation scenarios. This study can be beneficial to the community  
 378 to know what is the potential error made when selecting a specific ground truth scenario over another.  
 379 On the other hand, UAV videos induce new visual experiences. Consequently, observers exhibit  
 380 different behaviors towards this type of stimuli. Therefore, we investigate whether the center bias, one  
 381 of the main viewing tendencies [27], still applies to *EyeTrackUAV2* content.

##### 382 4.1. Six different ground truths

383 The first question we address concerns the method used to determine the ground truth. In a  
 384 number of papers, researchers use the ocular dominance theory in order to generate the ground truth.  
 385 This theory relies on the fact that the human visual system favors the input of one eye over the other  
 386 should binocular images be too disparate on the retinas. However, the cyclopean theory gains more  
 387 and more momentum [80,81]. It alleges that vision processes approximate a central point between  
 388 two eyes, from which an object is perceived. Furthermore, lately, manufacturers achieved major  
 389 improvements in eye-tracking systems. They are now able to record and calibrate the positions of  
 390 both eyes separately. This allows for exploring what are the best practices to create saliency ground  
 391 truth [80–82].

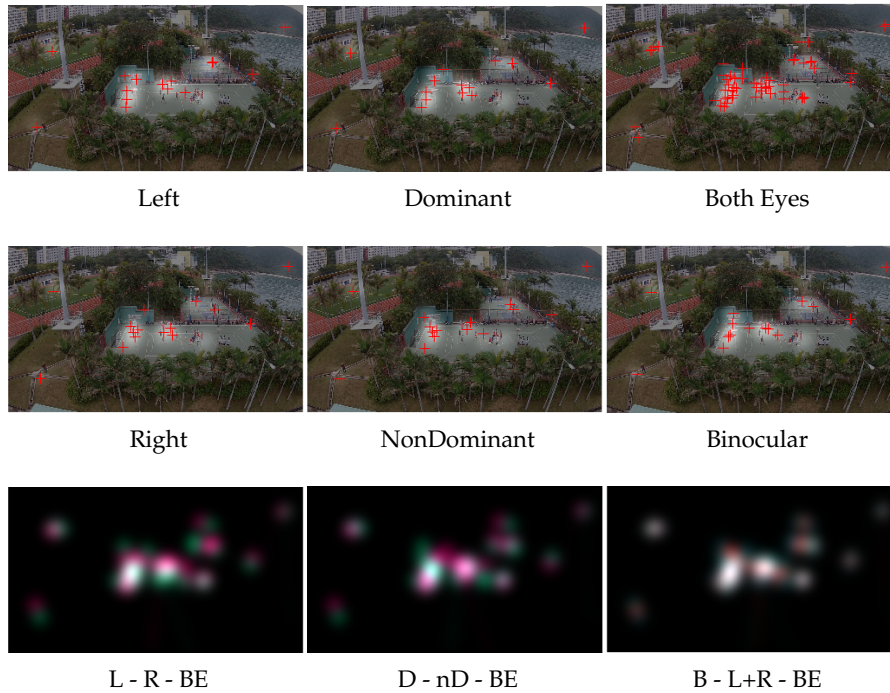
<sup>7</sup> <ftp://dissocie@ftp.ivc.polytech.univ-nantes.fr/EyeTrackUAV2/>

392 When examining the mean of absolute error between eye positions of all scenarios, we have found  
 393 a maximum value of about 0.6 degrees of visual angle. That value is rather small compared to the  
 394 Gaussian kernel convolved on eye positions. Thus, we question whether selecting a ground truth  
 395 scenario over another makes a significant difference for saliency studies. Consequently, we compare  
 396 gaze density maps generated for the six scenarios defined below.

397 We propose to evaluate the potential errors made when different methods for creating the ground  
 398 truth are used. Note that the true position of the user gaze is not available. Accordingly, we need to run  
 399 a cross-comparison between several well-selected and representative ground truths. Scenarios highly  
 400 similar to all others are the ones that will make fewer errors. In such a context, the more scenarios are  
 401 included, the more complete and reliable the study is.

402 We tested six methods, namely Left (L), Right (R), Binocular (B), Dominant (D), non Dominant  
 403 (nD), and Both Eyes (BE). B corresponds to the average position between the left and right eyes and  
 404 can be called version signal (see Equation 2). BE includes the positions of both L and R eyes, and  
 405 hence comprises twice more information than other scenarios (see Equation 3). nD has been added to  
 406 estimate the gain made when using dominant eye information. The two visual attention conditions  
 407 Free Viewing (FV) and surveillance-viewing Task (Task) were examined in this study. Illustrations of  
 408 scenarios gaze density maps and fixations as well as methods comparisons are presented in Figure 5.  
 409 Below is presented the quantitative evaluation.

$$410 \quad \begin{cases} x_B = \lfloor \frac{x_L + x_R}{2} \rfloor \\ y_B = \lfloor \frac{y_L + y_R}{2} \rfloor \end{cases} \quad (2) \quad \begin{cases} x_{BE} = x_L \cup x_R \\ y_{BE} = y_L \cup y_R \end{cases} \quad (3)$$



**Figure 5.** Qualitative comparison of gaze density maps for all scenarios on Basketball, frame 401. gaze density and fixations are displayed in transparency over the content. The last row compares scenarios: first scenario is attributed to the red channel, the second to green and the last to blue. When fully overlapping, the pixel turns white.

411 We ran a cross-comparison on six well-used saliency metrics: Correlation Coefficient (CC) ( $\in [-1, 1]$ ),  
 412 Similarity (SIM) ( $\in [0, 1]$ ) the intersection between histograms of saliency, AUC Judd and Borji ( $\in [0, 1]$ ),  
 413 NSS ( $\in ]-\infty, +\infty[$ ), and IG ( $\in [0, +\infty[$ ), which measures on average the gain in information contained in the  
 414 saliency map compared to a prior baseline ( $\in [0, +\infty[$ ). We did not report Kullback Leibler divergence (KL)  
 415 ( $\in [0, +\infty[$ ) as we favored symmetric metrics. Moreover, even though symmetric in absolute value, IG

SM1	SM2	FV				Task			
		CC ↑	SIM ↑	IG ↓	SM1-Fix1-SM2	SM2-Fix2-SM1	CC ↑	SIM ↑	IG ↓
Binocular	Dominant	0,94	0,83	0,377	0,300	0,952	0,850	0,276	0,148
Binocular	EyeNonDom	0,95	0,84	0,370	0,301	0,952	0,849	0,283	0,163
Binocular	Left	0,94	0,83	0,371	0,301	0,948	0,843	0,264	0,192
Binocular	Right	0,94	0,83	0,390	0,324	0,944	0,838	0,304	0,152
Binocular	BothEyes	<b>0,98</b>	<b>0,90</b>	0,246	<b>0,139</b>	<b>0,983</b>	<b>0,916</b>	0,177	<b>0,012</b>
Dominant	BothEyes	0,96	0,87	<b>0,158</b>	0,374	0,967	0,873	<b>0,143</b>	0,248
EyeNonDom	BothEyes	0,97	0,87	0,167	0,394	0,966	0,872	0,144	0,228
Left	BothEyes	0,96	0,86	0,166	0,387	0,960	0,861	0,174	0,232
Right	BothEyes	0,96	0,86	0,181	0,416	0,960	0,862	0,147	0,279
Dominant	EyeNonDom	0,87	0,74	1,115	1,069	0,873	0,747	0,743	0,781
Dominant	Left	0,95	0,88	0,341	0,339	0,903	0,792	0,520	0,582
Dominant	Right	0,91	0,79	0,810	0,757	0,957	0,884	0,256	0,233
EyeNonDom	Right	0,96	0,88	0,346	0,342	0,902	0,793	0,587	0,519
Left	EyeNonDom	0,91	0,79	0,792	0,754	0,957	0,884	0,256	0,231
Left	Right	<b>0,85</b>	<b>0,72</b>	<b>1,176</b>	<b>1,121</b>	<b>0,850</b>	<b>0,725</b>	<b>0,877</b>	<b>0,782</b>
Mean		0,937	0,832	0,467	0,488	0,938	0,839	0,343	0,319
Std		0,037	0,052	0,340	0,295	0,038	0,053	0,230	0,234

**Table 5.** CC, SIM and IG results for scenarios cross-comparison. Red indicates the best scores, blue the least.

		FV		Task	
		F-value	p-value	F-value	p-value
$p < 0.05$	CC	F(14,630) = 77.72	5e-127	F(14,630) = 172.55	9e-205
	SIM	F(14,630) = 200.07	5e-221	F(14,630) = 309.43	2e-271
	IG	F(14,630) = 158.96	5e-196	F(14,630) = 156.16	4e-194
$p > 0.05$	AUCJ	F(14,630) = 0.36	0.9857	F(14,630) = 0.4	0.9742
	AUCB	F(14,630) = 0.22	0.9989	F(14,630) = 0.05	1
	NSS	F(14,630) = 0.95	0.5036	F(14,630) = 0.92	0.5344

**Table 6.** ANOVA analysis.

416 provides different scores depending on fixations under consideration. We thus compared scenarios  
 417 for fixations of both methods, which leads to two IG measures. More details on metrics and metrics  
 418 behaviors are given in [35,70,83]. Table 5 presents the results of measures when comparing gaze  
 419 density maps of two scenarios.

420 Here are some insights extracted from the results:

- 421 • There is a high similarity between scenario gaze density maps. As expected, scores are pretty  
 422 high (respectively low for IG), which indicates the high similarity between scenarios.
- 423 • All metrics show the best results for comparisons including Binocular and BothEyes scenarios,  
 424 the highest being the Binocular-BothEyes comparison.
- 425 • Left-Right and Dominant-NonDominant comparisons achieve worst results.
- 426 • It is possible to know the population main dominant eye through scenarios comparisons (not  
 427 including two eyes information). When describing the population, we have seen that a majority  
 428 of left-dominant-eye subjects participated in the FV test, while the reverse happened for the Task  
 429 experiment. This fact is noticeable in metric scores.

430 To verify whether scenarios present statistically significant differences, we have conducted an  
 431 ANalysis Of VAriance (ANOVA) on the scores obtained by the metrics. ANOVA results are presented  
 432 in Table 6. All metrics show statistically different results ( $p < 0.05$ ) except for AUC Borji, AUC Judd,  
 433 and NSS. It shows that, with regard to these three metrics, using a scenario over another makes no  
 434 significant difference. This also explains why we did not report AUC and NNS results in Table 5. We  
 435 explore further the other metrics, namely CC, SIM, and IG, through multi-comparison analyses.

436 Results are presented in Figure 6. On the charts, we can see where stands mean and standard  
 437 deviation of metric scores for each scenario over the entire dataset. Scenarios having non-overlapping  
 438 confidence intervals are statistically different. This study has been conducted on four metrics, namely

439 CC, SIM and the two variants of IG. Results confirm the previously mentioned insights. Moreover, a  
 440 statistical difference is observed between Left-Right and Dominant-NonDominant comparisons under  
 441 the task condition.

442 Overall, over six metrics, three do not find significant differences between the scenarios' gaze  
 443 density maps. The four others do and indicate that using both eye information can be encouraged.  
 444 Then, if not possible, eye-dominance-based signals may be favored over left and right eye scenarios, in  
 445 particular under task-based attention. We stress out that overall, the difference between scenarios is  
 446 rather small, as three metrics could not differentiate them.

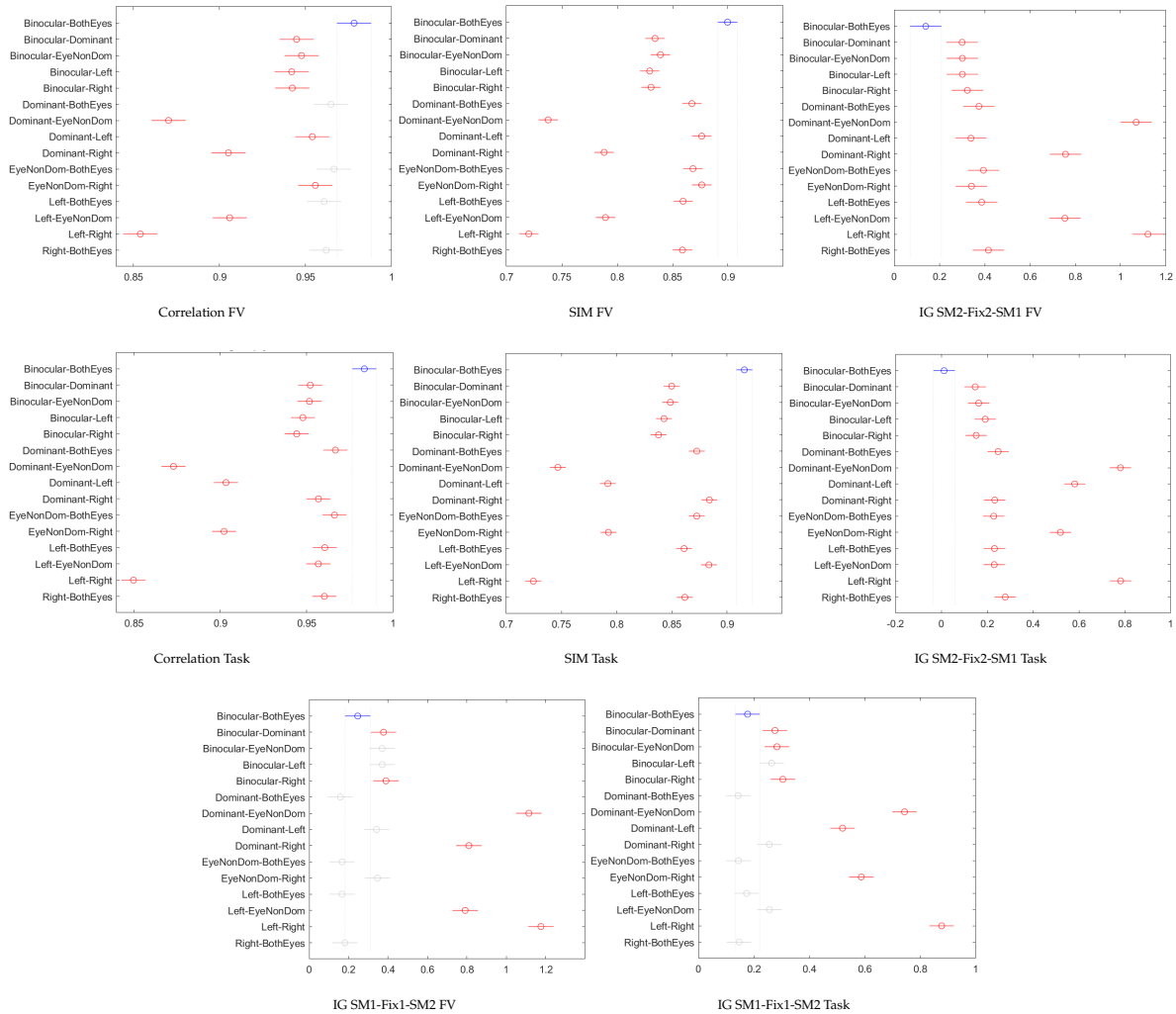


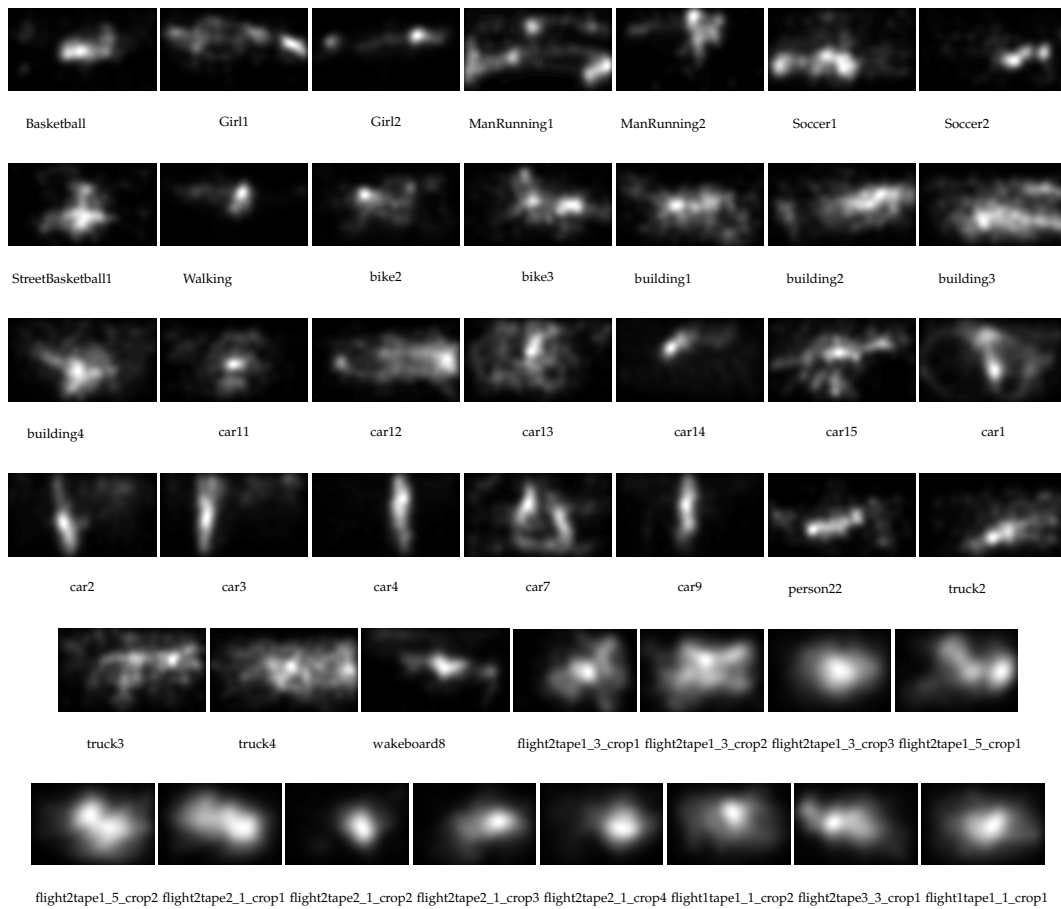
Figure 6. Multi-comparison on scenarios correlation measure.

## 447 4.2. Biases in UAV videos

### 448 4.2.1. Qualitative evaluation of biases in UAV videos

449 The importance of the center bias in visual saliency for conventional imaging has been shown  
 450 in Section 2. We wondered whether the center bias is systematically present in UAV content. This  
 451 section aims to shed light on this question. We evaluate the viewing tendencies of observers thanks to  
 452 the average gaze density map, computed over the entire sequence. It is representative of the average  
 453 position of gaze throughout the video. It is used to observe potential overall biases, as it could be the  
 454 case with the center bias. Figures 7 and 8 show the average gaze density map for all sequences of





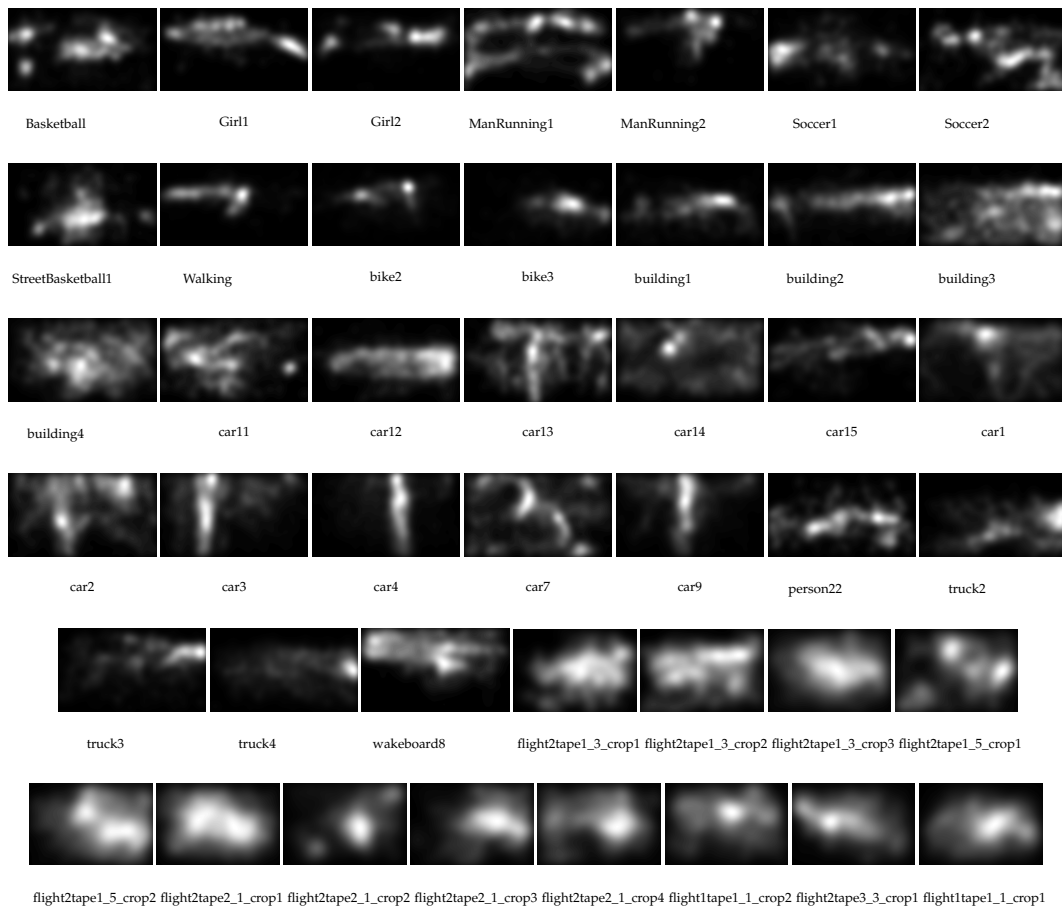
**Figure 7.** Average gaze density maps for all sequences of *EyeTrackUAV2* dataset, generated from D scenario, for the free-viewing condition.

455 *EyeTrackUAV2* dataset, generated from D scenario, for both free-viewing and task-viewing conditions.  
 456 Several observations can be made.

457 **Content-dependent center bias.** We verify here the content-dependence of the center bias in  
 458 UAV videos. For both attention conditions, the scene environment and movements exacerbates or not  
 459 UAV biases. For instance, in sequences *car 2-9* (fourth row), the aircraft is following cars on a road.  
 460 Associated average gaze density maps display the shape of the road and its direction, i.e. vertical  
 461 route for all and roundabout for *car7*. *Car 14* (third row), a semantically similar content except that it  
 462 displays only one object on the road with a constant reframing (camera movement) which keeps the  
 463 car at the same location, presents an average gaze density map centered on the tracked object.

464 **Original database-specific center bias.** We can observe that a center bias is present in VIRAT  
 465 sequences, while videos from other datasets, namely UAV123 and DTB70, do not present this bias  
 466 systematically. The original resolution of content and the experimental setup are possibly the sources  
 467 of this result. Indeed, the proportion of content seen at once is not the same for all sequences: 1,19%  
 468 of a VIRAT content is seen per degree of visual angle, whereas it is 0,44% for the two other original  
 469 databases. VIRAT gaze density maps are thus smoother, which results in higher chances to present a  
 470 center bias. To verify this assumption based on qualitative assessment, we have computed the overall  
 471 gaze density maps for sequences coming from each original dataset, namely DTB70, UAV123 and  
 472 VIRAT. These maps are shown in Figure 9. VIRAT gaze density maps are much more concentrated and  
 473 centered. This corroborates that biases can be original-database-specific.

474 **Task-related gaze density maps seem more spread out.** Task-based gaze density maps cover  
 475 more content when compared to free-viewing condition for most sequences (e.g. in about 58% of



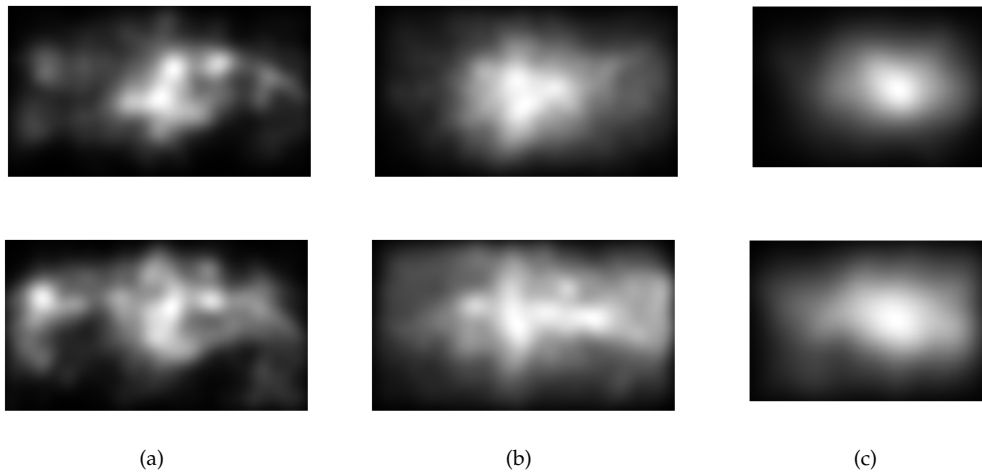
**Figure 8.** Average gaze density maps for all sequences of *EyeTrackUAV2* dataset, generated from D scenario, for the task-viewing condition.

476 videos such as *Basketball*, *car11*, *car2*, and *wakeboard*). This behavior is also illustrated in Figure 9. We  
 477 correlate this response with the object detection task. Visual search scanning implies an extensive  
 478 exploration of the content. However, 21% of the remaining sequences (i.e. *soccer1*, *bike2-3*, *building*  
 479 *1-2*, *car1,15*, and *truck3-4*) show less discrepancies in the task-viewing condition than in free-viewing  
 480 condition. We do not find correlation between such behavior and sequences characteristics given in  
 481 Table 3. This leaves room for further exploration of differences between task-based and free viewing  
 482 attention.

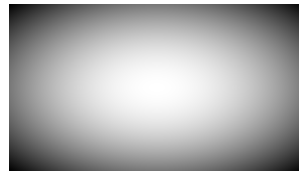
483 **Overall, there is no generalization of center bias for UAV content.** As stated earlier, we do not  
 484 observe a systematic center bias, except for VIRAT sequences. This is especially true for task-related  
 485 viewing. However, we observe specific patterns. Indeed, vertical and horizontal potatoe-shaped  
 486 salient areas are quite present in average gaze density maps of *EyeTrackUAV2*. Such patterns are also  
 487 visible in UAV2 and DTB70 overall gaze density maps, especially in task-viewing condition. This  
 488 indicates future axes of developments for UAV saliency-based applications. For instance, instead of  
 489 using a center bias, one may introduce priors as a set of prevalent saliency area shapes with different  
 490 directions and sizes [84].

#### 491 4.2.2. Quantitative evaluation of the central bias in UAV videos

492 To go further into content-dependencies, we investigate quantitatively the similarity of  
 493 dominant-eye-generated gaze density maps with a pre-defined center bias. Figure 10 presents the  
 494 center bias baseline created in this purpose as suggested in [34,35].



**Figure 9.** Overall average gaze density maps per original dataset, generated from D scenario, in free-viewing (top-row) and Task-viewing (bottom row) for original datasets: (a) DTB70; (b) UAV123; (c) VIRAT.



**Figure 10.** Center prior baseline.

495 We performed the evaluation based on four well-used saliency metrics: CC, SIM, KL, and IG.  
 496 Results are presented in Table 7. They support the observations we made in the previous section.  
 497 Overall scores do not reveal a high similarity with the center prior (e.g. maximum CC and SIM of  
 498 about 0.5, high KL and IG). On the other hand, we observe content-specific center prior in UAV123  
 499 and DTB70. For instance, videos more prone to center bias includes sequences extracted from VIRAT  
 500 and *building1,3,4*, and *car13*. On the contrary, sequences *Girl1-2*, *ManRunning1-2*, *Walking*, *car4*, and  
 501 *wakeboard8* are not likely to present center bias. This confirms there is no generalization of center bias  
 502 for UAV content. Regarding differences between free-viewing and task-viewing conditions, results are  
 503 inconclusive as no systematic behavior is clearly visible from this analysis.

## 504 5. Conclusion

505 UAV imaging modifies the perceptual clues of typical scenes due to its bird point of view, the  
 506 presence of camera movements and the high distance and angle to the scene. For instance, low-level  
 507 visual features, and size of objects change and depth information is flattened or disappears (e.g.  
 508 presence of sky). To understand observers' behaviors toward these new features, especially in terms of  
 509 visual attention and deployment, there is a need for large-scale eye-tracking databases for saliency in  
 510 UAV videos. This dataset is also a key factor in the field of computational models of visual attention,  
 511 in which large scale datasets are required to train the latest generation of deep-based models.

512 This need is even stronger with the fast expansion of applications related to UAVs, for leisure and  
 513 professional civilian activities and a wide range of military services. Combining UAV imagery with  
 514 one of the most dynamic research fields in vision, namely saliency, is highly promising, especially for  
 515 videos that are gaining more and more attention these last years.

516 This work addresses the need for such a dedicated dataset. An experimental process has  
 517 been designed in order to build a new dataset, *EyeTrackUAV2*. Gaze data were collected during  
 518 the observation of UAV videos under controlled laboratory conditions for both free viewing and

	FV				Task			
	CC ↑	SIM ↑	KL ↓	IG ↓	CC ↑	SIM ↑	KL ↓	IG ↓
VIRAT_09152008flight2tape1_3_crop1	<b>0,50</b>	<b>0,48</b>	<b>7,17</b>	<b>1,53</b>	<b>0,46</b>	<b>0,48</b>	<b>6,85</b>	<b>1,62</b>
VIRAT_09152008flight2tape1_3_crop2	<b>0,49</b>	<b>0,52</b>	<b>5,59</b>	<b>1,50</b>	<b>0,36</b>	<b>0,48</b>	<b>6,42</b>	<b>1,75</b>
VIRAT_09152008flight2tape1_3_crop3	<b>0,46</b>	<b>0,43</b>	<b>8,46</b>	<b>1,91</b>	<b>0,37</b>	<b>0,43</b>	<b>7,98</b>	<b>1,99</b>
VIRAT_09152008flight2tape1_5_crop1	0,27	<b>0,38</b>	<b>9,77</b>	<b>2,29</b>	0,18	<b>0,36</b>	<b>10,14</b>	<b>2,49</b>
VIRAT_09152008flight2tape1_5_crop2	<b>0,42</b>	<b>0,44</b>	<b>8,05</b>	<b>1,90</b>	<b>0,30</b>	<b>0,45</b>	<b>7,41</b>	<b>1,87</b>
VIRAT_09152008flight2tape2_1_crop1	<b>0,41</b>	<b>0,39</b>	<b>9,34</b>	<b>2,05</b>	<b>0,38</b>	<b>0,42</b>	<b>8,55</b>	<b>1,97</b>
VIRAT_09152008flight2tape2_1_crop2	<b>0,40</b>	<b>0,35</b>	<b>10,90</b>	<b>2,50</b>	<b>0,32</b>	<b>0,42</b>	<b>8,01</b>	<b>2,01</b>
VIRAT_09152008flight2tape2_1_crop3	<b>0,42</b>	<b>0,40</b>	<b>9,46</b>	<b>2,11</b>	<b>0,28</b>	<b>0,39</b>	<b>9,30</b>	<b>2,24</b>
VIRAT_09152008flight2tape2_1_crop4	<b>0,36</b>	<b>0,36</b>	<b>10,35</b>	<b>2,34</b>	<b>0,28</b>	<b>0,38</b>	<b>9,79</b>	<b>2,30</b>
VIRAT_09152008flight2tape3_3_crop1	<b>0,42</b>	<b>0,43</b>	<b>8,16</b>	<b>1,96</b>	<b>0,35</b>	<b>0,43</b>	<b>7,84</b>	<b>2,03</b>
VIRAT_09162008flight1tape1_1_crop1	<b>0,47</b>	<b>0,45</b>	<b>7,76</b>	<b>1,80</b>	<b>0,37</b>	<b>0,42</b>	<b>8,40</b>	<b>2,00</b>
VIRAT_09162008flight1tape1_1_crop2	<b>0,40</b>	<b>0,40</b>	<b>9,14</b>	<b>2,14</b>	<b>0,27</b>	<b>0,40</b>	<b>8,91</b>	<b>2,22</b>
UAV123_bike2	<b>0,39</b>	<b>0,34</b>	<b>11,51</b>	<b>2,43</b>	<b>0,34</b>	0,29	13,21	2,82
UAV123_bike3	<b>0,39</b>	<b>0,34</b>	<b>11,71</b>	<b>2,37</b>	<b>0,29</b>	0,26	14,34	2,96
UAV123_building1	<b>0,40</b>	<b>0,37</b>	<b>10,64</b>	<b>2,18</b>	<b>0,32</b>	0,31	12,74	<b>2,69</b>
UAV123_building2	0,30	<b>0,33</b>	<b>11,89</b>	<b>2,43</b>	0,18	0,27	13,87	3,06
UAV123_building3	0,27	<b>0,34</b>	<b>11,50</b>	<b>2,42</b>	0,17	<b>0,32</b>	<b>11,82</b>	<b>2,56</b>
UAV123_building4	<b>0,39</b>	<b>0,36</b>	<b>10,82</b>	<b>2,20</b>	<b>0,35</b>	<b>0,39</b>	<b>9,72</b>	<b>2,10</b>
UAV123_car11	<b>0,37</b>	<b>0,32</b>	12,37	<b>2,58</b>	0,21	0,30	12,68	<b>2,67</b>
UAV123_car12	0,21	0,28	13,35	2,80	<b>0,26</b>	0,29	13,12	2,69
UAV123_car13	0,30	<b>0,34</b>	<b>11,48</b>	<b>2,39</b>	0,20	<b>0,33</b>	<b>11,50</b>	<b>2,44</b>
UAV123_car14	0,20	0,25	14,47	3,16	0,12	0,31	12,28	2,71
UAV123_car15	<b>0,31</b>	<b>0,34</b>	<b>11,52</b>	<b>2,47</b>	0,10	0,30	12,70	2,81
UAV123_car1	0,21	0,26	14,33	3,10	0,13	0,30	12,61	2,77
UAV123_car2	0,22	0,27	13,91	3,02	0,13	0,30	12,68	2,80
UAV123_car3	0,16	0,24	14,77	3,19	0,14	0,28	13,39	2,93
UAV123_car4	0,22	0,20	16,27	3,55	0,20	0,24	14,76	3,23
UAV123_car7	0,22	0,23	15,11	3,16	0,11	0,28	13,13	2,92
UAV123_car9	0,26	0,23	15,41	3,27	0,21	0,28	13,69	2,86
UAV123_person22	<b>0,35</b>	0,31	12,44	<b>2,60</b>	<b>0,27</b>	0,31	12,45	2,68
UAV123_truck2	0,27	<b>0,32</b>	12,29	<b>2,56</b>	0,09	0,27	13,66	3,01
UAV123_truck3	0,27	<b>0,35</b>	<b>11,14</b>	<b>2,34</b>	0,12	0,31	12,23	2,73
UAV123_truck4	0,29	<b>0,36</b>	<b>10,71</b>	<b>2,34</b>	0,16	0,29	13,18	3,03
UAV123_wakeboard8	0,23	0,21	15,91	3,45	0,11	0,24	14,93	3,29
DTB70_Basketball	<b>0,38</b>	0,27	14,13	2,89	<b>0,30</b>	0,31	12,30	<b>2,59</b>
DTB70_Girl1	0,16	0,28	13,47	2,90	0,15	0,25	14,54	3,18
DTB70_Girl2	0,20	0,20	16,04	3,60	0,19	0,23	15,04	3,34
DTB70_ManRunning1	0,02	0,16	17,45	4,09	0,00	0,20	16,11	3,73
DTB70_ManRunning2	0,12	0,13	18,40	4,31	0,10	0,15	17,99	4,24
DTB70_Soccer1	0,21	0,26	14,23	3,04	0,17	0,26	14,03	3,18
DTB70_Soccer2	0,21	0,22	15,56	3,33	0,22	<b>0,32</b>	<b>11,86</b>	<b>2,69</b>
DTB70_StreetBasketball1	<b>0,33</b>	0,26	14,29	2,94	<b>0,28</b>	0,26	14,29	3,00
DTB70_Walking	0,29	0,20	16,14	3,51	<b>0,27</b>	0,22	15,81	3,51
mean	0,31	0,32	12,27	2,67	0,23	0,32	12,01	2,69

**Table 7.** Comparison of gaze density maps with the center bias presented in Figure 10. Are displayed in red the numbers over (or under for KL and IG) measures average, indicated in the last row.

519 object-detection surveillance task conditions. Gaze positions have been collected on 30 participants  
520 for each attention condition, on 43 UAV videos in 30 fps, 1280x720 or 720x480, consisting in 42 241  
521 frames and 1408 seconds. Overall, 1 239 157 fixations in free-viewing and 1 269 433 in task-viewing  
522 were extracted from the dominant eye positions. Test stimuli were carefully selected from three  
523 original datasets, i.e. UAV123, VIRAT, and DTB70, to be representative as much as possible of the UAV  
524 ecosystem. Accordingly, they present variations in terms of environments, camera movement, size of  
525 objects, aircraft flight heights and angles to the ground, video duration, resolution, and quality. Also,  
526 three sequences were recorded in infra-red.

527 The collected gaze data were analyzed and transformed into fixation and saccade eye movements  
528 using an I-DT based identification algorithm. Moreover, the eye-tracking system high frequency of  
529 acquisition enabled the production of gaze density maps for each experimental frame of the examined  
530 video stimuli directly from raw data. The dataset is publicly available and includes, for instance, raw  
531 binocular eye positions, fixation, and gaze density maps generated from the dominant eye and both  
532 eyes information.

533 We further characterized the dataset considering two different aspects. On one hand, six scenarios,  
534 namely binocular, both eyes, dominant eye, non-dominant eye, left, and right can be envisioned to  
535 generate gaze density maps. We wondered whether a scenario should be favored over another or  
536 not. Comparisons of scenarios have been conducted on six typical saliency metrics for gaze density  
537 maps. Overall, all scenarios are pretty similar: over the six evaluated metrics, three could not make a  
538 distinction between scenarios. The three last metrics present mild but statistically significant differences.  
539 Accordingly, the information of both eyes may be favored to study saliency. If not possible, choosing  
540 information from the dominant eye is encouraged. This advice is not a strict recommendation. On the  
541 other hand, we notice that conventional biases in saliency do not necessarily apply to UAV content.  
542 Indeed, the center bias is not systematic in UAV sequences. This bias is content-dependent as well as  
543 and task-condition-dependent. We observed new prior patterns that must be examined in the future.

544 In conclusion, the *EyeTrackUAV2* dataset enables in-depth studies of visual attention through the  
545 exploration of new salience biases and prior patterns. It establishes in addition a solid basis on which  
546 dynamic salience for UAV imaging can build upon, in particular for the development of deep-learning  
547 saliency models.

## 548 6. Acknowledgment

549 The presented work is funded by the ongoing research project ANR ASTRID DISSOCIE  
550 (Automated Detection of SaliencieS from Operators' Point of View and Intelligent Compression  
551 of DronE videos) referenced as ANR-17-ASTR-0009. Specifically, the LS2N team ran the experiment,  
552 created and made available the *EyeTrackUAV2* dataset. The Univ Rennes team added binocular and  
553 both-eyed scenarios information to the dataset, conducted analyses, and reported it.

554

- 555 1. Zhao, Y.; Ma, J.; Li, X.; Zhang, J. Saliency detection and deep learning-based wildfire identification in UAV  
556 imagery. *Sensors* **2018**, *18*, 712.
- 557 2. van Gemert, J.C.; Verschoor, C.R.; Mettes, P.; Epema, K.; Koh, L.P.; Wich, S. Nature conservation drones  
558 for automatic localization and counting of animals. Workshop at the European Conference on Computer  
559 Vision. Springer, 2014, pp. 255–270.
- 560 3. Postema, S. News Drones: An Auxiliary Perspective, 2015.
- 561 4. Agbeyangi, A.O.; Odieta, J.O.; Olorunlome, A.B. Review on UAVs used for aerial surveillance. *Journal*  
562 *of Multidisciplinary Engineering Science and Technology (JMEST)* **2016**, *3*, 5713–5719.
- 563 5. Lee-Morrison, L. *State of the Art Report on Drone-Based Warfare*; Citeseer, 2014.
- 564 6. Zhou, Y.; Tang, D.; Zhou, H.; Xiang, X.; Hu, T. Vision-Based Online Localization and Trajectory Smoothing  
565 for Fixed-Wing UAV Tracking a Moving Target. Proceedings of the IEEE International Conference on  
566 Computer Vision Workshops, 2019, pp. 0–0.

- 567 7. Zhu, P.; Du, D.; Wen, L.; Bian, X.; Ling, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; others.  
568 VisDrone-VID2019: The Vision Meets Drone Object Detection in Video Challenge Results. Proceedings of  
569 the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.
- 570 8. Aguilar, W.G.; Luna, M.A.; Moya, J.F.; Abad, V.; Ruiz, H.; Parra, H.; Angulo, C. Pedestrian detection for  
571 UAVs using cascade classifiers and saliency maps. International Work-Conference on Artificial Neural  
572 Networks. Springer, 2017, pp. 563–574.
- 573 9. Dang, T.; Khattak, S.; Papachristos, C.; Alexis, K. Anomaly Detection and Cognizant Path Planning  
574 for Surveillance Operations using Aerial Robots. 2019 International Conference on Unmanned Aircraft  
575 Systems (ICUAS). IEEE, 2019, pp. 667–673.
- 576 10. Edney-Browne, A. Vision, visuality, and agency in the US drone program. *Technology and Agency in  
577 International Relations* 2019, p. 88.
- 578 11. Krassanakis, V.; Perreira Da Silva, M.; Ricordel, V. Monitoring Human Visual Behavior during the  
579 Observation of Unmanned Aerial Vehicles (UAVs) Videos. *Drones* 2018, 2, 36.
- 580 12. Howard, I.P.; Rogers, B. Depth perception. *Stevens Handbook of Experimental Psychology* 2002, 6, 77–120.
- 581 13. Foulsham, T.; Kingstone, A.; Underwood, G. Turning the world around: Patterns in saccade direction vary  
582 with picture orientation. *Vision research* 2008, 48, 1777–1790.
- 583 14. Papachristos, C.; Khattak, S.; Mascarich, F.; Dang, T.; Alexis, K. Autonomous Aerial Robotic Exploration of  
584 Subterranean Environments relying on Morphology-aware Path Planning. 2019 International Conference  
585 on Unmanned Aircraft Systems (ICUAS). IEEE, 2019, pp. 299–305.
- 586 15. Itti, L.; Koch, C. Computational modelling of visual attention. *Nature reviews neuroscience* 2001, 2, 194.
- 587 16. Katsuki, F.; Constantinidis, C. Bottom-up and top-down attention: different processes and overlapping  
588 neural systems. *The Neuroscientist* 2014, 20, 509–521.
- 589 17. Krasovskaya, S.; MacInnes, W.J. Saliency Models: A Computational Cognitive Neuroscience Review. *Vision  
590* 2019, 3, 56.
- 591 18. Kummerer, M.; Wallis, T.S.; Bethge, M. Saliency benchmarking made easy: Separating models, maps and  
592 metrics. Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 770–787.
- 593 19. Riche, N.; Duvinage, M.; Mancas, M.; Gosselin, B.; Dutoit, T. Saliency and human fixations: State-of-the-art  
594 and study of comparison metrics. Proceedings of the IEEE international conference on computer vision,  
595 2013, pp. 1153–1160.
- 596 20. Guo, C.; Zhang, L. A novel multiresolution spatiotemporal saliency detection model and its applications  
597 in image and video compression. *IEEE transactions on image processing* 2009, 19, 185–198.
- 598 21. Jain, S.D.; Xiong, B.; Grauman, K. Fusionseg: Learning to combine motion and appearance for fully  
599 automatic segmentation of generic objects in videos. 2017 IEEE conference on computer vision and pattern  
600 recognition (CVPR). IEEE, 2017, pp. 2117–2126.
- 601 22. Wang, W.; Shen, J.; Shao, L. Video salient object detection via fully convolutional networks. *IEEE  
602 Transactions on Image Processing* 2017, 27, 38–49.
- 603 23. Li, G.; Xie, Y.; Wei, T.; Wang, K.; Lin, L. Flow guided recurrent neural encoder for video salient object  
604 detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp.  
605 3243–3252.
- 606 24. Le Meur, O.; Coutrot, A.; Liu, Z.; Rămă, P.; Le Roch, A.; Helo, A. Visual attention saccadic models  
607 learn to emulate gaze patterns from childhood to adulthood. *IEEE Transactions on Image Processing* 2017,  
608 26, 4777–4789.
- 609 25. Brunye, T.T.; Martis, S.B.; Horner, C.; Kirejczyk, J.A.; Rock, K. Visual salience and biological motion interact  
610 to determine camouflaged target detectability. *Applied ergonomics* 2018, 73, 1–6.
- 611 26. Perrin, A.F.; Zhang, L.; Le Meur, O. How well current saliency prediction models perform on UAVs videos?  
612 International Conference on Computer Analysis of Images and Patterns. Springer, 2019, pp. 311–323.
- 613 27. Bindemann, M. Scene and screen center bias early eye movements in scene viewing. *Vision research* 2010,  
614 50, 2577–2587.
- 615 28. Tseng, P.H.; Carmi, R.; Cameron, I.G.; Munoz, D.P.; Itti, L. Quantifying center bias of observers in free  
616 viewing of dynamic natural scenes. *Journal of vision* 2009, 9, 4–4.
- 617 29. Van Opstal, A.; Hepp, K.; Suzuki, Y.; Henn, V. Influence of eye position on activity in monkey superior  
618 colliculus. *Journal of Neurophysiology* 1995, 74, 1593–1610.

- 619 30. Tatler, B.W. The central fixation bias in scene viewing: Selecting an optimal viewing position independently  
620 of motor biases and image feature distributions. *Journal of vision* **2007**, *7*, 4–4.
- 621 31. Le Meur, O.; Liu, Z. Saccadic model of eye movements for free-viewing condition. *Vision research* **2015**,  
622 *116*, 152–164.
- 623 32. Vigier, T.; Da Silva, M.P.; Le Callet, P. Impact of visual angle on attention deployment and robustness  
624 of visual saliency models in videos: From SD to UHD. 2016 IEEE International Conference on Image  
625 Processing (ICIP). IEEE, 2016, pp. 689–693.
- 626 33. Zhang, K.; Chen, Z. Video Saliency Prediction based on Spatial-temporal Two-stream Network. *IEEE*  
627 *Transactions on Circuits and Systems for Video Technology* **2018**, *PP*, 1–1. doi:10.1109/TCSVT.2018.2883305.
- 628 34. Le Meur, O.; Le Callet, P.; Barba, D.; Thoreau, D. A coherent computational approach to model bottom-up  
629 visual attention. *IEEE transactions on pattern analysis and machine intelligence* **2006**, *28*, 802–817.
- 630 35. Bylinskii, Z.; Judd, T.; Oliva, A.; Torralba, A.; Durand, F. What Do Different Evaluation Metrics Tell Us  
631 About Saliency Models? *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2019**, *41*, 740–757.
- 632 36. Paglin, M.; Rufolo, A.M. Heterogeneous human capital, occupational choice, and male-female earnings  
633 differences. *Journal of Labor Economics* **1990**, *8*, 123–144.
- 634 37. Ehinger, K.A.; Hidalgo-Sotelo, B.; Torralba, A.; Oliva, A. Modelling search for people in 900 scenes: A  
635 combined source model of eye guidance. *Visual cognition* **2009**, *17*, 945–978.
- 636 38. Liu, H.; Heynderickx, I. Studying the added value of visual attention in objective image quality metrics  
637 based on eye movement data. 2009 16th IEEE international conference on image processing (ICIP). IEEE,  
638 2009, pp. 3097–3100.
- 639 39. Judd, T.; Durand, F.; Torralba, A. A benchmark of computational models of saliency to predict human  
640 fixations, 2012.
- 641 40. Ma, K.T.; Sim, T.; Kankanhalli, M. VIP: A unifying framework for computational eye-gaze research.  
642 International Workshop on Human Behavior Understanding. Springer, 2013, pp. 209–222.
- 643 41. Koehler, K.; Guo, F.; Zhang, S.; Eckstein, M.P. What do saliency models predict? *Journal of vision* **2014**,  
644 *14*, 14–14.
- 645 42. Borji, A.; Itti, L. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint*  
646 *arXiv:1505.03581* **2015**.
- 647 43. Bylinskii, Z.; Isola, P.; Bainbridge, C.; Torralba, A.; Oliva, A. Intrinsic and extrinsic effects on image  
648 memorability. *Vision research* **2015**, *116*, 165–178.
- 649 44. Fan, S.; Shen, Z.; Jiang, M.; Koenig, B.L.; Xu, J.; Kankanhalli, M.S.; Zhao, Q. Emotional attention: A study  
650 of image sentiment and visual attention. Proceedings of the IEEE Conference on Computer Vision and  
651 Pattern Recognition, 2018, pp. 7521–7531.
- 652 45. McCamy, M.B.; Otero-Millan, J.; Di Stasi, L.L.; Macknik, S.L.; Martinez-Conde, S. Highly informative  
653 natural scene regions increase microsaccade production during visual scanning. *Journal of neuroscience*  
654 **2014**, *34*, 2956–2966.
- 655 46. Gitman, Y.; Erofeev, M.; Vatolin, D.; Andrey, B.; Alexey, F. Semiautomatic visual-attention modeling and its  
656 application to video compression. 2014 IEEE International Conference on Image Processing (ICIP). IEEE,  
657 2014, pp. 1105–1109.
- 658 47. Coutrot, A.; Guyader, N. How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal*  
659 *of vision* **2014**, *14*, 5–5.
- 660 48. Coutrot, A.; Guyader, N. An efficient audiovisual saliency model to predict eye positions when looking at  
661 conversations. 2015 23rd European Signal Processing Conference (EUSIPCO). IEEE, 2015, pp. 1531–1535.
- 662 49. Wang, W.; Shen, J.; Xie, J.; Cheng, M.M.; Ling, H.; Borji, A. Revisiting video saliency prediction in the deep  
663 learning era. *IEEE transactions on pattern analysis and machine intelligence* **2019**.
- 664 50. Oh, S.; Hoogs, A.; Perera, A.; Cuntoor, N.; Chen, C.C.; Lee, J.T.; Mukherjee, S.; Aggarwal, J.; Lee, H.; Davis,  
665 L.; others. A large-scale benchmark dataset for event recognition in surveillance video. CVPR 2011. IEEE,  
666 2011, pp. 3153–3160.
- 667 51. Layne, R.; Hospedales, T.M.; Gong, S. Investigating open-world person re-identification using a drone.  
668 European Conference on Computer Vision. Springer, 2014, pp. 225–240.
- 669 52. Bonetto, M.; Korshunov, P.; Ramponi, G.; Ebrahimi, T. Privacy in mini-drone based video surveillance.  
670 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG).  
671 IEEE, 2015, Vol. 4, pp. 1–6.

- 672 53. Shu, T.; Xie, D.; Rothrock, B.; Todorovic, S.; Chun Zhu, S. Joint inference of groups, events and human  
673 roles in aerial videos. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*,  
674 2015, pp. 4576–4584.
- 675 54. Mueller, M.; Smith, N.; Ghanem, B. A benchmark and simulator for uav tracking. *European conference on*  
676 *computer vision*. Springer, 2016, pp. 445–461.
- 677 55. Robicquet, A.; Sadeghian, A.; Alahi, A.; Savarese, S. Learning social etiquette: Human trajectory  
678 understanding in crowded scenes. *European conference on computer vision*. Springer, 2016, pp. 549–565.
- 679 56. Li, S.; Yeung, D.Y. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion  
680 models. *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- 681 57. Barekatin, M.; Martí, M.; Shih, H.F.; Murray, S.; Nakayama, K.; Matsuo, Y.; Prendinger, H. Okutama-action:  
682 An aerial view video dataset for concurrent human action detection. *Proceedings of the IEEE Conference*  
683 *on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 28–35.
- 684 58. Hsieh, M.R.; Lin, Y.L.; Hsu, W.H. Drone-based object counting by spatially regularized regional proposal  
685 network. *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4145–4153.
- 686 59. Ribeiro, R.; Cruz, G.; Matos, J.; Bernardino, A. A dataset for airborne maritime surveillance environments.  
687 *IEEE Trans. Circuits Syst. Video Technol* **2017**.
- 688 60. Hsu, H.J.; Chen, K.T. DroneFace: an open dataset for drone research. *Proceedings of the 8th ACM on*  
689 *Multimedia Systems Conference*. ACM, 2017, pp. 187–192.
- 690 61. Božić-Štulić, D.; Marušić, Ž.; Gotovac, S. Deep Learning Approach in Aerial Imagery for Supporting Land  
691 Search and Rescue Missions. *International Journal of Computer Vision* **2019**, pp. 1–23.
- 692 62. Fu, K.; Li, J.; Shen, H.; Tian, Y. How drones look: Crowdsourced knowledge transfer for aerial video  
693 saliency prediction. *arXiv preprint arXiv:1811.05625* **2018**.
- 694 63. Zhu, P.; Wen, L.; Bian, X.; Ling, H.; Hu, Q. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*  
695 **2018**.
- 696 64. Nyström, M.; Andersson, R.; Holmqvist, K.; Van De Weijer, J. The influence of calibration method and eye  
697 physiology on eyetracking data quality. *Behavior research methods* **2013**, *45*, 272–288.
- 698 65. ITU-T RECOMMENDATION, P. Subjective video quality assessment methods for multimedia applications.  
699 *International telecommunication union* **2008**.
- 700 66. Rec, I. BT. 710-4. *Subjective assessment methods for image quality in high-definition television* **1998**.
- 701 67. Cornelissen, F.W.; Peters, E.M.; Palmer, J. The Eyelink Toolbox: eye tracking with MATLAB and the  
702 Psychophysics Toolbox. *Behavior Research Methods, Instruments, & Computers* **2002**, *34*, 613–617.
- 703 68. Rec, I. BT. 500-13. *Methodology for the subjective assessment of the quality of television pictures* **2012**.
- 704 69. Wandell, B.; Thomas, S. Foundations of vision. *Psychocritiques* **1997**, *42*.
- 705 70. Le Meur, O.; Baccino, T. Methods for comparing scanpaths and saliency maps: strengths and weaknesses.  
706 *Behavior research methods* **2013**, *45*, 251–266.
- 707 71. Guznov, S.; Matthews, G.; Warm, J.S.; Pfahler, M. Training techniques for visual search in complex task  
708 environments. *Human factors* **2017**, *59*, 1139–1152.
- 709 72. Shah, M.; Javed, O.; Shafique, K. Automated visual surveillance in realistic scenarios. *IEEE MultiMedia*  
710 **2007**, *14*, 30–39.
- 711 73. Snellen, H. *Test-types for the determination of the acuteness of vision*; Williams and Norgate, 1868.
- 712 74. Ishihara, S. *Test for colour-blindness*; Kanehara Tokyo, Japan, 1987.
- 713 75. Salvucci, D.D.; Goldberg, J.H. Identifying fixations and saccades in eye-tracking protocols. *Proceedings of*  
714 *the 2000 symposium on Eye tracking research & applications*. ACM, 2000, pp. 71–78.
- 715 76. Krassanakis, V.; Filippakopoulou, V.; Nakos, B. EyeMMV toolbox: An eye movement post-analysis tool  
716 based on a two-step spatial dispersion threshold for fixation identification. *Journal of eye movement research*  
717 **2014**.
- 718 77. Krassanakis, V.; Misthos, L.M.; Menegaki, M. LandRate toolbox: An adaptable tool for eye movement  
719 analysis and landscape rating. *Eye Tracking for Spatial Research*, *Proceedings of the 3rd International*  
720 *Workshop*. ETH Zurich, 2018.
- 721 78. Krassanakis, V.; Filippakopoulou, V.; Nakos, B. Detection of moving point symbols on cartographic  
722 backgrounds. *Journal of Eye Movement Research* **2016**, *9*.
- 723 79. Ooms, K.; Krassanakis, V. Measuring the Spatial Noise of a Low-Cost Eye Tracker to Enhance Fixation  
724 Detection. *Journal of Imaging* **2018**, *4*, 96.



- 725 80. Cui, Y.; Hondzinski, J.M. Gaze tracking accuracy in humans: Two eyes are better than one. *Neuroscience*  
726 *letters* **2006**, *396*, 257–262.
- 727 81. Holmqvist, K.; Nyström, M.; Mulvey, F. Eye tracker data quality: what it is and how to measure it.  
728 Proceedings of the symposium on eye tracking research and applications. ACM, 2012, pp. 45–52.
- 729 82. Hooge, I.T.; Holleman, G.A.; Haukes, N.C.; Hessels, R.S. Gaze tracking accuracy in humans: One eye is  
730 sometimes better than two. *Behavior Research Methods* **2018**, pp. 1–10.
- 731 83. Bylinskii, Z.; Judd, T.; Borji, A.; Itti, L.; Durand, F.; Oliva, A.; Torralba, A. Mit saliency benchmark, 2015.
- 732 84. Le Meur, O.; Coutrot, A. Introducing context-dependent and spatially-variant viewing biases in saccadic  
733 models. *Vision research* **2016**, *121*, 72–84.

734 © 2019 by the authors. Submitted to *Drones* for possible open access publication under the terms and conditions  
735 of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).