



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, 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. 2019. hal-02391832v1

HAL Id: hal-02391832

<https://univ-rennes.hal.science/hal-02391832v1>

Preprint submitted on 3 Dec 2019 (v1), 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.

Article

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

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

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

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

³ Univ Rennes, INSA Rennes, CNRS, IETR - UMR 6164, 35000 Rennes, France; lu.lu@insa-rennes.fr

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

Version December 3, 2019 submitted to Journal Not Specified

Abstract: Unmanned Aerial Vehicles (UAVs) achieved a lot of momentum through their fast and tremendous evolution over the last decade. A multiplication of applications results from the use of the UAV imagery 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 compression, retargeting, and decision-making tools. To conduct such studies, in this era of big data and deep learning, we identified the need for new large-scale eye-tracking datasets for visual salience in UAV content. To address this need, we introduce here the dataset *EyeTrackUAV2* consisting of the collection of binocular gaze information through visualization of UAV videos for both free viewing and task-based attention conditions. 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 applications from making vacation movies to drone races for mainstream civilian applications, from fire detection [1], wildlife counting [2] to journalism [3], precision agriculture and delivery services for professional applications, and from military aerial surveillance [4], drone-based warfare [5] to tracking moving targets [6], object, person or anomaly detection [7–9] for military applications.

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]. Also, their high autonomy in conjunction with large-field of view camera permit to cover large areas in limited time duration. Besides, UAV sensors can be multi-modal and can include RGB, thermal, Infra-Red (IR), or multi-spectral sensors. Multiplying imagery modalities allows overcoming possible weaknesses of RGB-only [10]. For instance, occlusions

30 may be compensated by thermal information, and the capture of IR is desired for low-luminance
31 environments [12].

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

40 Visual attention occurs to filter and sort out visual clues. Indeed, it is impossible to process all
41 the information one perceives. Particular consideration should be dedicated to identifying which
42 attentional processes are involved as they are diverse and aim at specific behaviors. For instance, one
43 must make the distinction between overt and covert attention [13]. The former refers to a direct focus
44 onto where points eyes and head. The latter relates more to the peripheral vision, where attention is
45 directed without eye movements towards it. In practice, when an object of interest is detected in the
46 area covered by the covert attention, one may make a saccade movement to direct the eyes from the
47 overt area to this position. The context of visualization is also important. For instance, we make a
48 distinction between two content exploration processes [14]: (1) A no constraint examination named
49 free viewing. The observer is rather free from cognitive loads and is supposed to mainly use bottom-up
50 or exogenous attention processes driven by external factors, e.g. content and environment stimuli.
51 (2) A task-based visualization, such as surveillance for instance. Cognitive processes such as prior
52 knowledge, willful plans, and current goals guide the viewer's attention. This is known as top-down
53 or endogenous attention. A strict division is slightly inaccurate in that both top-down and bottom-up
54 processes are triggered during a visual stimuli in a very intricate interaction [15]. Both processes are
55 important and need to be studied through salience.

56 Visual salience means to represent attention in multimedia content as a probability distribution per
57 pixels [16]. Saliency analyses rest on the relation of visual attention to eye movements, and these latter
58 are obtained through gaze collection with eye-trackers [17]. Saliency predictions help to understand
59 computational cognitive neuroscience as it reveals attention behaviors such as center bias and spatial
60 and temporal inhibition of return [15]. Multiple applications derive from saliency predictions such as
61 compression [18], content-aware re-targeting, object segmentation [19], and detection [20,21].

62 Recently, there has been a growing interest on one particular application, which combines visual
63 salience and UAV content. Information overload in the drone program and fatigue in military operators
64 may have disastrous consequences for military applications [10]. New methods and approaches are
65 required to detect anomaly in UAV footages and to ease the decision-making. Among them, we believe
66 that computational models of visual attention could be used to simulate operators' behaviors [22].
67 Eventually, thanks to predictions, operators' workloads can be reduced by eliminating unnecessary
68 footages segments.

69 To that end, it is necessary to develop new dynamic saliency models tailored to UAV content.
70 The gain brought by deep-learning saliency models this last decade [18–21,23] has been more than
71 significant. This improvement comes with the definition and the design of large-scale eye tracking
72 datasets, from which a ground truth can be defined. However, in the context of UAV content, there are
73 very few eye-tracking datasets. This is the reason why we propose and present in this paper a new
74 large-scale eye-tracking dataset, freely downloadable from internet.

75 The paper is organized as follows. In section 2, we first justify and elaborate on the need for
76 large-scale eye-tracking databases for UAV videos. Then, we introduce the entire process of dataset
77 creation in section 3. It describes the content selection, the experiment set up, and the implementation
78 of fixations, saccades, and saliency maps. Section 4 presents an in-depth analysis of the dataset. The
79 study is two-fold: it explores what ground truth should be used for salience studies, and brings to light

80 the fading of conventional biases in visual salience for UAV stimuli. Finally, conclusions are provided
81 in section 5.

82 2. Related Work

83 While it is now rather easy to find eye tracking data on typical images [24–33] or videos [34–38],
84 and that there are many UAV content datasets [7,39–50], it turns out to be extremely difficult to find
85 eye-tracking data on UAV content. Indeed, very few works are dealing with eye-tracking data related
86 to UAV content. This is even truer when we consider dynamic salience, which refers to salience for
87 video content.

88 To the best of our knowledge, *EyeTrackUAV1* dataset released in 2018 [11] is the only public dataset
89 available for studying the visual deployment over UAV video. There exist another dataset *AVS1K* [51].
90 However, *AVS1K* is, to the present day, not publicly available. We thus focus here on *EyeTrackUAV1*,
91 with the awareness that all points below but the last apply to *AVS1K*.

92 *EyeTrackUAV1* consists in 19 sequences (1280x720 and 30 frame per second (fps)) extracted from
93 the *UAV123* database [43]. The sequence selection relied on content characteristics, which are the
94 diversity of environment, distance and angle to the scene, size of the principal object, and the presence
95 of sky. Precise binocular gaze data (1000 Hz) of 14 observers were recorded under free viewing
96 condition, for every content. Overall, the dataset comprises eye-tracking information on 26599 frames,
97 which represents 887 seconds of video. In spite of a number of merits, this dataset presents several
98 limitations for saliency prediction applications. These limitations have been listed in [23]. We briefly
99 summarize them below:

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

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

124 3. *EyeTrackUAV2* dataset

125 This section introduces the new dataset *EyeTrackUAV2* aiming at tackling issues mentioned above.
126 *EyeTrackUAV2* includes more video content than its predecessor *EyeTrackUAV1*. It involves more

127 participants, and considers both free and task-based viewing. In the following subsections, we first
128 elaborate on the selection of video content, followed by a description of the eye-tracking experiment.
129 It includes the presentation of the eye-tracking apparatus, the experiment procedure and setup, and
130 the characterization of population samples. Finally, we describe the generation of the human ground
131 truth, i.e. algorithms for fixation and saccade detection as well as saliency map computation.

132 3.1. Content selection

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

139 After examining a number of UAV datasets (UCF's dataset ¹, VIRAT [39], MRP [40], the
140 privacy-based mini-drones dataset [41], the aerial videos dataset described in [42], UAV123 [43],
141 DTB70 [45], Okutama-Action [46], VisDrone [52], CARPK [47], SEAGULL [48], DroneFace [49], and the
142 aerial video dataset described in [44]), a total of 43 videos (RGB, 30 fps, 1280x720 or 720x480) have been
143 selected from 3 different databases, *VIRAT*, *UAV123* and *DTB70*. These three databases are exhibiting
144 different content for various applications, which makes the final selection representative of the UAV
145 ecosystem. Table 2 reports the number of sequences selected from each database and details their
146 native resolution, duration and frame number. We present below the main characteristics of the three
147 selected datasets:

- 148 • *UAV123* includes challenging UAV content annotated for object tracking. We restrict the content
149 selection to the first set, which includes 103 sequences (1280x720 and 30 fps) captured by an
150 off-the-shelf professional-grade UAV (DJI S1000) tracking various objects in a range of altitudes
151 comprised between 5-25 meters. Sequences include a large variety of environments (e.g. urban
152 landscapes, roads, and marina), objects (e.g. cars, boats, and persons) and activities (e.g. walking,
153 biking, and swimming) as well as present many challenges for object tracking (e.g. long- and
154 short-term occlusions, illumination variations, viewpoint change, background clutter, and camera
155 motion).
- 156 • Aerial videos in the VIRAT dataset were manually selected (for smooth camera motion and
157 good weather conditions) from rushes of a total amount of 4 hours in outdoor areas with broad
158 coverage of realistic scenarios for real-world surveillance. Content includes "single person",
159 "person and vehicle", and "person and facility" events, with changes in viewpoints, illumination,
160 and visibility. The dataset comes with annotations of moving object tracks and event examples
161 in sequences. These videos (720x480 and 30 fps) exhibit quite low quality and include content
162 shot in infra-red.
- 163 • The 70 videos (RGB, 1280x720 and 30 fps) from DTB70 dataset are manually annotated with
164 bounding boxes for tracked objects. Sequences were shot with a DJI Phantom 2 Vision+ drone or
165 were collected from YouTube to add diversity in environments and target types (mostly humans,
166 animals, and rigid objects). There is also a variety of camera movements (both translation and
167 rotation), short- and long-term occlusions, and target deformability.

168 Table 1 presents the sequences which have been extracted from their original datasets. Video
169 characteristics such as duration, spatial and temporal complexity are also given. Native resolutions are
170 provided by original dataset in Table 2.

¹ http://csrcv.ucf.edu/data/UCF_Aerial_Action.php

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 - IR	Small	False	Oblique	
2	09152008flight2tape1_3 (crop 2)		367	137	503	12234	0,474	35	High	Urban military - IR	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 - IR	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	

Table 1. Stimuli ID and name, their original dataset, number of frames together with starting and ending frame number, duration and native resolution.

Dataset	Native resolution	Proportion of content seen per degree of visual angle (%)	Videos number	Frames number (30 fps)	Duration (sec)
VIRAT [39]	720 x 480	1,19	12	17851	595,03
UAV123 [43]	1280 x 720	0,44	22	20758	691,93
DTB70 [45]	1280 x 720	0,44	9	3632	121,07
Overall			43	42241	1408,03 (23:28 min)

Table 2. 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 3. Basic statistics on selected videos.

171 3.2. Content Diversity

172 To present the diversity of selected UAV sequences, Figure 1 illustrates the first frame of every
 173 content. Visual stimuli cover a variety of visual scenes in different environments (e.g. public and
 174 military environments, roads, buildings, sports, and port areas, etc.) and different moving or fixed
 175 objects (e.g. people, groups of people, cars, boats, bikes, motorbikes, etc.). Selected videos were
 176 captured from various flight heights and different angles between the UAV and the ground (allowing
 177 or not the presence of sky during their observation). This information is reported per sequence in
 178 Table 1. Additionally, we considered various video duration as the length of the video may possibly
 179 impact the behavior of observers due to fatigue, resulting in a lack of attention and more blinking
 180 artifacts [10,53].

181 To quantitatively show the diversity of selected videos, we have computed temporal and spatial
 182 complexity [54], named TI ($\in [0, +\infty)$) and SI ($\in [0, +\infty)$), respectively. These features are commonly used
 183 in image quality domain for characterizing the properties of selected images. They characterize the
 184 maximum standard deviation of spatial and temporal discrepancies over the entire sequence. The
 185 higher a measure is, the more complex is the content. TI and SI are reported per sequence in Table 1.
 186 The range of temporal complexity in sequences is broad, displaying the variety of movements present
 187 in sequences. Spatial measures are more homogeneous. Indeed, the spatial complexity is due to
 188 the bird point of view of the sensor. The aircraft high up position offers access to a large amount of
 189 information.

190 Three sequences, extracted from the VIRAT dataset, were captured by IR cameras. As a side note,
 191 finding non-natural content for UAV of sufficient quality in publicly available datasets was difficult.

192 Finally, Table 3 presents basic statistics of the database in terms of number of frames and duration.
 193 The 43 selected videos are now referred to as test stimuli in the following section, which presents the
 194 experiment design.

195 3.3. Experimental design

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

198 3.3.1. Eye-tracking apparatus

199 A specific setup is designed to capture eye-tracking information on video stimuli. It includes a
 200 rendering monitor, an eye-tracking system, a control operating system, and a controlled laboratory test
 201 room.

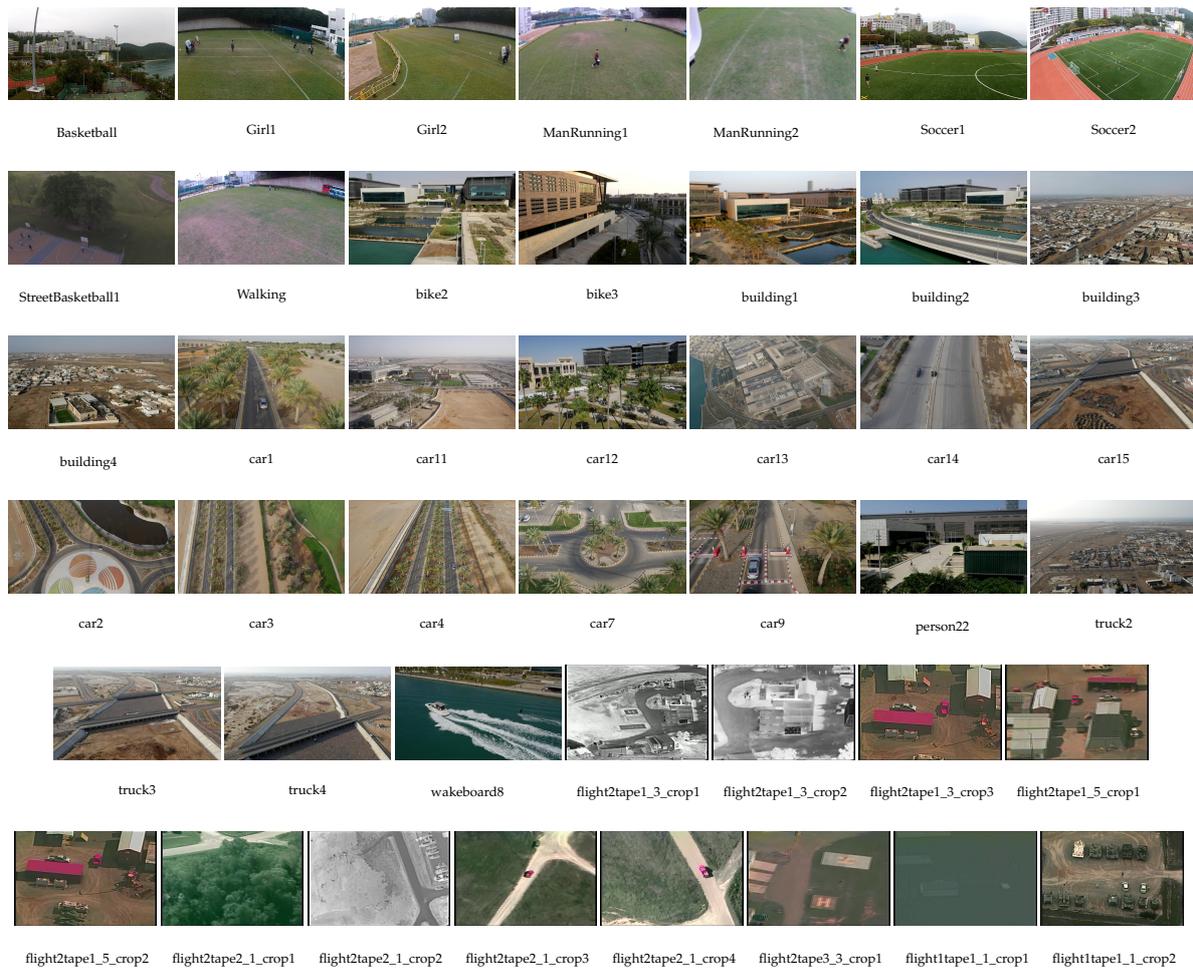


Figure 1. *EyeTrackUAV2* dataset: first frame of each sequence.

202 To run the experiment and collect gaze information, we used the EyeLink® 1000 Plus eye-tracking
 203 system, in the head free-to-move *remote mode*, taking advantage of its embedded 25mm camera lens.
 204 The eye tracker principle is to detect and record the IR illuminator reflection rays on the observer's
 205 pupil. This system enables the collection of highly precise gaze data at a temporal frequency of
 206 1000 Hz and a spatial accuracy between the visual angle range of 0.25 and 0.50 degree, according to
 207 the manufacturer. To prevent errors in the robust algorithm for the detection of observers' pupils [53],
 208 participants were asked to remove any excess of mascara if need be. Also, the eye tracker's camera
 209 was configured for each subject, without affecting the corresponding distance between them. This
 210 configuration guarantees to achieve an optimal detection of the observer's eyes and head sticker.

211 The experimental monitor which displayed stimuli was a 23.8 inches (52,70 × 29,65 cm) DELL
 212 P2417H computer monitor display ² with full HD resolution (1920×1080) at 60 Hz and with a response
 213 time of 6 ms. As suggested by both the International Telecommunication Union (ITU)-Broadcasting
 214 service (Television) (BT).710 [55] and manufacturer, observers sited in distance of about 3H (1m ±
 215 10cm) from the monitor, where H corresponds to the stimuli display height so that observers have an
 216 assumed spatial visual angle acuity of one degree. Moreover, the eye tracker camera was placed 43 cm
 217 away from the experimental display, and thus about 67 cm from participants. Based on this setting,

² <https://www.dell.com/cd/business/p/dell-p2417h-monitor/pd>

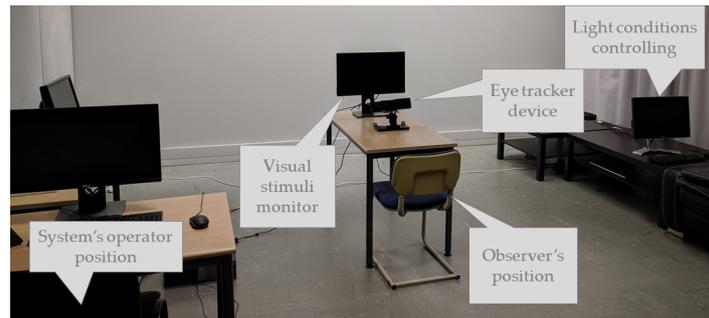


Figure 2. Experiment setup.

218 there are 64 pixels per degree of visual angle in each dimension, and the display resolution is about
 219 30x17 visual degrees.

220 Regarding software, the MPC-HC video player ³, considered as one of the most lightweight
 221 open-source video players, rendered the experimental video stimuli. Also, we took advantage of the
 222 Eyelink toolbox [56] as it is part of the 3rd version of Psychophysics Toolbox Psychtoolbox-3 (PTB-3) ⁴
 223 and added in-house communication processes ⁵ for sync between control and display systems. The
 224 control system consists of an additional computer, used by the experimenter to configure and control
 225 the eye-tracking system with an Ethernet connection.

226 Eventually, eye-tracking tests were performed in a room with controlled constant light conditions.
 227 The performed calibration set the constant ambient light conditions at approximately 36.5 cd/m², i.e.
 228 15% of the maximum stimuli monitor brightness - 249 cm/m² - as recommended by the ITU-BT.500 [57],
 229 with the i1 Display Pro X-Rite® system.

230 Figure 2 illustrates the experimental setup used during the collection of gaze information. We can
 231 observe the arrangement of all the systems described above.

232 3.3.2. Stimuli presentation

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

239 Before starting the experiment, a training session is organized to get the subject familiar with the
 240 experiment design. It includes a calibration procedure and its validation followed by the visualization
 241 of one video. This UAV video, the sequence *car4* from the DTB70 dataset, is additional to test stimuli
 242 to avoid any memory bias. Once subjects completed the training session, they could ask questions to
 243 experimenters before taking part into test sessions.

244 Regarding test sessions, they start with calibration and its validation before the visualization
 245 of 9 videos, during which subjects do or do not perform a task. To ensure the optimal quality of
 246 the collected gaze data, each participant took part in five test sessions. Splitting the experiment into
 247 sessions decreases the tiredness and lack of attention in observers. Also, this design enables frequent
 248 calibration so that recordings do not suffer from the decrease of accuracy in gaze recordings with
 249 time [53].

³ <https://mpc-hc.org/>

⁴ <http://psychtoolbox.org/>

⁵ LS2N, University of Nantes

250 With regards to calibration, the eye-tracking system is calibrated for each participant, following a
251 typical 13 fixed-point detection procedure [53]. Actually, experimenters started tests with a 9-point
252 strategy for calibration (subject 1 to 17 in Free Viewing (FV)) but realized that gaze collection is more
253 accurate with a 13-point calibration. The calibration reaches validation when the overall deviation
254 of both eye positions is approximately below the fovea vision accuracy (e.g. a degree of visual
255 angle [53,58]). The calibration procedure is repeated until validation.

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

259 3.3.3. Visual tasks to perform

260 *EyeTrackUAV2* aims to investigate two visual tasks. Indeed, we want to be able to witness visual
261 attention processes triggered by top-down (or goal-directed) and bottom-up (or stimulus-driven)
262 attention. Accordingly, we defined two visual tasks participants have to perform: the first condition is
263 a Free Viewing (FV) task while the second relates to a surveillance-viewing Task (Task). The former
264 task is rather common in eye-tracking tests [24,31,33,38,59,60]. Observers were simply asked to observe
265 visual video stimuli without performing any task. For the surveillance-viewing task, participants
266 were required to watch video stimuli and to push a specific button on a keyboard each time they
267 observe a new - meaning not presented before - moving object (e.g. people, vehicle, bike, etc.) in
268 the video. The purpose of this task is to simulate one of the basic surveillance procedures in which
269 targets could be located anywhere when the visual search process was performed [61]. After reviewing
270 typical surveillance systems' abilities [62], we have decided to define our task as object detection. The
271 defined object detection task is compelling in that it encompasses target-specific training (repeated
272 discrimination of targets and non-targets) and visual search scanning (targets potentially located
273 anywhere) [61]. The only task-related behavior not explored within this task is cue training (targets
274 likely to be co-located with more salient objects or events). The surveillance-viewing task is especially
275 interesting for a military context, in which operators have to detect discrepancies in drone videos.

276 3.3.4. Population

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

284 Each observer has been tested for visual acuity and color vision with Ishihara and Snellen
285 tests [63,64]. Any failure to these tests motivated the dismissal of the person from the experiment.
286 Before running the test, the experimenter provided subjects with written consents and information
287 forms, together with oral instructions. This process made sure of the consent of participants and their
288 understanding of the experiment process. It also ensures an anonymous data collection.

289

Sample statistics	FV	Task	Total
Participants	30	30	60
Females	16	16	32
Males	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, the conversion of collected raw signals into the pixel coordinate system of the original sequence 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 saliency maps (also called grayscale heatmaps) 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 this remapping has been done for both eyes, the spatial binocular coordinates is simply given by the average of the spatial coordinates of left and right eyes.

For the surveillance-viewing task, each gaze point was assigned with the relative information of the button reaction. We denote in raw data a button activation (respectively no button reaction) with the Boolean value 1 (respectively 0). Finally, for convenience, we have sorted the positions of the observer's dominant eyes and included them in raw gaze data.

3.4.2. Fixation and saccade event detection

To retrieve fixations from eye positions, we used the Dispersion-Threshold Identification (I-DT) [65] from the EyeMMV and LandRate toolboxes [66,67]. This algorithm performs "two-step" spatial and temporal thresholds. As exposed in [67,68], thanks to the very high precision of our eye-tracking equipment, we can combine the two-step spatial thresholds in one operation, as both thresholds have the same value. Ultimately, in our context, this algorithm conceptually implements a spatial noise removal filter and a temporal threshold indicating the minimum fixation duration. We have selected

316 the minimum threshold values from the state of the art to ensure the performance of the fixation
317 detection algorithm. Accordingly, spatial and temporal thresholds were selected to be equal to 0.7
318 degree of the visual angle and 80 ms [69], respectively. Finally, saccade events were calculated based
319 on the computed fixations considering that a saccade corresponds to eye movements between two
320 successive fixation points.

321 When considering raw data of the dominant eye, I-DT exhibits a total number of fixations of 1 239
322 157 in FV and 1 269 433 in Task.

323 3.4.3. Saliency maps

324 Saliency maps are a 2D topographic representation indicating the ability of an area to attract
325 observers' attention. It is common to represent the salience of an image thanks to either its saliency map
326 or by its colored representation, called heatmap. Saliency maps are usually computed by convolving
327 the fixation map, gathering observers' fixations, with a Gaussian kernel representing the foveal part of
328 our retina. More details can be found in [60].

329 For video sequence, there is one saliency map for each frame of test sequences. It could be
330 debatable to do so for temporal analyses, but it is current practices to deal with videos as a succession
331 of frames in visual media processing (e.g. compression, High Dynamic Range (HDR) video tone
332 mapping, and dynamic saliency).

333 We took benefit from the high frequency of acquisition of the eye-tracker system to compute
334 saliency maps directly from raw gaze data (in pixel coordinates). Thus, our saliency maps are free
335 from any biases that could be introduced by any fixation extractor algorithms. Hence, the generated
336 saliency maps include fixation and saccade information, without distinction.

337 To indicate salient regions of each frame, we followed the method described in [66], with
338 parameters derived from the experimental setup (e.g., a grid size of a pixel, a standard deviation of 0.5
339 degree of angle i.e. $\sigma = 32$ pixels, and a kernel size of 6σ). For visualization purposes, heat maps were
340 normalized between 0 and 255.

341 Figure 4 presents saliency maps obtained for both attention conditions in frame 100 of seven
342 sequences. We have selected frame 100 to get free from the initial center-bias in video exploration
343 occurring during the first seconds. These examples illustrate, for instance, the sparsity of salience in
344 videos in free viewing, while task-based attention usually presents more salient points, more dispersed
345 in the content than FV, depending on the task and attention-grabbing objects.

346 3.5. *EyeTrackUAV2 in brief*

347 We have created a dataset containing binocular gaze information collected during two viewing
348 conditions (free viewing and task) over 43 UAV videos (30 fps, 1280x720 and 720x480 - 42241 frames,
349 1408 seconds) observed by 30 participants per condition, leading to 1 239 157 fixations in free viewing
350 and 1 269 433 in task-viewing for dominant eyes positions. Notably, selected UAV videos sowing
351 diversity in rendered environments, movement and size of objects, aircraft flight heights and angles to
352 the ground, duration, size, and quality. This dataset overcomes the limitations of *EyeTrackUAV1* in that
353 it enables investigations of salience in more test sequences, on larger population samples, and for both
354 free-viewing and task-based attention. Additionally, and even though they are still too few, three IR
355 videos are part of visual stimuli.

356 Fixations, saccades, and saliency maps were computed - for both eyes in additive and averaged
357 fashions (see Binocular and BothEyes scenarios described later) and for the dominant eye - and are
358 publicly available with original content and raw data on our FTP ⁶. The code in MATLAB to generate
359 all ground truth information is also made available.

⁶ <ftp.ivic.polytech.univ-nantes.fr>

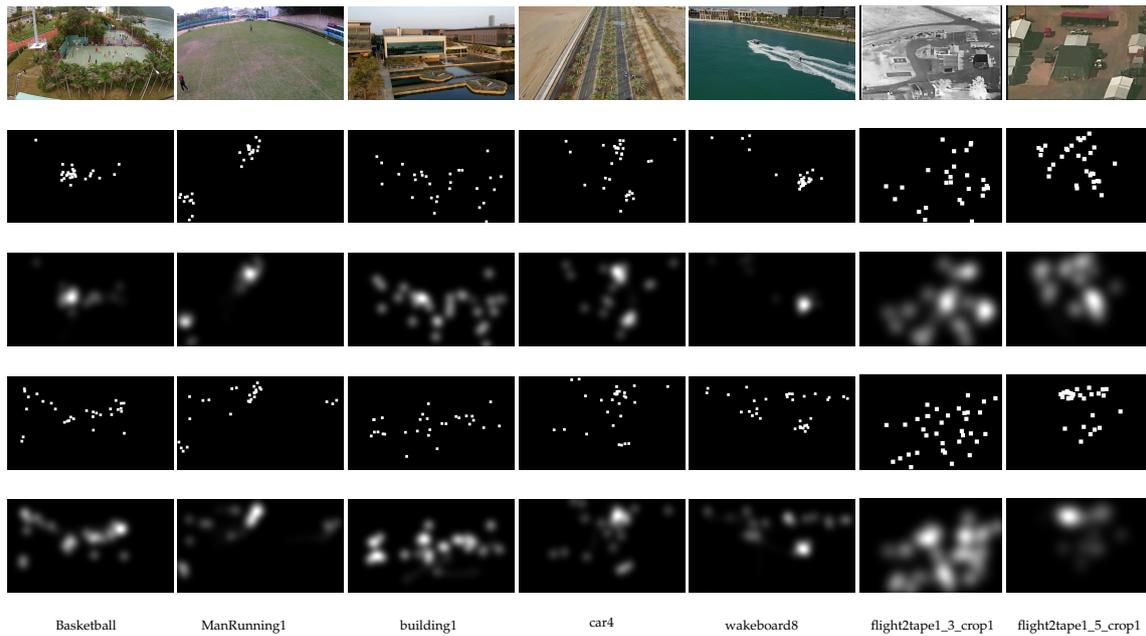


Figure 4. Frame 100 of seven sequences of *EyeTrackUAV2* dataset, together with saliency and fixation maps generated based on gaze data of dominant eye. Results are presented for both types of attention. The first row presents sequences hundredth frame, the second fixations for FV, the third saliency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

360 4. Analyses

361 In this section, we characterize the proposed *EyeTrackUAV2* database. On one hand, we compare
 362 eye positions and salience between the different attention conditions. Such information may be of
 363 great importance to generate the ground truth on a case-by-case basis. On the other hand, UAV videos
 364 induce new visual experiences. Consequently, observers exhibit different behaviors towards this type
 365 of stimuli. Therefore, we investigate whether the center bias, one of the main viewing tendencies [70],
 366 still applies to *EyeTrackUAV2* content.

367 4.1. Six different ground truths

368 The first question we address concerns the method used to determine the ground truth. In a
 369 number of papers, researchers use the ocular dominance theory in order to generate a ground truth.
 370 This theory relies on the fact that the human visual system favors the input of one eye over the other
 371 should binocular images be too disparate on the retinas. However, the cyclopean theory gains more
 372 and more momentum [71,72]. It alleges that vision processes approximate a central point between
 373 two eyes, from which an object is perceived. Furthermore, lately, manufacturers achieved major
 374 improvements in eye-tracking systems. They are now able to record and calibrate the positions of
 375 both eyes separately. This allows for exploring what are the best practices to create salience ground
 376 truth [71–73].

377 We therefore propose to evaluate the potential errors made when different methods for creating the
 378 ground truth are used. We tested six methods, namely Left (L), Right (R), Binocular (B), Dominant (D),
 379 non Dominant (nD), Both Eyes (BE), under the two visual attention conditions, Free Viewing (FV) and
 380 surveillance-viewing Task (Task). B corresponds to the average position between the left and right
 381 eyes and can be called version signal. BE includes the positions of both L and R positions, and hence
 382 comprises twice more information than other scenarios. nD has been added to estimate the gain made
 383 when using dominant eye information. Estimating the relevance of the aforementioned methods will
 384 help to decide which ground truth scenario should be used depending on the precision and accuracy

		FV						Task					
		Average		max		min		Average		max		min	
		hor	vert										
		x	y	x	y	x	y	x	y	x	y	x	y
Binocular	EyeNonDom	0,23	0,26	0,29	0,31	0,19	0,21	0,23	0,20	0,28	0,23	0,20	0,18
Binocular	Right	0,23	0,26	0,29	0,31	0,19	0,21	0,23	0,20	0,28	0,23	0,20	0,18
Binocular	Left	0,23	0,26	0,29	0,31	0,19	0,21	0,23	0,20	0,28	0,23	0,20	0,18
Binocular	Dominant	0,23	0,26	0,29	0,31	0,19	0,21	0,23	0,20	0,28	0,23	0,20	0,18
EyeNonDom	Right	0,16	0,17	0,21	0,24	0,12	0,10	0,31	0,28	0,37	0,34	0,25	0,25
EyeNonDom	Left	0,31	0,34	0,40	0,44	0,22	0,27	0,15	0,12	0,20	0,15	0,11	0,10
EyeNonDom	Dominant	0,46	0,51	0,57	0,63	0,39	0,42	0,46	0,40	0,56	0,45	0,39	0,36
Right	Left	0,46	0,51	0,57	0,63	0,39	0,42	0,46	0,40	0,56	0,45	0,39	0,36
Right	Dominant	0,31	0,34	0,40	0,44	0,22	0,27	0,15	0,12	0,20	0,15	0,11	0,10
Left	Dominant	0,16	0,17	0,21	0,24	0,12	0,10	0,31	0,28	0,37	0,34	0,25	0,25
Binocular	BothEyes	0,23	0,26	0,29	0,31	0,19	0,21	0,23	0,20	0,28	0,23	0,20	0,18
EyeNonDom	BothEyes	0,23	0,26	0,29	0,31	0,19	0,21	0,23	0,20	0,28	0,23	0,20	0,18
Right	BothEyes	0,23	0,26	0,29	0,31	0,19	0,21	0,23	0,20	0,28	0,23	0,20	0,18
Left	BothEyes	0,23	0,26	0,29	0,31	0,19	0,21	0,23	0,20	0,28	0,23	0,20	0,18
Dominant	BothEyes	0,23	0,26	0,29	0,31	0,19	0,21	0,23	0,20	0,28	0,23	0,20	0,18

Table 5. MAE in eye positions depending on scenarios for Free Viewing and Task viewing, in degree per pixels. Results are provided in average, minimum and maximum over sequences and observers. Highest values are emphasised in red, the least in blue.

385 one requires [73]. We performed two evaluations: the first directly on eye positions and the second on
 386 human saliency maps.

387 4.1.1. Mean of Absolute Error of eye positions

388 To characterize the error made when choosing a scenario over another, we compute the Mean of
 389 Absolute Error (MAE) between eye positions for both viewing conditions:

$$MAE_x^{(i,j)} = \frac{1}{N} \sum_n |x_{s,n}^{M_i} - x_{s,n}^{M_j}|$$

$$MAE_y^{(i,j)} = \frac{1}{N} \sum_n |y_{s,n}^{M_i} - y_{s,n}^{M_j}|$$

390 With $MAE_x^{(i,j)}$ and $MAE_y^{(i,j)}$ the Mean Absolute Error for x and y axis, respectively, and for methods
 391 M_i and M_j . Method refers here to Left (L), Right (R), Binocular (B), Dominant (D), non Dominant (nD),
 392 Both Eyes (BE). N denotes the number of gaze samples for one sequence, for all observers. Note that
 393 $MAE_x^{(i,i)} = MAE_y^{(i,i)} = 0$.

394 Table 5 reports the average, maximum and minimum MAE over sequences for all comparisons
 395 of scenarios. We observe several interesting behaviors, such as the decrease of error when using the
 396 dominant eye in a single-eyed method, or the fixed error made when using both eyes positions. But
 397 what really is noteworthy is that the maximum average MAE is about half of a degree of visual angle,
 398 the maximum error value being 0,63.

399 Half a degree of visual angle is usually the least value of Gaussian kernel used to filter eye
 400 positions (or fixations) when creating saliency maps. We thus extend the analysis with a comparison of
 401 methods in terms of similarity between saliency maps. We want to know if it makes a difference to
 402 use different eye positions to generate saliency ground truth. Indeed, this would help the scientific
 403 community to know how to use eye-tracking data for saliency depending on which information is
 404 available. It is necessary to apprehend, when downloading a dataset, what is the possible error made
 405 by using only the dominant eye for instance.

406 4.1.2. Similarity of human saliency maps

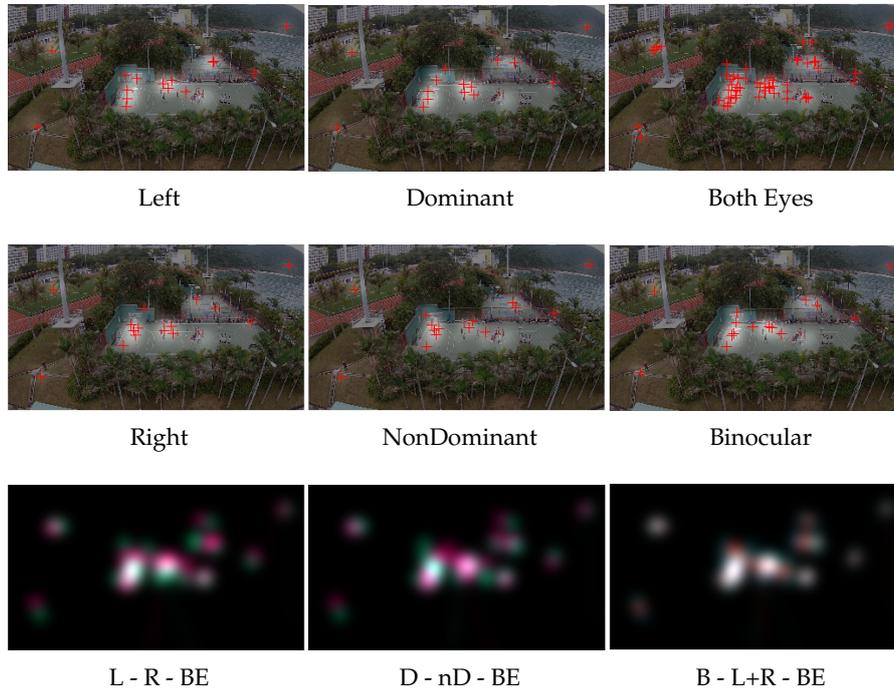


Figure 5. Qualitative comparison of saliency maps for all scenarios on Basketball, frame 401. Saliency and fixations are displayed in transparency over the content. The last row compares scenarios: first scenario is attributed to the red channel, the second to green and the last to blue. When fully overlapping,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

407 Generating saliency maps implies Gaussian filtering with a rather large kernel when compared to
 408 MAE values. For instance, the Gaussian kernel we used here is 3° . We thus question whether selecting
 409 a ground truth scenario over another makes a significant difference for saliency studies.

410 Thus, we decided to compare saliency maps generated for the six scenarios defined above.
 411 Illustrations of scenarios saliency maps and fixations as well as methods comparisons are presented in
 412 Figure 5. Below is presented the quantitative evaluation.

413 We run a cross-comparison on six well-used quality metrics: Correlation Coefficient (CC) ($\in [-1, 1]$),
 414 Similarity (SIM) ($\in [0, 1]$) the intersection between histograms of saliency, AUC Judd and Borji ($\in [0, 1]$),
 415 NSS ($\in]-\infty, +\infty[$), and IG ($\in [0, +\infty[$), which measures on average the gain in information contained in the
 416 saliency map compared to a prior baseline ($\in [0, +\infty[$). We did not report Kullback Leibler divergence (KL)
 417 ($\in [0, +\infty[$) as we favored symmetric metrics. Moreover, even though symmetric in absolute value, IG
 418 provides different scores depending on fixation maps. We thus compared scenarios for fixation maps
 419 of both methods, which leads to two IG measures. More details on metrics and metrics behaviors are
 420 given in [60,74,75].

421 To verify if scenarios are different, we have conducted ANOVA and multi-comparison analyses
 422 on the scores obtained by measures. All metrics show statistically different results ($p \ll 0.001$) except
 423 for AUC Borji, AUC Judd, and NSS. Thus, the analysis discards those metrics. However, it shows that
 424 for those metrics using a scenario over another makes no significant difference.

425 Table 6 presents the results of measures that present a significant difference when comparing
 426 saliency maps of two scenarios. We confirm here the results hinted by MAE scores:

- 427 • There is a high similarity between scenarios saliency maps. As expected, scores are pretty high
 428 (or low for IG), which indicates the high similarity between scenarios.

SM1	SM2	FV				Task			
		CC ↑	SIM ↑	IG ↓		CC ↑	SIM ↑	IG ↓	
				SM1-Fix1-SM2	SM2-Fix2-SM1			SM1-Fix1-SM2	SM2-Fix2-SM1
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	0,98	0,90	0,246	0,139	0,983	0,916	0,177	0,012
Dominant	BothEyes	0,96	0,87	0,158	0,374	0,967	0,873	0,143	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	0,85	0,72	1,176	1,121	0,850	0,725	0,877	0,782
	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 6. CC, SIM and IG results for scenarios cross-comparison. Red indicates the best scores, blue the least.

- 429 • Two-eyes-based saliency maps reach the best results. All metrics show the best results
430 for comparisons including Binocular and BothEyes scenarios, the highest being the
431 Binocular-BothEyes comparison.
432 • Left-Right and Dominant-NonDominant comparisons achieve worst results.
433 • It is possible to know the population main dominant eye through scenarios comparisons (not
434 including two eyes information). When describing the population, we have seen that a majority
435 of left-dominant-eye subjects participated in the FV test, while the reverse happened for the Task
436 experiment. This fact is easy to notice in metric scores.

437 ANOVA and multi-comparison analyses characterize the differences between scenarios in Figure
438 6. We can see where stands the mean and standard deviation of CC scores for each scenario over the
439 entire dataset. Scenarios having non-overlapping confidence intervals are statistically different.

440 We can see that for this metric, and this result applies to SIM and IG, it is recommended to
441 use both eyes information, BothEyes and Binocular, as significantly higher similarity is reached for
442 these two scenarios. The worst scenario, significantly in Task, is to favor the position of one eye over
443 the other. Also, scenarios based on the dominant eye are obviously biased towards one eye, thus
444 generating more errors than two-eyes but less than one-eye scenarios.

445 Overall, over six metrics, three do not find significant differences between the scenarios' saliency
446 maps. The three others do and indicate that using both eye information must be favored. **Accordingly,**
447 **the cyclopean theory takes a slight precedence over the ocular dominance theory in salience.**
448 Moreover, it is recommended to favor datasets that record both eyes, and if not possible these that
449 collect the dominant eye positions.

450 4.2. Biases in UAV videos

451 In conventional imaging, the center position is the best location to have access to most visual
452 information of a content [70]. This fact leads to a well-known bias in visual attention named central
453 bias. This effect may be associated with various causes. For instance, Tseng et al. [76] showed a
454 contribution of photographer bias, viewing strategy, and to a lesser extent, motor, re-centering, and
455 screen center biases to the center bias. They are briefly described below:

- 456 • The **photographer bias** often emphasizes objects in the content center through composition and
457 artistic intent [76].
458 • Directly related to photographer bias, observers tend to learn the probability of finding salient
459 objects at the content center. We refer to this behavior as a **viewing strategy**.

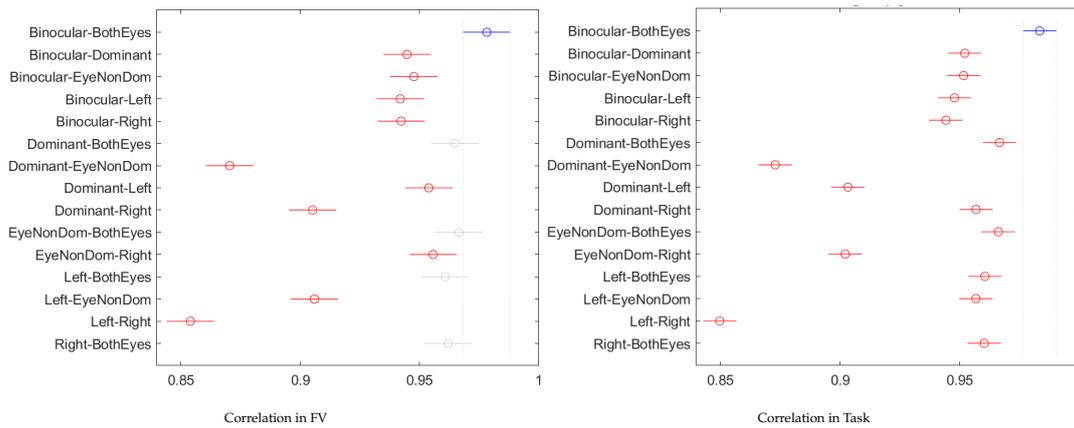


Figure 6. Multi-comparison on scenarios correlation measure.

- 460 • With regards to the Human Visual System (HVS), the central orbital position, that is when
 461 looking straight ahead, is the most comfortable eye position [77], leading to a **recentering bias**.
 462 • Additionally, there is a **motor bias**, in which one prefers making short saccades and horizontal
 463 displacements [59,78].
 464 • Lastly, onscreen presentation of visual content pushes observers to stare at the center of the
 465 screen frame [70]. This experimental bias is named the **screen center bias**.

466 The central bias is so critical in the computational modelling of visual attention that saliency
 467 models include this bias as prior knowledge or use it as a baseline to which saliency predictions are
 468 being compared. The center bias is often represented by a centered isotropic Gaussian stretched to the
 469 video frame aspect ratio [25,75].

470 The presence of this bias in UAV videos has already been questioned in [23]. In [23], authors
 471 showed that saliency models that heavily rely on the center bias were less efficient on UAV videos than
 472 on conventional video sequences. Therefore, we believe that the central bias could be less significant in
 473 drone footage as a result of the lack of photographer bias or due to UAV content characteristics. These
 474 latter comprise, but are not restricted to, the camera bird-point-of-view that changes objects semantic
 475 and size [11], the loss of pictorial depth cues [79] such as horizontal line [80], and the presence of camera
 476 movements [11]. To make this point clear, we propose to evaluate qualitatively and quantitatively the
 477 center bias for UAV videos.

478 4.2.1. Qualitative evaluation of biases in UAV videos

479 We evaluate the viewing tendency of observer thanks to the average saliency map, computed over
 480 the entire sequence. It is representative of the average position of gaze throughout the video sequence.
 481 It is used to observe potential overall biases, as it could be the case with the center bias. Figures 7 and 8
 482 show the average saliency map for all sequences of *EyeTrackUAV2* dataset, generated from D scenario,
 483 for both free-viewing and task-viewing conditions. Several observations can be made.

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

491 **Original database-specific center bias.** We can observe that a center bias is present in VIRAT
 492 sequences, while videos from other datasets, namely UAV123 and DTB70, do not present this systematic
 493 bias. The original resolution of content and the experimental setup are possibly the sources of this

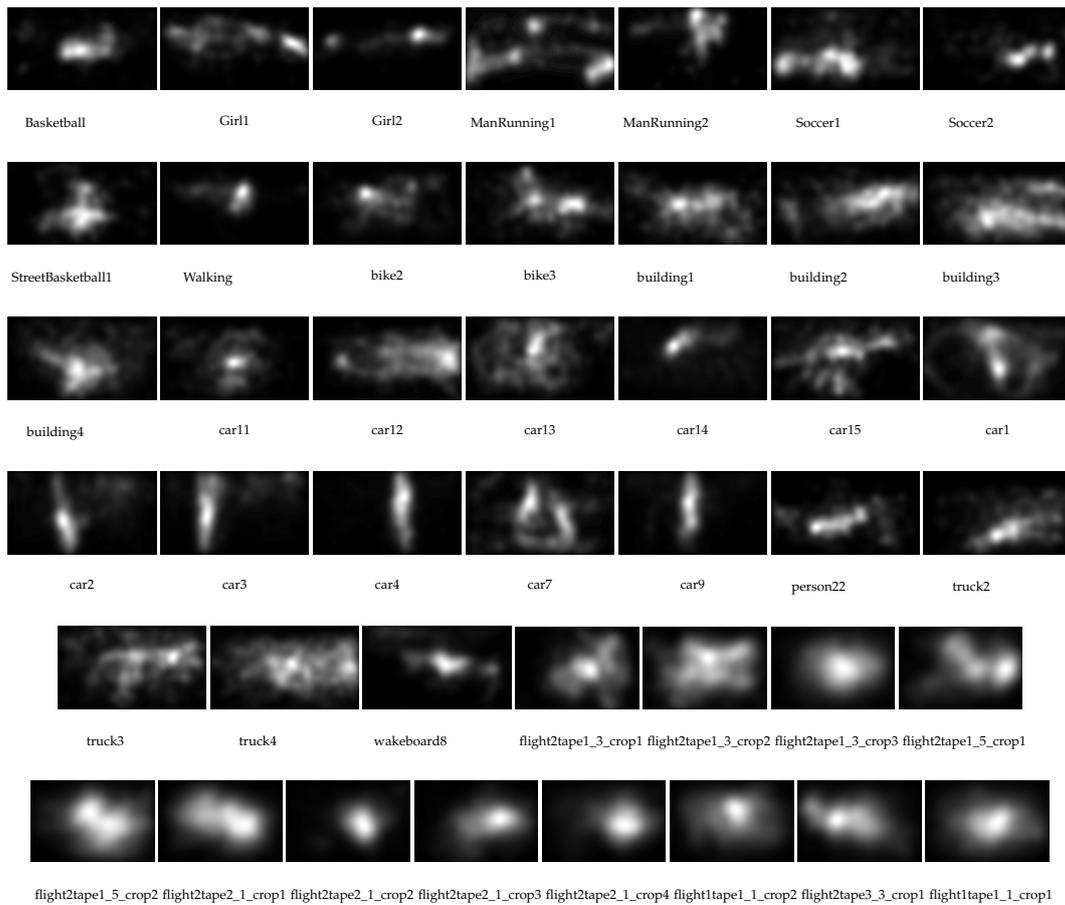


Figure 7. Average saliency maps for all sequences of *EyeTrackUAV2* dataset, generated from D scenario, for the free-viewing condition.

494 result. Indeed, the proportion of content seen at once is not the same for all sequences: 1,19% of
 495 a VIRAT content is seen per degree of visual angle, whereas it is 0,44% for the two other original
 496 databases. VIRAT saliency maps are thus smoother, which results in higher chances to present a center
 497 bias. To verify this assumption based on qualitative assessment, we have computed the overall human
 498 saliency maps for sequences coming from original dataset, namely DTB70, UAV123 and VIRAT. These
 499 maps are shown in Figure 9. VIRAT saliency maps are much more concentrated and centered. This
 500 corroborates that biases can be original-database-specific.

501 **Task-related human saliency maps are more spread out.** Task-based saliency maps cover more
 502 content when compared to free-viewing condition for most sequences (e.g. in about 58% of videos
 503 such as *Basketball*, *car11*, *car2*, and *wakeboard*). This behavior is also illustrated in Figure 9. We correlate
 504 this response with the object detection task. Visual search scanning implies an extensive exploration of
 505 the content. However, 21% of the remaining sequences (i.e. *soccer1*, *bike2-3*, *building 1-2*, *car1,15*, and
 506 *truck3-4*) show less discrepancies in the task-viewing condition than in free-viewing condition. We do
 507 not find correlation between such behavior and sequences characteristics given in Table 1. This leaves
 508 room for further exploration of differences between task-based and free viewing attention.

509 **Overall, there is no generalization of center bias for UAV content.** As stated earlier, we do not
 510 observe a systematic center bias, except for VIRAT sequences. This is especially true for task-related
 511 viewing. However, we observe specific patterns. Indeed, vertical and horizontal potatoe-shaped
 512 saliency areas are quite present in average human saliency maps of *EyeTrackUAV2*. Such patterns
 513 are also visible in UAV2 and DTB70 overall mean maps, especially in task-viewing condition. This
 514 indicates future axes of developments for UAV saliency-based applications. For instance, instead of

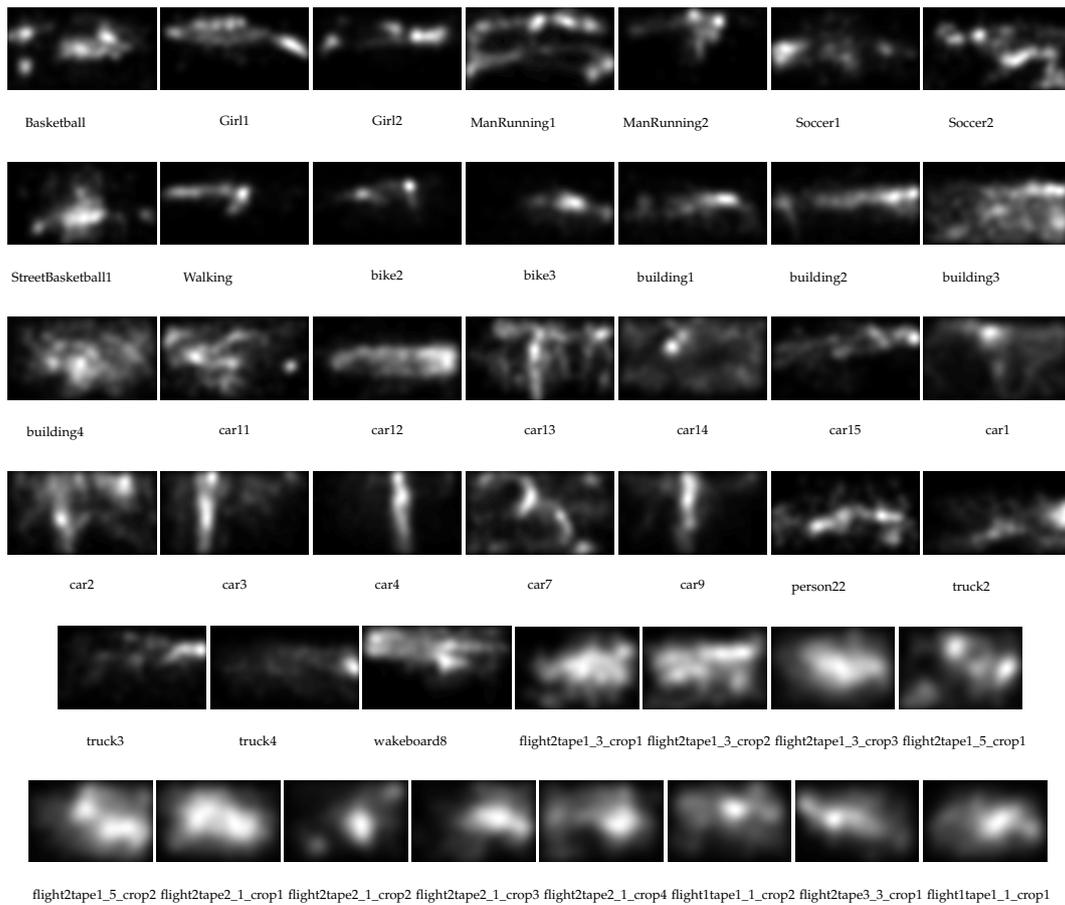


Figure 8. Average saliency maps for all sequences of *EyeTrackUAV2* dataset, generated from D scenario, for the task-viewing condition.

515 using a center bias, one may introduce priors as a set of prevalent saliency area shapes with different
 516 directions and sizes [81].

517 4.2.2. Quantitative evaluation of the central bias in UAV videos

518 To go further into content-dependencies, we investigate quantitatively the similarity of
 519 dominant-eye-generated saliency maps, called ground truth in the remaining, with a pre-defined
 520 center bias. Figure 10 presents the center bias baseline created in this purpose as suggested in [25,75].

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

530 5. Conclusion

531 UAV imaging modifies the perceptual clues of typical scenes due to its bird point of view, the
 532 presence of camera movements and the high distance and angle to the scene. For instance, low-level
 533 visual features and size of objects changes and depth information is flattened or disappears (e.g.

	FV				Task			
	CC ↑	SIM ↑	KL ↓	IG ↓	CC ↑	SIM ↑	KL ↓	IG ↓
VIRAT_09152008flight2tape1_3_crop1	0,50	0,48	7,17	1,53	0,46	0,48	6,85	1,62
VIRAT_09152008flight2tape1_3_crop2	0,49	0,52	5,59	1,50	0,36	0,48	6,42	1,75
VIRAT_09152008flight2tape1_3_crop3	0,46	0,43	8,46	1,91	0,37	0,43	7,98	1,99
VIRAT_09152008flight2tape1_5_crop1	0,27	0,38	9,77	2,29	0,18	0,36	10,14	2,49
VIRAT_09152008flight2tape1_5_crop2	0,42	0,44	8,05	1,90	0,30	0,45	7,41	1,87
VIRAT_09152008flight2tape2_1_crop1	0,41	0,39	9,34	2,05	0,38	0,42	8,55	1,97
VIRAT_09152008flight2tape2_1_crop2	0,40	0,35	10,90	2,50	0,32	0,42	8,01	2,01
VIRAT_09152008flight2tape2_1_crop3	0,42	0,40	9,46	2,11	0,28	0,39	9,30	2,24
VIRAT_09152008flight2tape2_1_crop4	0,36	0,36	10,35	2,34	0,28	0,38	9,79	2,30
VIRAT_09152008flight2tape3_3_crop1	0,42	0,43	8,16	1,96	0,35	0,43	7,84	2,03
VIRAT_09162008flight1tape1_1_crop1	0,47	0,45	7,76	1,80	0,37	0,42	8,40	2,00
VIRAT_09162008flight1tape1_1_crop2	0,40	0,40	9,14	2,14	0,27	0,40	8,91	2,22
UAV123_bike2	0,39	0,34	11,51	2,43	0,34	0,29	13,21	2,82
UAV123_bike3	0,39	0,34	11,71	2,37	0,29	0,26	14,34	2,96
UAV123_building1	0,40	0,37	10,64	2,18	0,32	0,31	12,74	2,69
UAV123_building2	0,30	0,33	11,89	2,43	0,18	0,27	13,87	3,06
UAV123_building3	0,27	0,34	11,50	2,42	0,17	0,32	11,82	2,56
UAV123_building4	0,39	0,36	10,82	2,20	0,35	0,39	9,72	2,10
UAV123_car11	0,37	0,32	12,37	2,58	0,21	0,30	12,68	2,67
UAV123_car12	0,21	0,28	13,35	2,80	0,26	0,29	13,12	2,69
UAV123_car13	0,30	0,34	11,48	2,39	0,20	0,33	11,50	2,44
UAV123_car14	0,20	0,25	14,47	3,16	0,12	0,31	12,28	2,71
UAV123_car15	0,31	0,34	11,52	2,47	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	0,35	0,31	12,44	2,60	0,27	0,31	12,45	2,68
UAV123_truck2	0,27	0,32	12,29	2,56	0,09	0,27	13,66	3,01
UAV123_truck3	0,27	0,35	11,14	2,34	0,12	0,31	12,23	2,73
UAV123_truck4	0,29	0,36	10,71	2,34	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	0,38	0,27	14,13	2,89	0,30	0,31	12,30	2,59
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	0,32	11,86	2,69
DTB70_StreetBasketball1	0,33	0,26	14,29	2,94	0,28	0,26	14,29	3,00
DTB70_Walking	0,29	0,20	16,14	3,51	0,27	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 saliency 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.

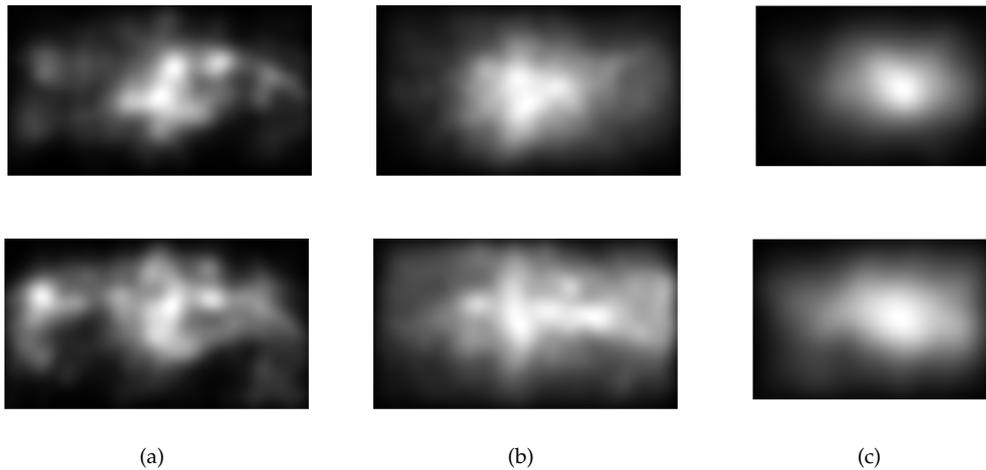


Figure 9. Overall average saliency maps per original dataset, generated from D scenario, in free-viewing (top-row) and Task-viewing (bottom row) for original datasets: (a) DTB70; (b) UAV123; (c) VIRAT.

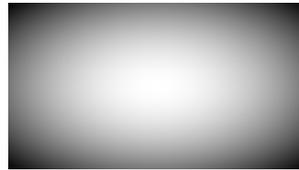


Figure 10. Center prior baseline.

534 presence of sky). To understand observers' behaviors toward these new features, especially in terms of
 535 visual attention and deployment, there is a need for large-scale eye-tracking databases for saliency in
 536 UAV videos. This dataset is also a key factor in the field of computational models of visual attention,
 537 in which large scale datasets are required to train the latest generation of deep-based models.

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

542 This work addresses the need for such a dedicated dataset. An experimental process has
 543 been designed in order to build a new dataset, *EyeTrackUAV2*. Gaze data were collected during
 544 the observation of UAV videos under controlled laboratory conditions for both free viewing and
 545 object-detection surveillance task conditions. Gaze positions have been collected on 30 participants
 546 for each attention condition, on 43 UAV videos in 30 fps, 1280x720 or 720x480, consisting in 42 241
 547 frames and 1408 seconds. Overall, 1 239 157 fixations in free-viewing and 1 269 433 in task-viewing
 548 were extracted from the dominant eye positions. Test stimuli were carefully selected from three
 549 original datasets, i.e. UAV123, VIRAT, and DTB70, to be representative as much as possible of the UAV
 550 ecosystem. Accordingly, they present variations in terms of environments, camera movement, size of
 551 objects, aircraft flight heights and angles to the ground, video duration, resolution, and quality. Also,
 552 three sequences were recorded in infra-red.

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

559 Then, we further characterized the dataset considering two different aspects. On one hand,
560 six scenarios, namely binocular, both eyes, dominant eye, non-dominant eye, left, and right can be
561 envisioned to generate human saliency maps. We wondered whether a scenario should be favored over
562 another or not. Comparisons of scenarios have been conducted, first in terms of the mean of absolute
563 errors for eye positions, and secondly on six typical saliency metrics for saliency maps. Results indicate
564 that the cyclopean theory prevails over the ocular dominance theory in saliency. That means that
565 information of both eyes should be favored to study saliency. If not possible, choosing information
566 from the dominant eye allows us to commit fewer errors when compared to other one-eye scenarios.
567 On the other hand, we notice that conventional biases in saliency do not necessarily apply to UAV
568 content. Indeed, the center bias is not systematic in UAV sequences. This bias is content-dependent as
569 well as and task-condition-dependent. We observed new prior patterns that must be examined in the
570 future.

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

575 6. Acknowledgment

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

581

- 582 1. Zhao, Y.; Ma, J.; Li, X.; Zhang, J. Saliency detection and deep learning-based wildfire identification in UAV
583 imagery. *Sensors* **2018**, *18*, 712.
- 584 2. van Gemert, J.C.; Verschoor, C.R.; Mettes, P.; Epema, K.; Koh, L.P.; Wich, S. Nature conservation drones
585 for automatic localization and counting of animals. Workshop at the European Conference on Computer
586 Vision. Springer, 2014, pp. 255–270.
- 587 3. Postema, S. News Drones: An Auxiliary Perspective, 2015.
- 588 4. Agbeyangi, A.O.; Odiete, J.O.; Olorunlome, A.B. Review on UAVs used for aerial surveillance. *Journal*
589 *of Multidisciplinary Engineering Science and Technology (JMEST)* **2016**, *3*, 5713–5719.
- 590 5. Lee-Morrison, L. *State of the Art Report on Drone-Based Warfare*; Citeseer, 2014.
- 591 6. Zhou, Y.; Tang, D.; Zhou, H.; Xiang, X.; Hu, T. Vision-Based Online Localization and Trajectory Smoothing
592 for Fixed-Wing UAV Tracking a Moving Target. Proceedings of the IEEE International Conference on
593 Computer Vision Workshops, 2019, pp. 0–0.
- 594 7. Zhu, P.; Du, D.; Wen, L.; Bian, X.; Ling, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; others.
595 VisDrone-VID2019: The Vision Meets Drone Object Detection in Video Challenge Results. Proceedings of
596 the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.
- 597 8. Aguilar, W.G.; Luna, M.A.; Moya, J.F.; Abad, V.; Ruiz, H.; Parra, H.; Angulo, C. Pedestrian detection for
598 UAVs using cascade classifiers and saliency maps. International Work-Conference on Artificial Neural
599 Networks. Springer, 2017, pp. 563–574.
- 600 9. Dang, T.; Khattak, S.; Papachristos, C.; Alexis, K. Anomaly Detection and Cognizant Path Planning
601 for Surveillance Operations using Aerial Robots. 2019 International Conference on Unmanned Aircraft
602 Systems (ICUAS). IEEE, 2019, pp. 667–673.
- 603 10. Edney-Browne, A. Vision, visuality, and agency in the US drone program. *Technology and Agency in*
604 *International Relations* **2019**, p. 88.
- 605 11. Krassanakis, V.; Perreira Da Silva, M.; Ricordel, V. Monitoring Human Visual Behavior during the
606 Observation of Unmanned Aerial Vehicles (UAVs) Videos. *Drones* **2018**, *2*, 36.

- 607 12. Papachristos, C.; Khattak, S.; Mascarich, F.; Dang, T.; Alexis, K. Autonomous Aerial Robotic Exploration of
608 Subterranean Environments relying on Morphology-aware Path Planning. 2019 International Conference
609 on Unmanned Aircraft Systems (ICUAS). IEEE, 2019, pp. 299–305.
- 610 13. Itti, L.; Koch, C. Computational modelling of visual attention. *Nature reviews neuroscience* **2001**, *2*, 194.
- 611 14. Katsuki, F.; Constantinidis, C. Bottom-up and top-down attention: different processes and overlapping
612 neural systems. *The Neuroscientist* **2014**, *20*, 509–521.
- 613 15. Krasovskaya, S.; MacInnes, W.J. Saliency Models: A Computational Cognitive Neuroscience Review. *Vision*
614 **2019**, *3*, 56.
- 615 16. Kummerer, M.; Wallis, T.S.; Bethge, M. Saliency benchmarking made easy: Separating models, maps and
616 metrics. Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 770–787.
- 617 17. Riche, N.; Duvinage, M.; Mancas, M.; Gosselin, B.; Dutoit, T. Saliency and human fixations: State-of-the-art
618 and study of comparison metrics. Proceedings of the IEEE international conference on computer vision,
619 2013, pp. 1153–1160.
- 620 18. Guo, C.; Zhang, L. A novel multiresolution spatiotemporal saliency detection model and its applications
621 in image and video compression. *IEEE transactions on image processing* **2009**, *19*, 185–198.
- 622 19. Jain, S.D.; Xiong, B.; Grauman, K. Fusionseg: Learning to combine motion and appearance for fully
623 automatic segmentation of generic objects in videos. 2017 IEEE conference on computer vision and pattern
624 recognition (CVPR). IEEE, 2017, pp. 2117–2126.
- 625 20. Wang, W.; Shen, J.; Shao, L. Video salient object detection via fully convolutional networks. *IEEE*
626 *Transactions on Image Processing* **2017**, *27*, 38–49.
- 627 21. Li, G.; Xie, Y.; Wei, T.; Wang, K.; Lin, L. Flow guided recurrent neural encoder for video salient object
628 detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp.
629 3243–3252.
- 630 22. Le Meur, O.; Coutrot, A.; Liu, Z.; Rämä, P.; Le Roch, A.; Helo, A. Visual attention saccadic models
631 learn to emulate gaze patterns from childhood to adulthood. *IEEE Transactions on Image Processing* **2017**,
632 *26*, 4777–4789.
- 633 23. Perrin, A.F.; Zhang, L.; Le Meur, O. How well current saliency prediction models perform on UAVs videos?
634 International Conference on Computer Analysis of Images and Patterns. Springer, 2019, pp. 311–323.
- 635 24. Paglin, M.; Rufolo, A.M. Heterogeneous human capital, occupational choice, and male-female earnings
636 differences. *Journal of Labor Economics* **1990**, *8*, 123–144.
- 637 25. Le Meur, O.; Le Callet, P.; Barba, D.; Thoreau, D. A coherent computational approach to model bottom-up
638 visual attention. *IEEE transactions on pattern analysis and machine intelligence* **2006**, *28*, 802–817.
- 639 26. Ehinger, K.A.; Hidalgo-Sotelo, B.; Torralba, A.; Oliva, A. Modelling search for people in 900 scenes: A
640 combined source model of eye guidance. *Visual cognition* **2009**, *17*, 945–978.
- 641 27. Liu, H.; Heynderickx, I. Studying the added value of visual attention in objective image quality metrics
642 based on eye movement data. 2009 16th IEEE international conference on image processing (ICIP). IEEE,
643 2009, pp. 3097–3100.
- 644 28. Judd, T.; Durand, F.; Torralba, A. A benchmark of computational models of saliency to predict human
645 fixations, 2012.
- 646 29. Ma, K.T.; Sim, T.; Kankanhalli, M. VIP: A unifying framework for computational eye-gaze research.
647 International Workshop on Human Behavior Understanding. Springer, 2013, pp. 209–222.
- 648 30. Koehler, K.; Guo, F.; Zhang, S.; Eckstein, M.P. What do saliency models predict? *Journal of vision* **2014**,
649 *14*, 14–14.
- 650 31. Borji, A.; Itti, L. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint*
651 *arXiv:1505.03581* **2015**.
- 652 32. Bylinskii, Z.; Isola, P.; Bainbridge, C.; Torralba, A.; Oliva, A. Intrinsic and extrinsic effects on image
653 memorability. *Vision research* **2015**, *116*, 165–178.
- 654 33. Fan, S.; Shen, Z.; Jiang, M.; Koenig, B.L.; Xu, J.; Kankanhalli, M.S.; Zhao, Q. Emotional attention: A study
655 of image sentiment and visual attention. Proceedings of the IEEE Conference on Computer Vision and
656 Pattern Recognition, 2018, pp. 7521–7531.
- 657 34. McCamy, M.B.; Otero-Millan, J.; Di Stasi, L.L.; Macknik, S.L.; Martinez-Conde, S. Highly informative
658 natural scene regions increase microsaccade production during visual scanning. *Journal of neuroscience*
659 **2014**, *34*, 2956–2966.

- 660 35. Gitman, Y.; Erofeev, M.; Vatolin, D.; Andrey, B.; Alexey, F. Semiautomatic visual-attention modeling and its
661 application to video compression. 2014 IEEE International Conference on Image Processing (ICIP). IEEE,
662 2014, pp. 1105–1109.
- 663 36. Coutrot, A.; Guyader, N. How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal*
664 *of vision* **2014**, *14*, 5–5.
- 665 37. Coutrot, A.; Guyader, N. An efficient audiovisual saliency model to predict eye positions when looking at
666 conversations. 2015 23rd European Signal Processing Conference (EUSIPCO). IEEE, 2015, pp. 1531–1535.
- 667 38. Wang, W.; Shen, J.; Xie, J.; Cheng, M.M.; Ling, H.; Borji, A. Revisiting video saliency prediction in the deep
668 learning era. *IEEE transactions on pattern analysis and machine intelligence* **2019**.
- 669 39. Oh, S.; Hoogs, A.; Perera, A.; Cuntoor, N.; Chen, C.C.; Lee, J.T.; Mukherjee, S.; Aggarwal, J.; Lee, H.; Davis,
670 L.; others. A large-scale benchmark dataset for event recognition in surveillance video. CVPR 2011. IEEE,
671 2011, pp. 3153–3160.
- 672 40. Layne, R.; Hospedales, T.M.; Gong, S. Investigating open-world person re-identification using a drone.
673 European Conference on Computer Vision. Springer, 2014, pp. 225–240.
- 674 41. Bonetto, M.; Korshunov, P.; Ramponi, G.; Ebrahimi, T. Privacy in mini-drone based video surveillance.
675 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG).
676 IEEE, 2015, Vol. 4, pp. 1–6.
- 677 42. Shu, T.; Xie, D.; Rothrock, B.; Todorovic, S.; Chun Zhu, S. Joint inference of groups, events and human
678 roles in aerial videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,
679 2015, pp. 4576–4584.
- 680 43. Mueller, M.; Smith, N.; Ghanem, B. A benchmark and simulator for uav tracking. European conference on
681 computer vision. Springer, 2016, pp. 445–461.
- 682 44. Robicquet, A.; Sadeghian, A.; Alahi, A.; Savarese, S. Learning social etiquette: Human trajectory
683 understanding in crowded scenes. European conference on computer vision. Springer, 2016, pp. 549–565.
- 684 45. Li, S.; Yeung, D.Y. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion
685 models. Thirty-First AAAI Conference on Artificial Intelligence, 2017.
- 686 46. Barekatin, M.; Martí, M.; Shih, H.F.; Murray, S.; Nakayama, K.; Matsuo, Y.; Prendinger, H. Okutama-action:
687 An aerial view video dataset for concurrent human action detection. Proceedings of the IEEE Conference
688 on Computer Vision and Pattern Recognition Workshops, 2017, pp. 28–35.
- 689 47. Hsieh, M.R.; Lin, Y.L.; Hsu, W.H. Drone-based object counting by spatially regularized regional proposal
690 network. Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4145–4153.
- 691 48. Ribeiro, R.; Cruz, G.; Matos, J.; Bernardino, A. A dataset for airborne maritime surveillance environments.
692 *IEEE Trans. Circuits Syst. Video Technol* **2017**.
- 693 49. Hsu, H.J.; Chen, K.T. DroneFace: an open dataset for drone research. Proceedings of the 8th ACM on
694 Multimedia Systems Conference. ACM, 2017, pp. 187–192.
- 695 50. Božić-Štulić, D.; Marušić, Ž.; Gotovac, S. Deep Learning Approach in Aerial Imagery for Supporting Land
696 Search and Rescue Missions. *International Journal of Computer Vision* **2019**, pp. 1–23.
- 697 51. Fu, K.; Li, J.; Shen, H.; Tian, Y. How drones look: Crowdsourced knowledge transfer for aerial video
698 saliency prediction. *arXiv preprint arXiv:1811.05625* **2018**.
- 699 52. Zhu, P.; Wen, L.; Bian, X.; Ling, H.; Hu, Q. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*
700 **2018**.
- 701 53. Nyström, M.; Andersson, R.; Holmqvist, K.; Van De Weijer, J. The influence of calibration method and eye
702 physiology on eyetracking data quality. *Behavior research methods* **2013**, *45*, 272–288.
- 703 54. ITU-T RECOMMENDATION, P. Subjective video quality assessment methods for multimedia applications.
704 *International telecommunication union* **2008**.
- 705 55. Rec, I. BT. 710-4. *Subjective assessment methods for image quality in high-definition television* **1998**.
- 706 56. Cornelissen, F.W.; Peters, E.M.; Palmer, J. The Eyelink Toolbox: eye tracking with MATLAB and the
707 Psychophysics Toolbox. *Behavior Research Methods, Instruments, & Computers* **2002**, *34*, 613–617.
- 708 57. Rec, I. BT. 500-13. *Methodology for the subjective assessment of the quality of television pictures* **2012**.
- 709 58. Wandell, B.; Thomas, S. Foundations of vision. *Psychocritiques* **1997**, *42*.
- 710 59. Le Meur, O.; Liu, Z. Saccadic model of eye movements for free-viewing condition. *Vision research* **2015**,
711 *116*, 152–164.

- 712 60. Le Meur, O.; Baccino, T. Methods for comparing scanpaths and saliency maps: strengths and weaknesses.
713 *Behavior research methods* **2013**, *45*, 251–266.
- 714 61. Guznov, S.; Matthews, G.; Warm, J.S.; Pfahler, M. Training techniques for visual search in complex task
715 environments. *Human factors* **2017**, *59*, 1139–1152.
- 716 62. Shah, M.; Javed, O.; Shafique, K. Automated visual surveillance in realistic scenarios. *IEEE MultiMedia*
717 **2007**, *14*, 30–39.
- 718 63. Snellen, H. *Test-types for the determination of the acuteness of vision*; Williams and Norgate, 1868.
- 719 64. Ishihara, S. *Test for colour-blindness*; Kanehara Tokyo, Japan, 1987.
- 720 65. Salvucci, D.D.; Goldberg, J.H. Identifying fixations and saccades in eye-tracking protocols. Proceedings of
721 the 2000 symposium on Eye tracking research & applications. ACM, 2000, pp. 71–78.
- 722 66. Krassanakis, V.; Filippakopoulou, V.; Nakos, B. EyeMMV toolbox: An eye movement post-analysis tool
723 based on a two-step spatial dispersion threshold for fixation identification. *Journal of eye movement research*
724 **2014**.
- 725 67. Krassanakis, V.; Misthos, L.M.; Menegaki, M. LandRate toolbox: An adaptable tool for eye movement
726 analysis and landscape rating. Eye Tracking for Spatial Research, Proceedings of the 3rd International
727 Workshop. ETH Zurich, 2018.
- 728 68. Krassanakis, V.; Filippakopoulou, V.; Nakos, B. Detection of moving point symbols on cartographic
729 backgrounds. *Journal of Eye Movement Research* **2016**, *9*.
- 730 69. Ooms, K.; Krassanakis, V. Measuring the Spatial Noise of a Low-Cost Eye Tracker to Enhance Fixation
731 Detection. *Journal of Imaging* **2018**, *4*, 96.
- 732 70. Bindemann, M. Scene and screen center bias early eye movements in scene viewing. *Vision research* **2010**,
733 *50*, 2577–2587.
- 734 71. Cui, Y.; Hondzinski, J.M. Gaze tracking accuracy in humans: Two eyes are better than one. *Neuroscience*
735 *letters* **2006**, *396*, 257–262.
- 736 72. Holmqvist, K.; Nyström, M.; Mulvey, F. Eye tracker data quality: what it is and how to measure it.
737 Proceedings of the symposium on eye tracking research and applications. ACM, 2012, pp. 45–52.
- 738 73. Hooge, I.T.; Holleman, G.A.; Haukes, N.C.; Hessels, R.S. Gaze tracking accuracy in humans: One eye is
739 sometimes better than two. *Behavior Research Methods* **2018**, pp. 1–10.
- 740 74. Bylinskii, Z.; Judd, T.; Borji, A.; Itti, L.; Durand, F.; Oliva, A.; Torralba, A. Mit saliency benchmark, 2015.
- 741 75. Bylinskii, Z.; Judd, T.; Oliva, A.; Torralba, A.; Durand, F. What Do Different Evaluation Metrics Tell Us
742 About Saliency Models? *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2019**, *41*, 740–757.
- 743 76. Tseng, P.H.; Carmi, R.; Cameron, I.G.; Munoz, D.P.; Itti, L. Quantifying center bias of observers in free
744 viewing of dynamic natural scenes. *Journal of vision* **2009**, *9*, 4–4.
- 745 77. Van Opstal, A.; Hepp, K.; Suzuki, Y.; Henn, V. Influence of eye position on activity in monkey superior
746 colliculus. *Journal of Neurophysiology* **1995**, *74*, 1593–1610.
- 747 78. Tatler, B.W. The central fixation bias in scene viewing: Selecting an optimal viewing position independently
748 of motor biases and image feature distributions. *Journal of vision* **2007**, *7*, 4–4.
- 749 79. Howard, I.P.; Rogers, B. Depth perception. *Stevens Handbook of Experimental Psychology* **2002**, *6*, 77–120.
- 750 80. Foulsham, T.; Kingstone, A.; Underwood, G. Turning the world around: Patterns in saccade direction vary
751 with picture orientation. *Vision research* **2008**, *48*, 1777–1790.
- 752 81. Le Meur, O.; Coutrot, A. Introducing context-dependent and spatially-variant viewing biases in saccadic
753 models. *Vision research* **2016**, *121*, 72–84.