Supplementary results for the ECCV'14 paper "Predicting Actions from Static Scenes"

Tuan-Hung Vu[1], Catherine Olsson[2], Ivan Laptev[1], Aude Oliva[2] and Josef Sivic[1]

[1]WILLOW, ENS/INRIA/CNRS UMR 8548, Paris, France,
[2]CSAIL, MIT, Cambridge, Massachusetts, USA

Paper link: `http://www.di.ens.fr/willow/pdfscurrent/Vu14eccv.pdf`

Dense geo-localized prediction of actions

Section 6 of the ECCV'14 paper demonstrates a successful application of the Image-based Geo-Mapping of Action (IGMA) to the country-scale prediction of certain actions such as "swim", "hike", etc. Prediction of other actions such as "drive", however, require much denser sampling of images

As demonstrated in Section 6 of the paper, the success of the Image-based Geo-Mapping of Action application (IGMA) roves that we could apply action predictors on scene images of country-size regions like France. For such large-scale areas, roughly knowing what type of activity people could probably do, for example "ski", "swim" or "hike", is preferable. However, when it comes to smaller-scale region, we urge for more information, i.e, how the actions are performed? For example, on a rough forest trail, people cycle more likely with a mountain bike than with a city bike. The question passionates us to build finer-grained action predictors and investigate how they could perform in real life. To do this, we densely collect trail images of Fontainebleau forest, annotate them with our predefined set of fine-grained actions, and use the same learning approach in the paper to build the action predictors. Section 1 of this document introduces the dataset and the action labels. In Section 2, we briefly show the experimental details and some quantitative results. In Section 3, we compare predicted action probability to the probability of action derived from manual image annotation, and show example images corresponding to high and low probabilities of each action.

Contents

1 Fontainebleau scene-action dataset **1**

2 Experiments **2**

2.1 Implementation details 2

2.2 Quantitative results 2

2.3 Qualitative results 3

2.3.1 Is it a good place to cycle with a city bike? 3

2.3.2 Is it a good place to cycle with a mountain bike? 5

2.3.3 Is it a good place to drive a Jeep/4x4 car? 6

2.3.4 Is it a good place to drive a regular car? 7

2.3.5 Is it a good place to have a picnic? 8

2.3.6 Is it a good place to jog? 9

2.3.7 Is it a good place to put up a tent for camping? 10

2.3.8 Is it a good place to ride a motorbike? 11

2.3.9 Is it a good place to walk with city shoes? 12

1 Fontainebleau scene-action dataset

The dataset consists of densely taken trail images of Fontainebleau forest, a famous hiking spot in France. Different from SUN Action dataset, we here pre-define a set of action labels X and collect answers for functional questions, such as ”Is it a good place to do X”. For each image and each functional question, the Yes/No answers are obtained from multiple people on the crowdsourcing service AMT. The list of nine action classes together with histograms of Yes/No responses for a typical image from our dataset are illustrated in Figure 1 and Table 1. In total the dataset contains 1084 images and 54400 Yes/No action annotations.

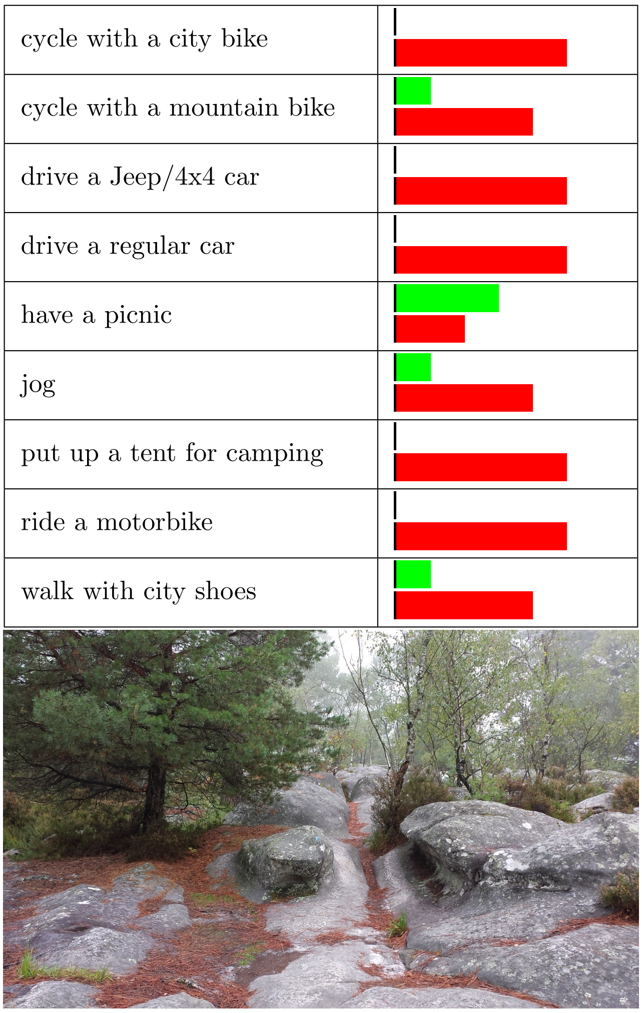


Figure 1: Example of trail image in Fontainebleau and corresponding Yes/No ratio for each action. Green and red color represent ”Yes” and ”No” respectively.

Once the dataset has been collected, to assign action labels, we use the following procedure. From the set of Yes/No responses $R_{Q,I}$ for a functional question Q and image I, we compute the positive ratio:

$$pr(Q, I) = \frac{|R_{Q,I} = \text{"Yes"}|}{|R_{Q,I}|}$$

A threshold is then put on the positive ratio to differentiate positive and negative samples. Trying to alleviate the limitation of hard-thresholding, we allow a ”grey area” for samples which are hard to decide neither as positive nor as negative.

1. cycle with a city bike
2. cycle with a mountain bike
3. drive a Jeep/4x4 car
4. drive a regular car
5. have a picnic
6. jog
7. put up a tent for camping
8. ride a motorbike
9. walk with city shoes

Table 1: List of 9 questions having the same format: "Is it a good place to ..."

2 Experiments

2.1 Implementation details

We split the dataset into the training and test sets such that both of them are equal in the number of images and do not share images from the same trail. Figure 2 illustrates locations of our training and test images by different colors on the map.

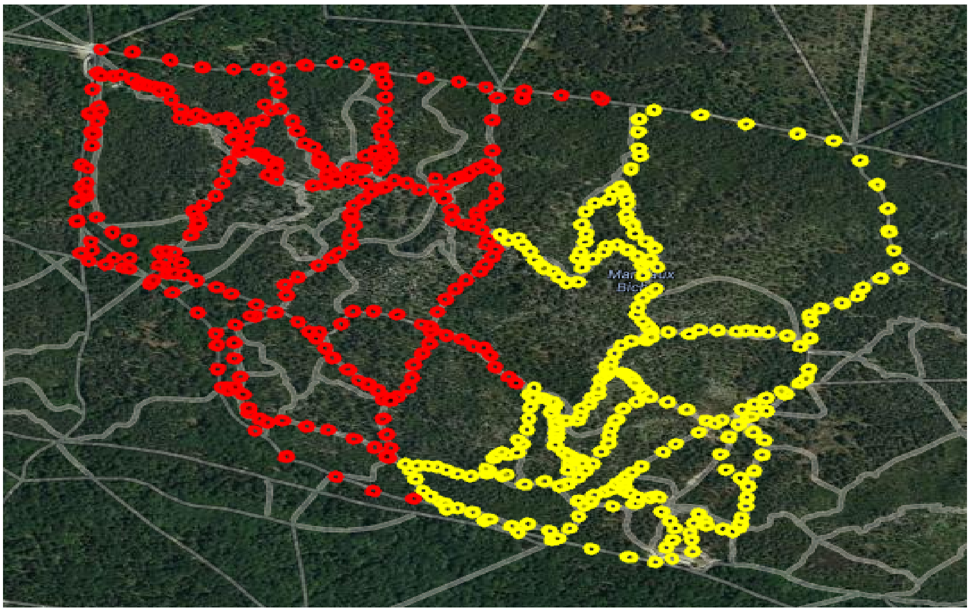


Figure 2: Yellow and red points stand for training and testing samples respectively.

For training and testing, we use image descriptors and classifiers as explained in Section 5.1 of the paper. We also add a new terrain descriptor as follows: we first determine the elevation heat image of examined regions, i.e. Fontainebleau, at multiple scales. At a given scale and a given position, we compute the differences in elevation magnitude with 8 corresponding neighbours, then aggregate them into a 8-dimensional vector. In practice, we consider 3 different scales, which results in 24-dimensional terrain descriptors. To build SVM classifiers using terrain descriptor, we exploit Histogram Intersection kernel.

2.2 Quantitative results

Table 2 compares performance using different descriptors. Although the terrain feature has the lowest accuracy compared to other visual features, itimproves overallperformance when combined with image features. HOG_FV has higher performance compared to other individual features. The best mAP of 91.46% is achieved with the combination of CSIFT_FV, HOG_FV and terrain features.

Method	q1	q2	q3	q4	q5	q6	q7	q8	q9	mAP
Chance level	26.11	73.61	31.12	17.81	34.82	53.26	25.24	40.52	32.04	37.17
Terrain	67.09	86.35	59.00	37.30	40.22	78.86	43.59	70.54	63.72	60.74
CSIFT_FV	90.21	97.11	91.18	90.64	86.06	93.20	75.59	90.60	93.29	89.79
HOG_FV	90.33	96.22	94.10	88.88	82.27	94.83	78.08	93.39	94.16	90.25
CSIFT_FV+Terrain	**91.69**	**97.13**	92.26	90.98	**86.18**	93.44	76.89	91.32	93.37	90.36
HOG_FV+Terrain	90.98	96.32	**94.32**	89.76	82.52	**94.86**	79.16	**93.51**	94.35	90.64
CSIFT_FV+HOG_FV	91.39	97.11	93.86	90.87	84.88	94.76	79.07	93.29	94.99	91.14
CSIFT_FV+ HOG_FV+ Terrain	91.98	97.05	94.17	**91.47**	85.04	94.77	**80.12**	93.48	**95.06**	**91.46**

Table 2: AP(%) per action class and mAP(%) of different methods on Direct action prediction. q1..q9 stand for 9 corresponding questions listed in Table 1

2.3 Qualitative results

2.3.1 Is it a good place to cycle with a city bike?

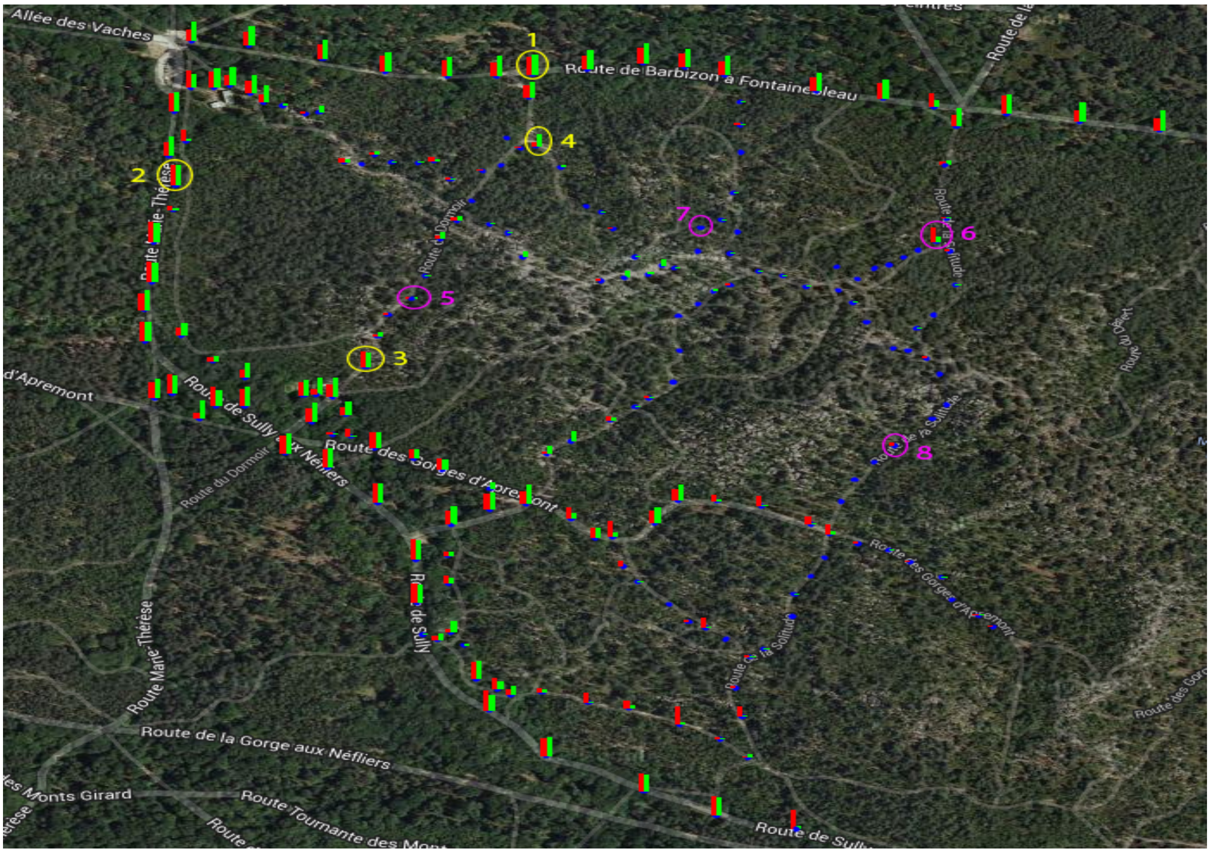
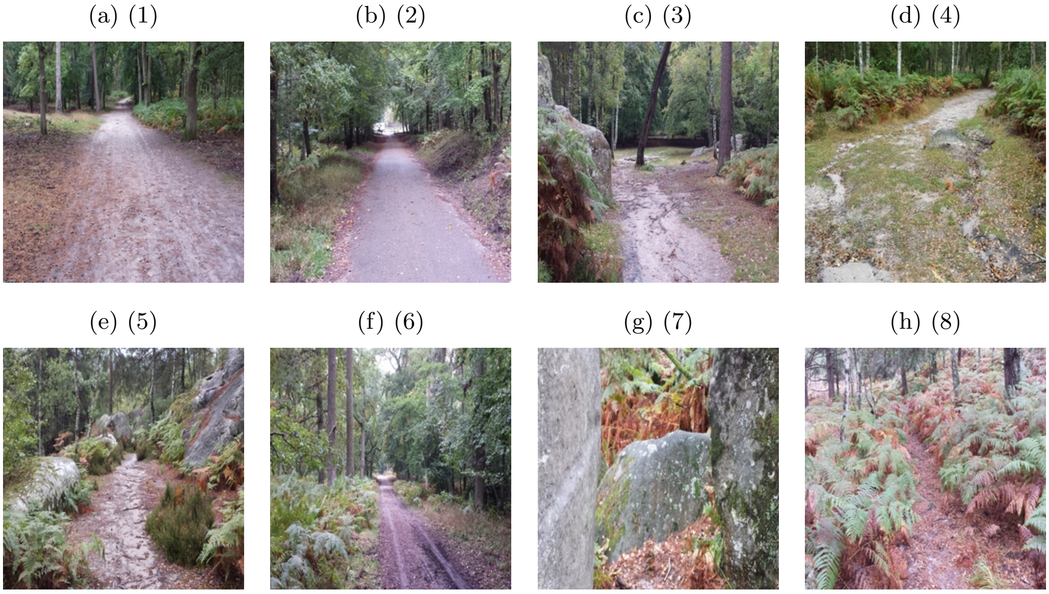

Allée des Vaches
Route de Barbizon à Fontainebleau
1
2
3
4
5
6
7
8
d'Apremont
Route de Sully aux Néfliers
Route des Gorges d'Apremont
Route du Dormoir
Route de Sully
Route Marie-Thérèse
Route de la Gorge aux Néfliers
des Monts Girard
Route Tournante des Monts
Route de Sully
Route de la Solitude

Figure 3: Red bars stand for annotated probability. Green bars stand for predicted probability



(a) (1)

(b) (2)

(c) (3)

(d) (4)

(e) (5)

(f) (6)

(g) (7)

(h) (8)

Figure 4: Top row: examples of high-scored scenes. Bottom row: examples of low-scored scenes

2.3.2 Is it a good place to cycle with a mountain bike?

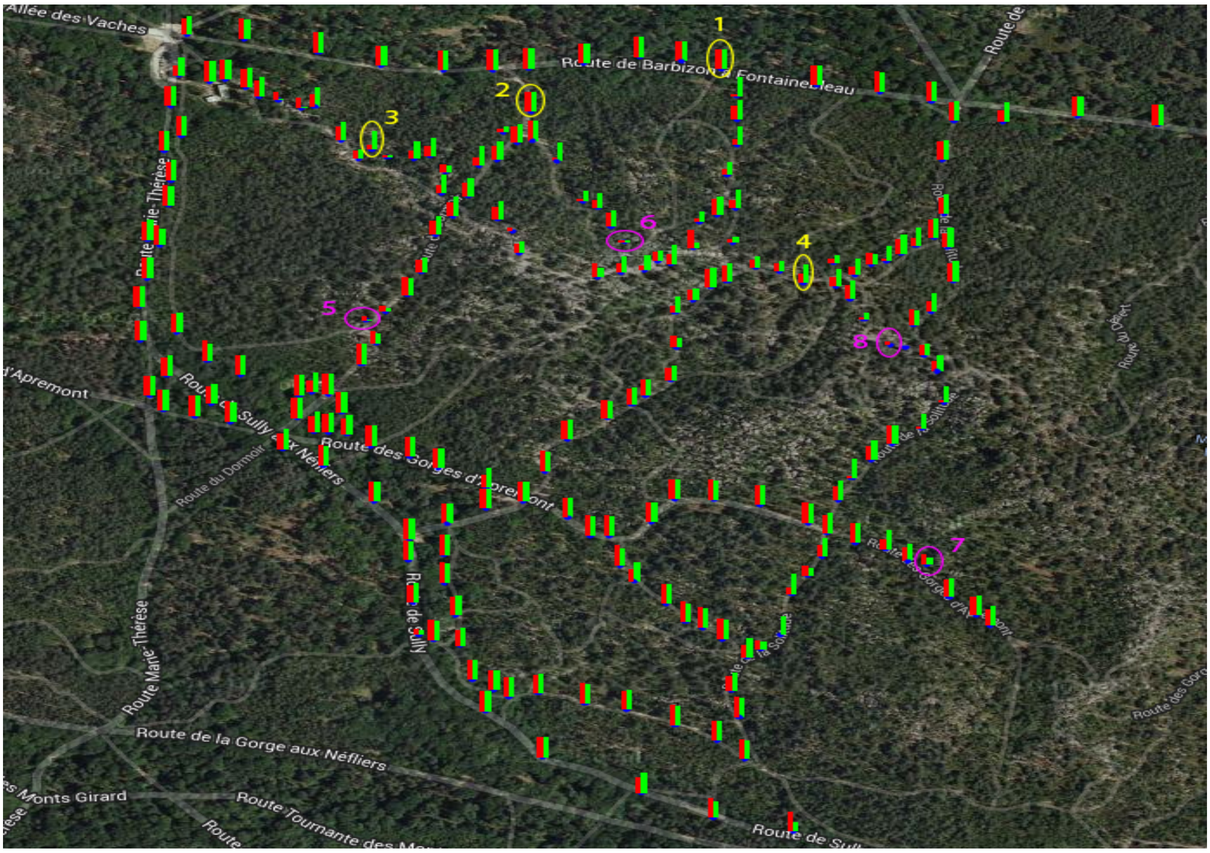
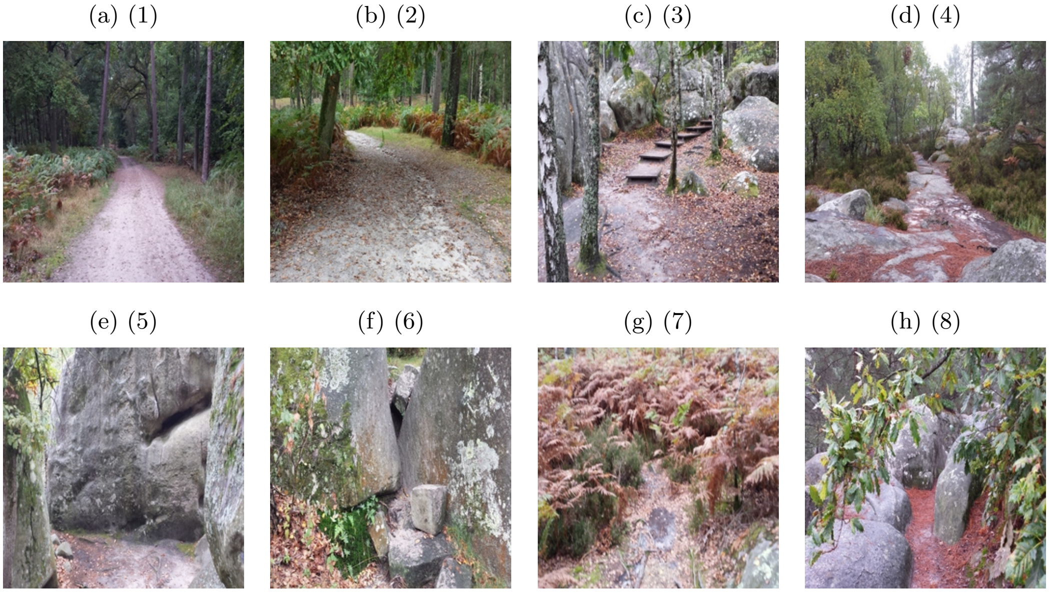


Figure 5: Red bars stand for annotated probability. Green bars stand for predicted probability


(a) (1)

(b) (2)

(c) (3)

(d) (4)

(e) (5)

(f) (6)

(g) (7)

(h) (8)

Figure 6: Top row: examples of high-scored scenes. Bottom row: examples of low-scored scenes

2.3.3 Is it a good place to drive a Jeep/4x4 car?

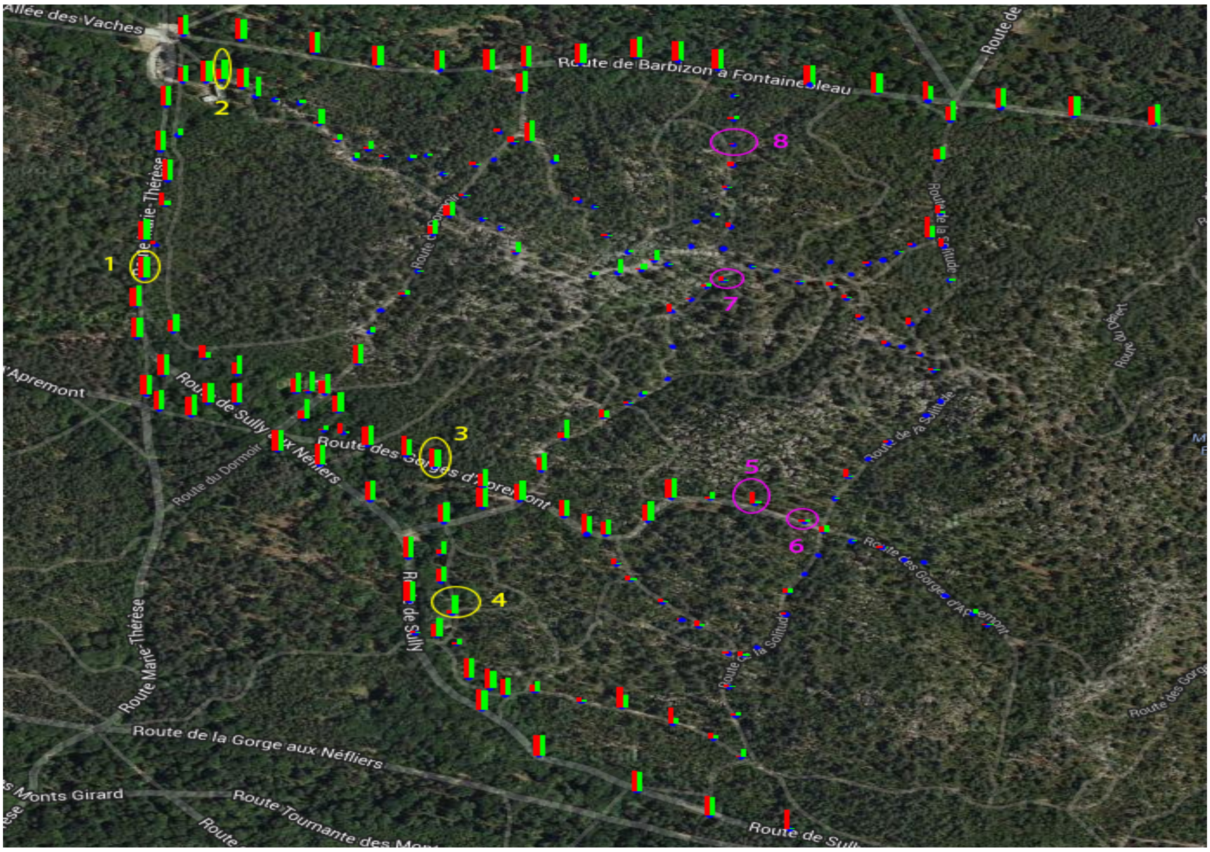


Figure 7: Red bars stand for annotated probability. Green bars stand for predicted probability

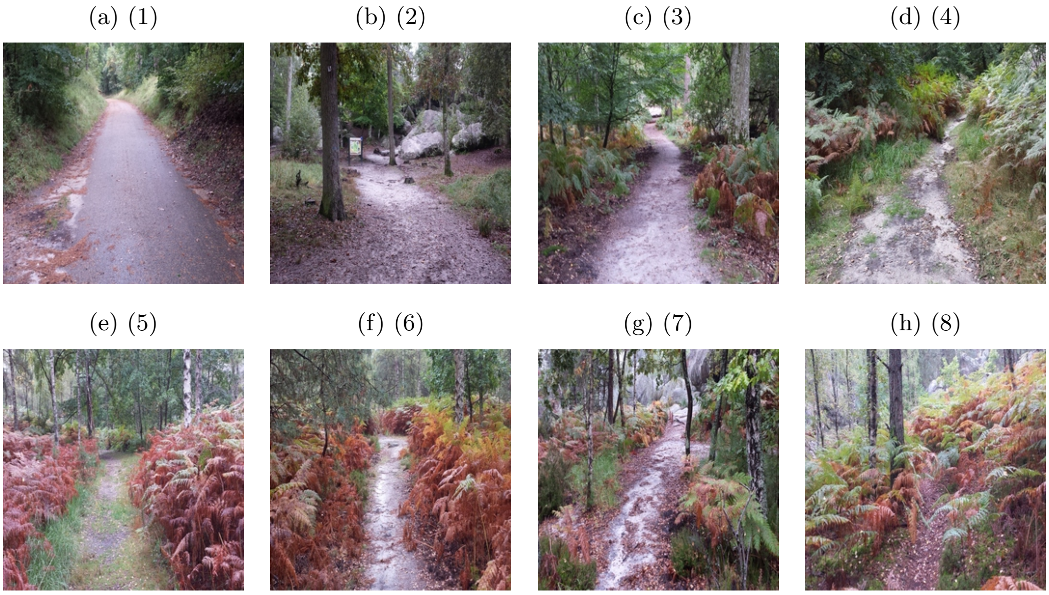

(a) (1)
(b) (2)
(c) (3)
(d) (4)
(e) (5)
(f) (6)
(g) (7)
(h) (8)

Figure 8: Top row: examples of high-scored scenes. Bottom row: examples of low-scored scenes

2.3.4 Is it a good place to drive a regular car?

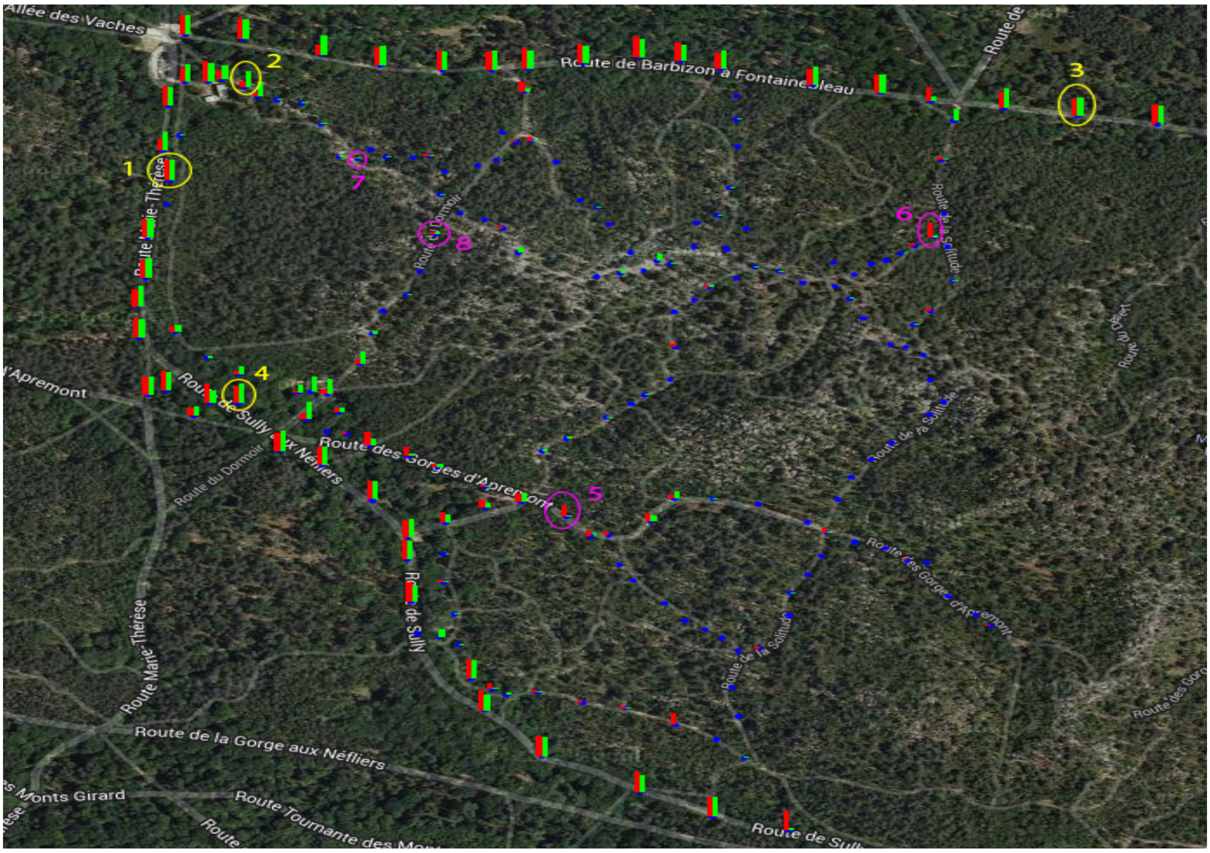
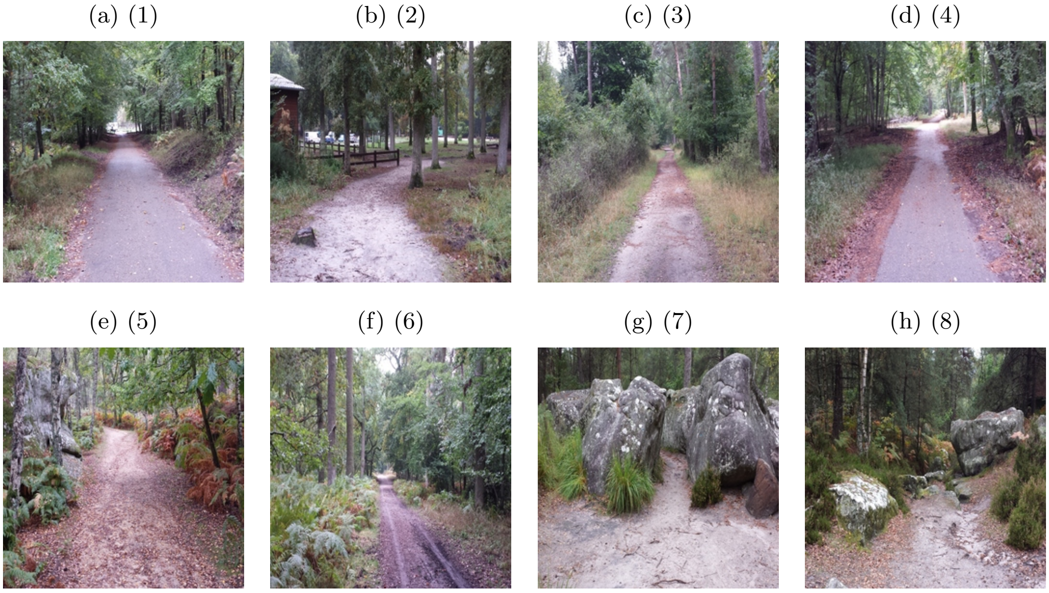


Figure 9: Red bars stand for annotated probability. Green bars stand for predicted probability


(a) (1)

(b) (2)

(c) (3)

(d) (4)

(e) (5)

(f) (6)

(g) (7)

(h) (8)

Figure 10: Top row: examples of high-scored scenes. Bottom row: examples of low-scored scenes

2.3.5 Is it a good place to have a picnic?

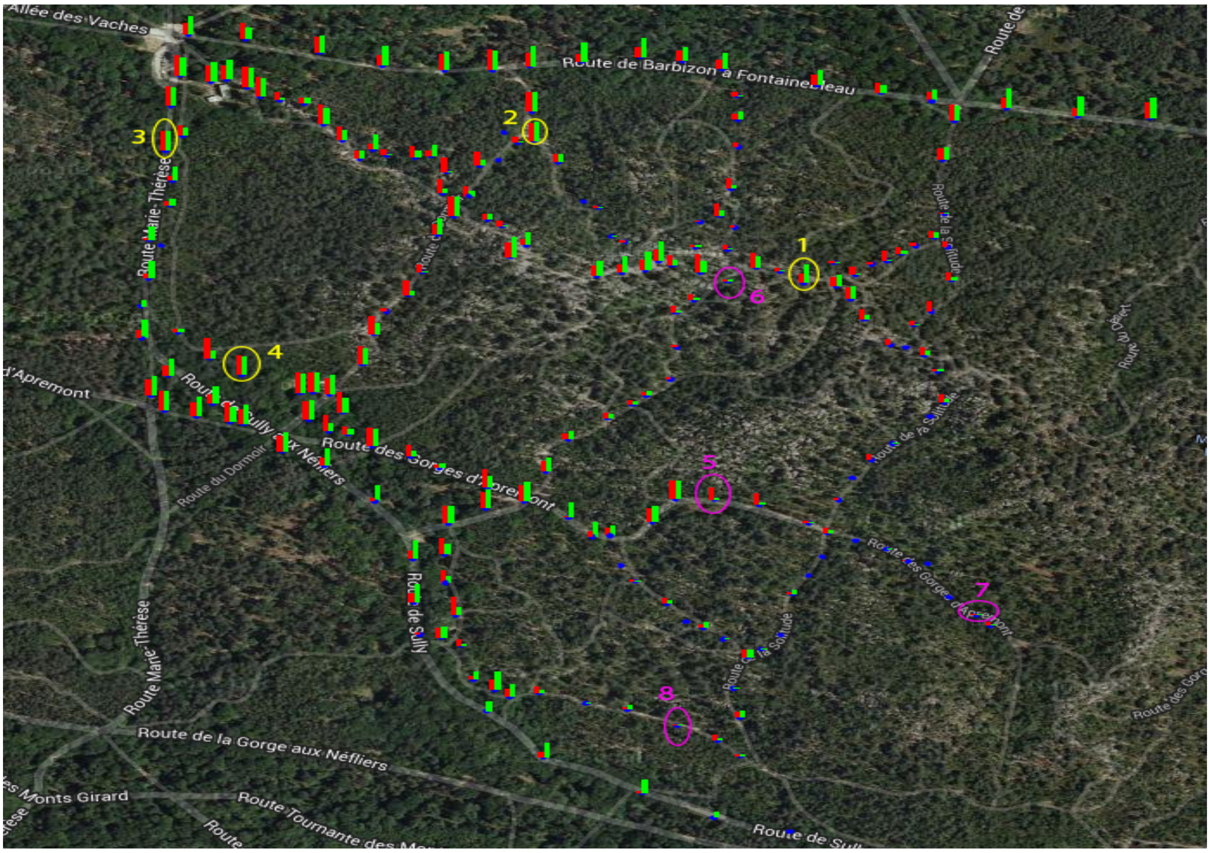
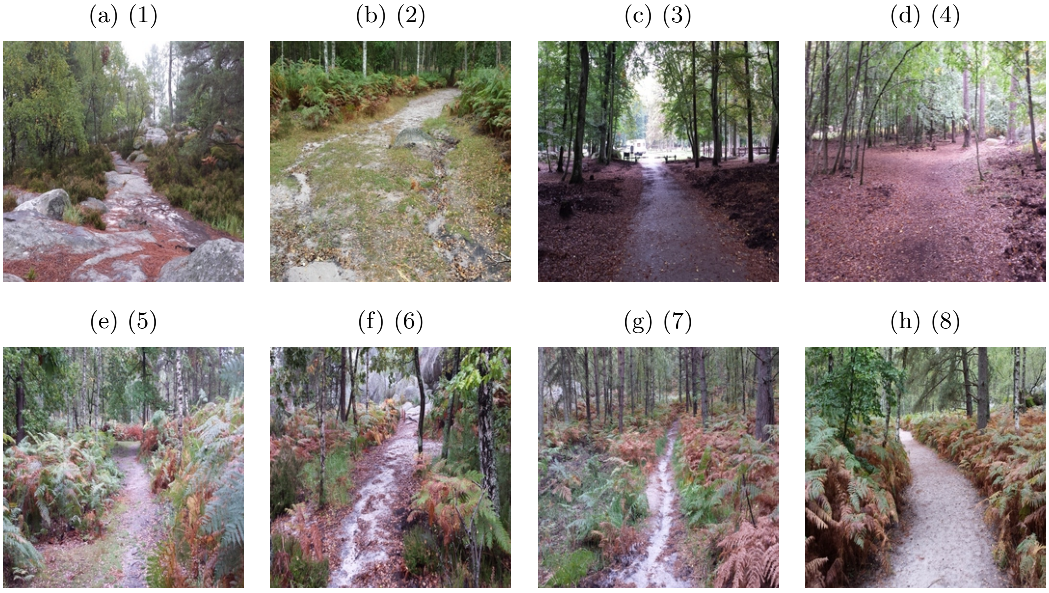


Figure 11: Red bars stand for annotated probability. Green bars stand for predicted probability



(a) (1)

(b) (2)

(c) (3)

(d) (4)

(e) (5)

(f) (6)

(g) (7)

(h) (8)

Figure 12: Top row: examples of high-scored scenes. Bottom row: examples of low-scored scenes

2.3.6 Is it a good place to jog?

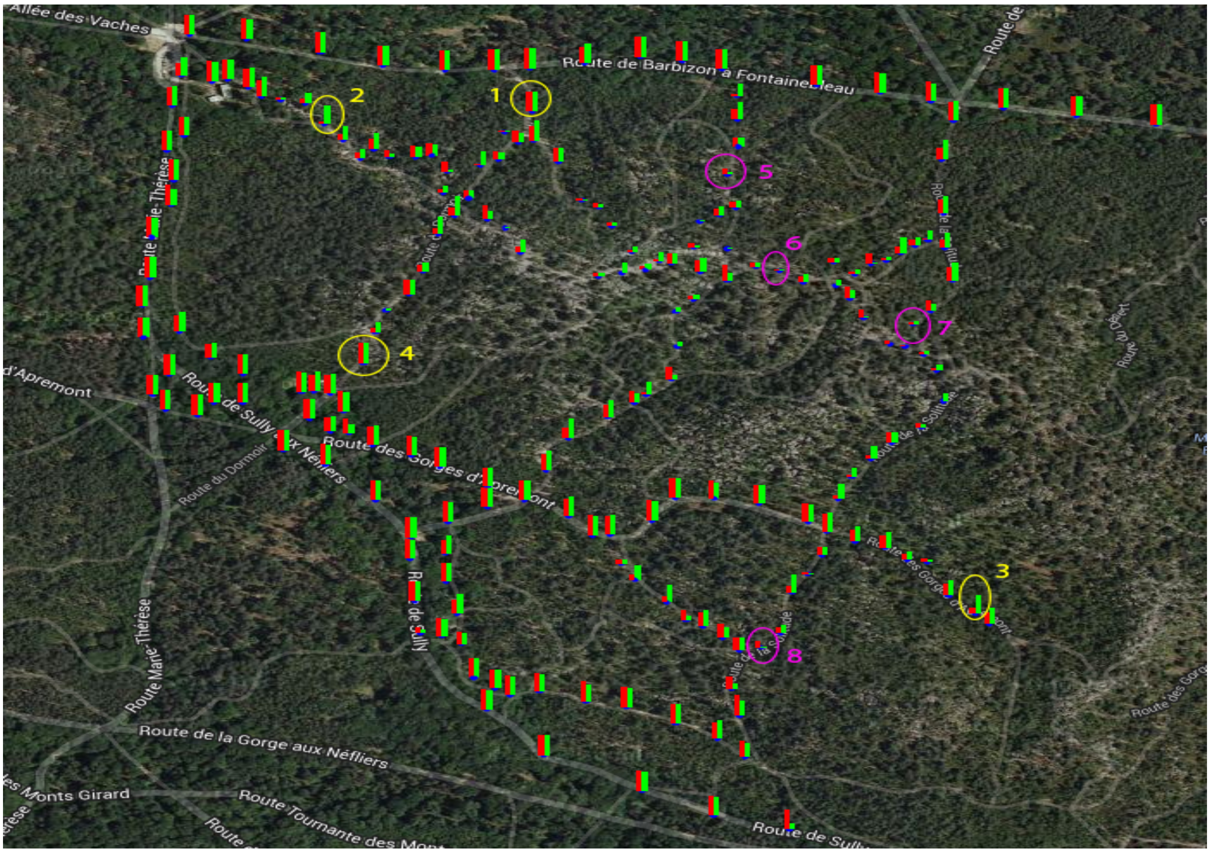
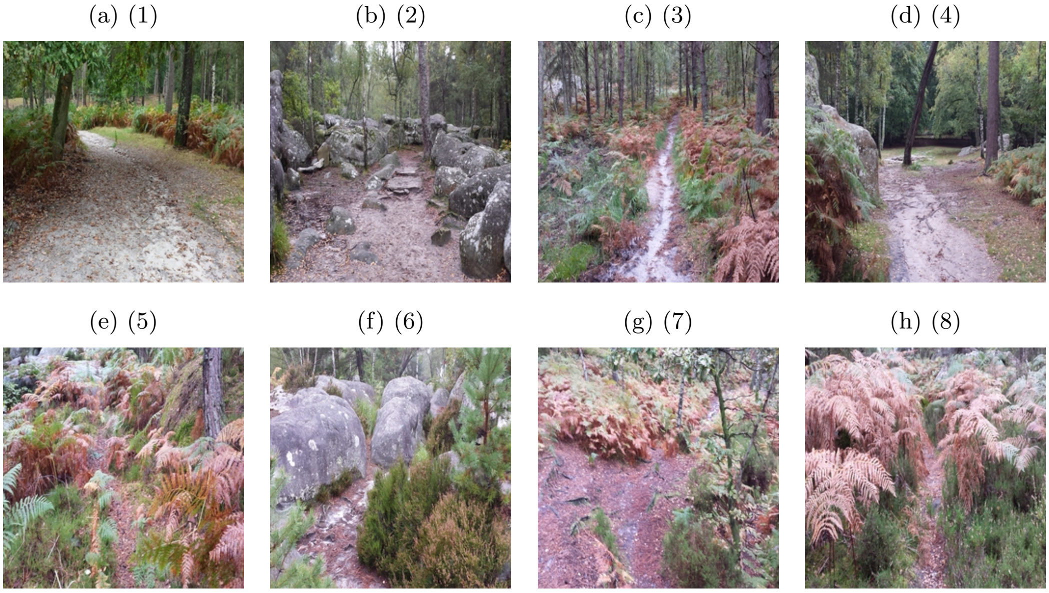


Figure 13: Red bars stand for annotated probability. Green bars stand for predicted probability



(a) (1)

(b) (2)

(c) (3)

(d) (4)

(e) (5)

(f) (6)

(g) (7)

(h) (8)

Figure 14: Top row: examples of high-scored scenes. Bottom row: examples of low-scored scenes

2.3.7 Is it a good place to put up a tent for camping?

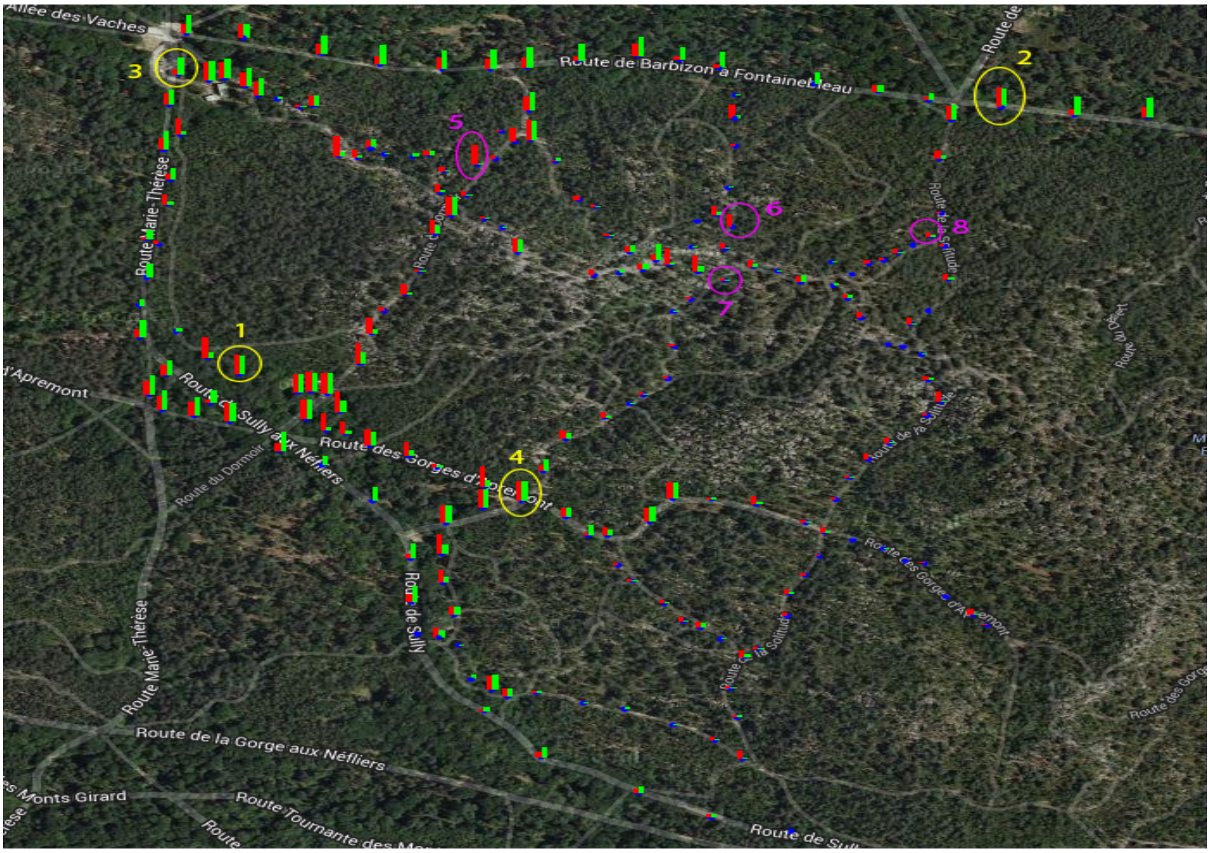
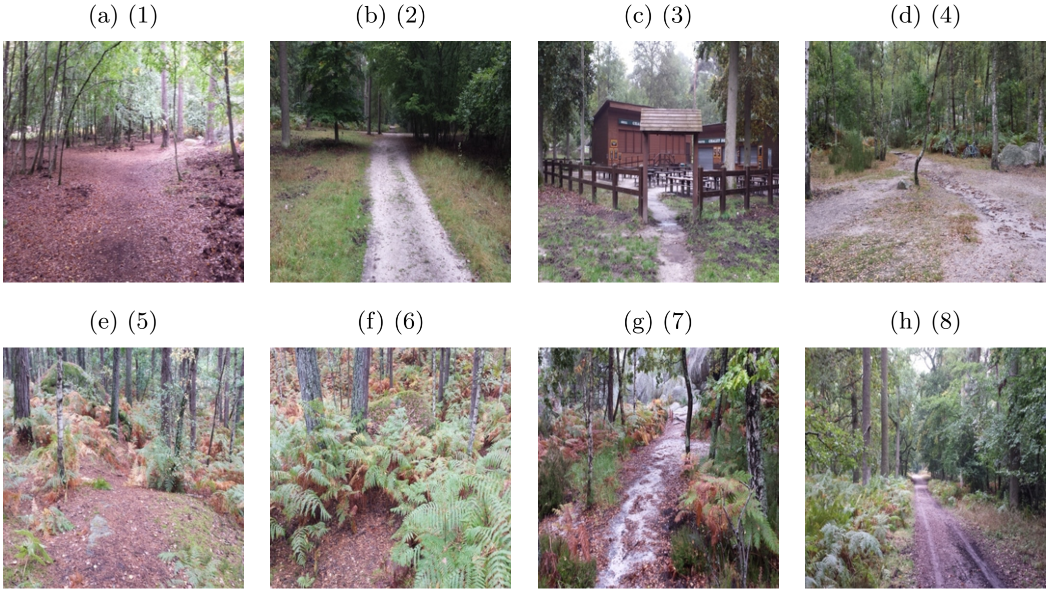


Figure 15: Red bars stand for annotated probability. Green bars stand for predicted probability


(a) (1) (b) (2) (c) (3) (d) (4)

(e) (5) (f) (6) (g) (7) (h) (8)

Figure 16: Top row: examples of high-scored scenes. Bottom row: examples of low-scored scenes

2.3.8 Is it a good place to ride a motorbike?

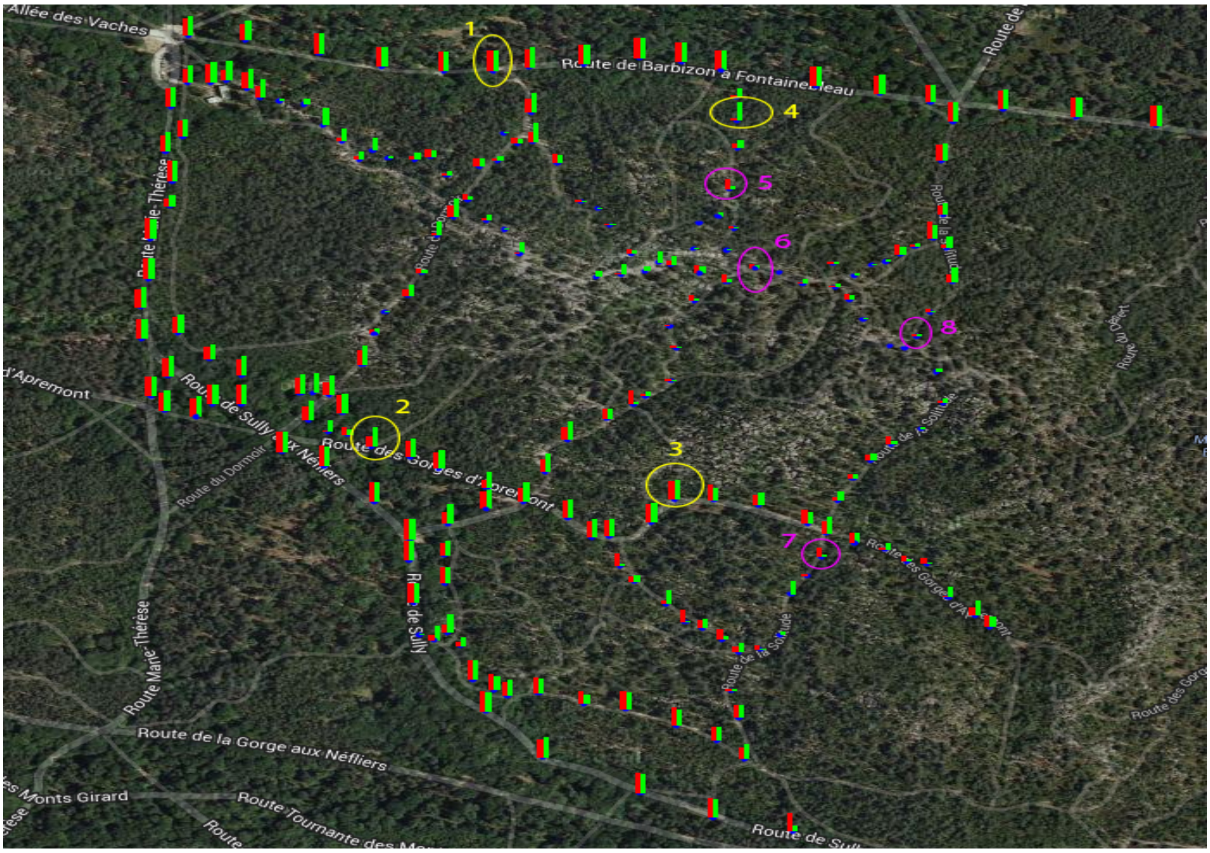
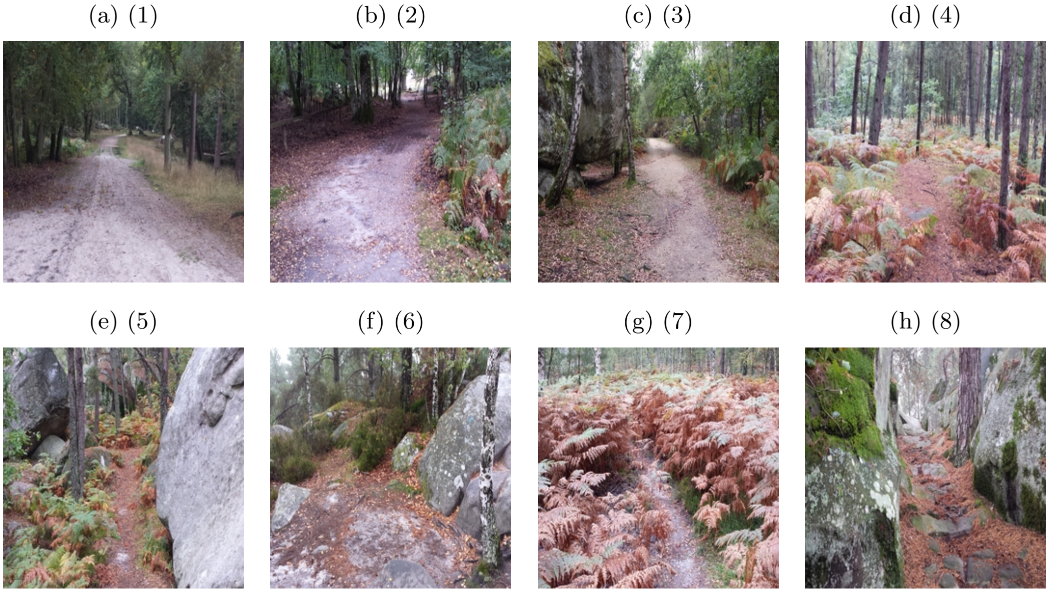


Figure 17: Red bars stand for annotated probability. Green bars stand for predicted probability


(a) (1)

(b) (2)

(c) (3)

(d) (4)

(e) (5)

(f) (6)

(g) (7)

(h) (8)

Figure 18: Top row: examples of high-scored scenes. Bottom row: examples of low-scored scenes

2.3.9 Is it a good place to walk with city shoes?

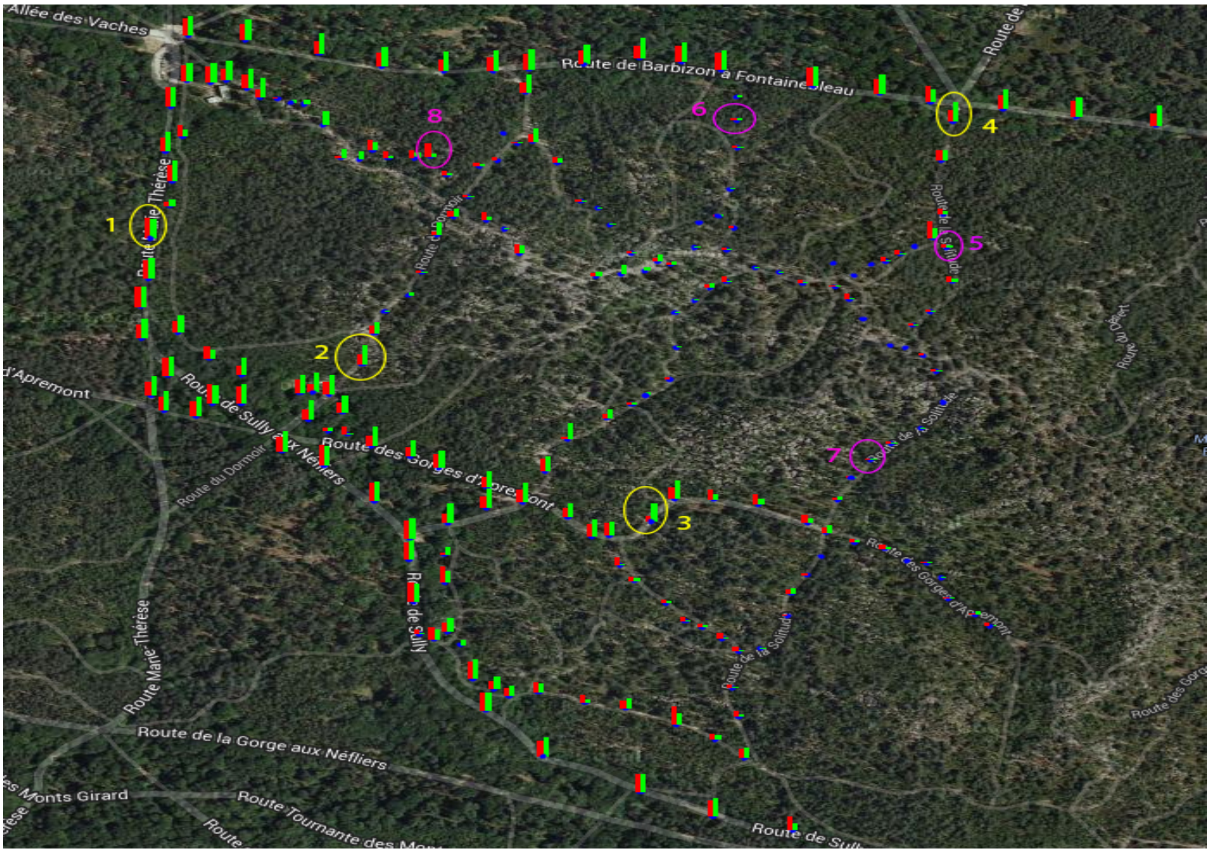


Figure 19: Red bars stand for annotated probability. Green bars stand for predicted probability

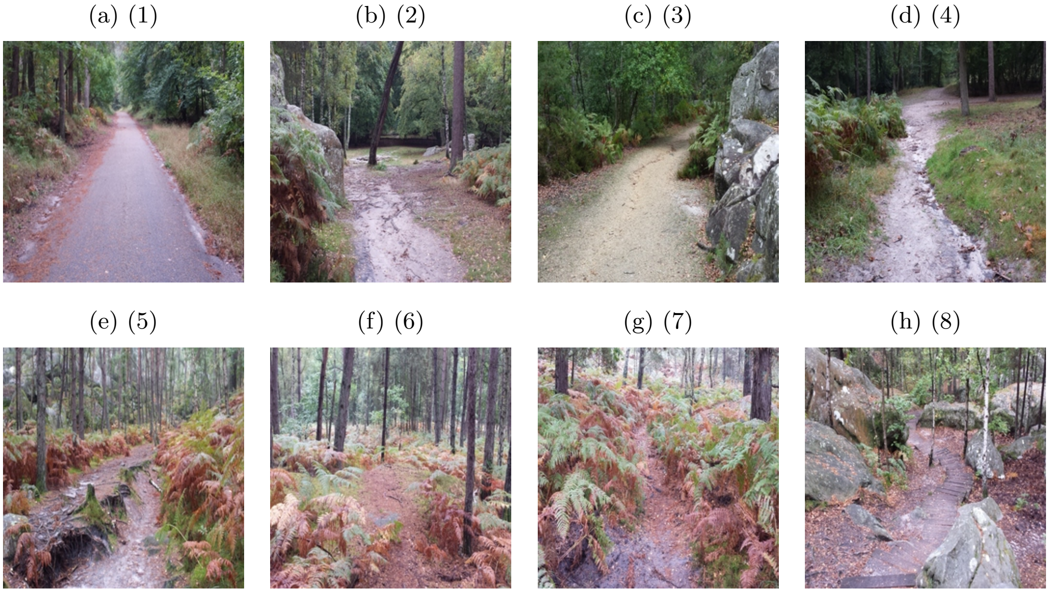

(a) (1)
(b) (2)
(c) (3)
(d) (4)
(e) (5)
(f) (6)
(g) (7)
(h) (8)

Figure 20: Top row: examples of high-scored scenes. Bottom row: examples of low-scored scenes