

000 054  
 001 055  
 002 056  
 003 057  
 004 058  
 005 059  
 006 060  
 007 061  
 008 062  
 009 063  
 010 064  
 011 065  
 012 066  
 013 067  
 014 068  
 015 069  
 016 070  
 017 071  
 018 072  
 019 073  
 020 074  
 021 075  
 022 076  
 023 077  
 024 078  
 025 079  
 026 080

## Viewpoint-aware Video Summarization (Supplemental Material)

Anonymous CVPR submission

Paper ID 2161

### 1. Relationship with other methods

The maximum bi-clique finding (MBF) technique [1] for video co-summarization builds a bi-partite graph for two videos, on which each segment corresponds to a node. Let  $\mathbf{u} \in \{0, 1\}^N$ ,  $\mathbf{v} \in \{0, 1\}^M$  be a vector indicating a selection of segments from video  $U$  and  $V$ , and  $C \in \mathbb{R}^{N \times M}$  be the similarity matrix between the segments of two videos used as an edge weight. This method finds a bi-clique from the graph with the maximum summation of weight. Formally, it maximizes  $\mathbf{u}^T C \mathbf{v}$  by using the constraint  $u_i + v_j \leq 1 + I(C_{ij} \geq \epsilon)$ , where the indicator is  $I(\cdot) = 1$  when the condition is met, otherwise it is 0, and  $\epsilon$  is the predefined threshold value.

The connection between this and our proposed methods can be observed. If we set  $\lambda_3 = 0$  in (10) in the main paper by ignoring the videos in other groups and assume that we treat only two samples (i.e.,  $n_k = 2$ ) denoting their selection vector as  $\mathbf{u} \in \{0, 1\}^N$ ,  $\mathbf{v} \in \{0, 1\}^M$ , the optimization problem in (10) in the main paper can be rewritten as

$$\max \begin{bmatrix} \mathbf{u}^T & \mathbf{v}^T \end{bmatrix} \begin{bmatrix} -(\lambda_1 - \frac{1}{2}\lambda_2)K_{UU} & \frac{1}{4}\lambda_2 K_{UV} \\ \frac{1}{4}\lambda_2 K_{UV}^\top & -(\lambda_1 - \frac{1}{2}\lambda_2)K_{VV} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}$$

where  $K_{UU}$ ,  $K_{UV}$ ,  $K_{VV}$  indicate the kernel matrices of shots features in the video  $U$  and  $U$ ,  $U$  and  $V$ , and  $V$  and  $V$  respectively. (In this paper, we utilized linear kernel instead of rbf kernel used in [1].) For simplicity, we assume features are normalized to meet  $k(\mathbf{x}, \mathbf{x}) = 1$  for all shot features  $\mathbf{x}$ . If we set  $\lambda_2 = 2\lambda_1$ , the block diagonal matrix will become 0, and the problem is simplified to the selection of a set of nodes from a bi-partite graph with the maximum inner weight, corresponding to  $\epsilon = 0$  in the MBF technique. From this, our algorithm can be regarded as a kind of generalization of MBF algorithm.

Furthermore, by only considering the first term (i.e.,  $(\lambda_2 = 0, \lambda_3 = 0)$ ), we can find an analogy to methods that aim to preserve diversity. For example, the DPP [3] extracts a subset whose determinant of the kernel matrix is the maximum, and Lu et al. [4] aims to minimize the similarity of consecutive frames in the summary. Our approach is different in that it minimizes the summation of all similarities in the summary, but it shares the same motivation as them.

### 2. Further results of user study

We fixed the *viewpoint* and we compared the generated summaries with the ones which is created based on one explicit concept in the main paper due to the difficulty of quantitative evaluation. We also conducted user study that measures the ability of estimating underlying *viewpoint* with weaker constraint using the same dataset. Experiments were conducted on the AMT-like web page we developed as shown in Fig. 1a and Fig. 1b.

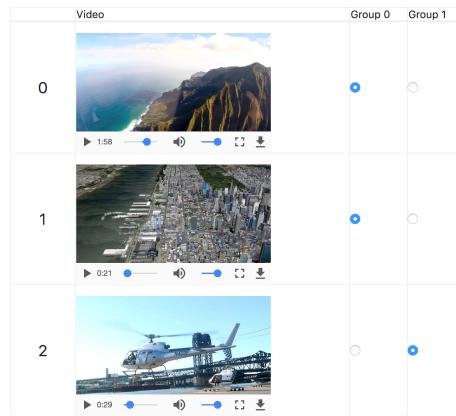
Firstly, four videos were randomly picked from each of **TG**, **OG1**, **OG2**, and they were shown to the subjects. Subjects were asked to split them to two groups based on one criteria which they decided on their own. Subsequently, they watched summaries of those videos belonging to **TG** generated by MBF [1], CVS [6], and ours (without feature learning). The summary which most reflects the criteria that was used to divide videos to groups was selected. (It was allowed to select multiple summaries. Moreover, if there were not appropriate one, subjects do not need to select anything.) For each task, five workers were assigned in this experiment.

Table 1: The ratio that the summary generated from each method were selected. N/A means no method were selected.

	N/A	MBF [1]	CVS [6]	ours
score	0.09	0.37	0.38	<b>0.50</b>

108 We show the number that each method were selected divided by the number of videos in the Table 1, and the score of our 162  
 109 method is better than the others in it. This result indicates that our method can generate the summary that explains the criteria 163  
 110 of grouping when the *viewpoint* changes person to person. 164  
 111  
 112  
 113  
 114  
 115  
 116

- Several videos are shown below.
- Please watch all of them and divide them into groups based on one aspect (e.g., location, activity,...) by checking either group 0 or group 1 of corresponding row.
- Please remember why videos are grouped the way they are because it will be used in the next step.



132 (a) The screenshot image of the web page used for dividing videos 186  
 133 to groups. 187

- Below are summaries of parts of videos you watched.
- Each row corresponds to one video, and different summaries from the same video are shown in Summary 0 - Summary 2 columns.
- Also, Group number you assigned in the previous page are shown in the last column.
- Please choose one which most reflects the aspect used for grouping in the previous page, and check button corresponding to that summary for each row. (You can choose multiple summaries.)
- If the evaluation is difficult, please check N/A columns.
- Most summaries has 5-10 seconds.

Summary 0	Summary 1	Summary 2	N/A Group
			<input type="checkbox"/> Group 1
			<input type="checkbox"/> Group 1
			<input type="checkbox"/> Group 1
			<input type="checkbox"/> Group 0
<a href="#">Submit</a>			

132 (b) The screenshot image of the web page used for the evaluation of 188  
 133 summaries. 189

134 Figure 1: The screenshot of web pages developed for the user study evaluation. 190

### 140 3. Further results of quantitative experiments 194

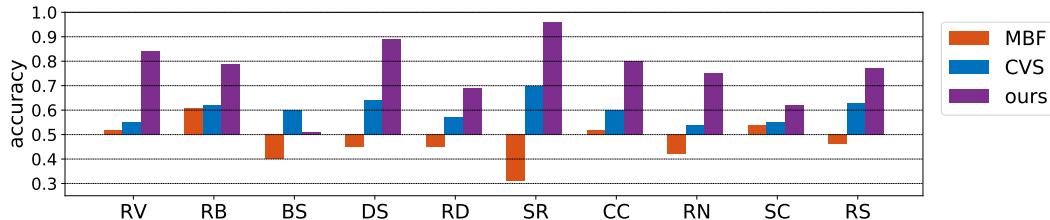
141 We also report the top-10 mean AP in Table 2. For the experimental settings, please refer the subsection 5.5 in the main 195  
 142 paper. 196

143 Table 2: top-10 Mean AP computed from human-created summary and predicted summary for each method. Results are 197  
 144 shown for each **target group**. For referring to the abbreviated names of groups, please see the Table 1 in the main paper. 198

	RV	RB	BS	DS	RD	SR	CC	RN	SC	RS	mean
SMRS [2]	0.354	0.370	0.373	0.335	0.320	0.344	0.309	0.374	0.365	0.344	0.349
CK	0.386	0.334	<u>0.393</u>	0.295	0.337	0.280	0.335	<b>0.434</b>	<u>0.400</u>	0.289	0.349
CS	0.344	0.344	0.326	0.333	0.319	0.304	0.330	0.384	<b>0.450</b>	0.310	0.344
MBF [1]	0.402	0.362	0.352	0.372	0.355	0.314	0.352	0.416	0.354	0.331	0.361
CVS [6]	0.370	0.382	<b>0.404</b>	0.381	<u>0.358</u>	0.387	0.374	0.376	0.408	<u>0.380</u>	0.382
WSVS [5]	0.358	0.303	0.356	0.353	<u>0.318</u>	0.368	0.359	0.344	0.349	0.323	0.343
WSVS (large) [5]	0.372	0.333	0.365	0.350	0.322	0.319	0.343	0.343	0.384	0.319	0.345
ours	<b>0.404</b>	<u>0.393</u>	0.366	<u>0.423</u>	0.338	<u>0.540</u>	<u>0.412</u>	0.386	0.387	0.375	0.402
ours (feature learning)	0.395	<b>0.407</b>	0.335	<b>0.430</b>	<b>0.363</b>	<b>0.545</b>	<b>0.423</b>	0.375	<u>0.399</u>	<b>0.393</b>	<b>0.406</b>

216 **4. Detailed result of topic selection task** 270  
217

218 Per-group accuracy of the topic selection task in the subsection 5.7 are displayed in the Fig. 2. We can see the topic of the  
 219 summary generated by our algorithm is correctly answered with higher probability than other methods, which demonstrates  
 220 the ability to recover the criteria of grouping. The performance of MBF was near random rate (0.5), and worse than that in  
 221 several groups. We conjecture the reason attributes to the fact that MBF uses only two videos to find the visual co-occurrence.  
 222 If the feature representation of shots which is representative to topics are similar each other, it may fail to find the common  
 223 pattern within the group.



233 Figure 2: Per-group accuracy of topic selection task. Each bar corresponds to the each method, namely, MBF [1] (orange),  
 234 CVS [6] (blue), and ours (purple). Please note 0.5 (random rate) are set to the center of this graph. For referring to the  
 235 abbreviated names of groups, please see the Table 1 in the main paper.

239 **5. Detailed derivation of equations** 293  
240241 **5.1. Trace of inner-video variance** 294  
242

$$Tr(S_i^V) = Tr\left(\sum_{t=1}^{T_i} z_t (\mathbf{x}_t - \mathbf{v}_i)(\mathbf{x}_t - \mathbf{v}_i)^\top\right) \quad (1)$$

$$= \sum_{t=1}^{T_i} Tr(z_t (\mathbf{x}_t - \mathbf{v}_i)(\mathbf{x}_t - \mathbf{v}_i)^\top) \quad (2)$$

$$= \sum_{t=1}^{T_i} z_t (\mathbf{x}_t - \mathbf{v}_i)^\top (\mathbf{x}_t - \mathbf{v}_i) \quad (3)$$

$$= \sum_{t=1}^{T_i} z_t \mathbf{x}_t^\top \mathbf{x}_t - 2\mathbf{v}_i^\top \sum_{t=1}^{T_i} z_t \mathbf{x}_t + \sum_{t=1}^{T_i} z_t \mathbf{v}_i^\top \mathbf{v}_i \quad (4)$$

$$= \sum_{t=1}^{T_i} z_t \mathbf{x}_t^\top \mathbf{x}_t - \frac{2}{s} \mathbf{z}_i^\top \mathbf{X}_i \mathbf{X}_i^\top \mathbf{z}_i + \frac{1}{s} \mathbf{z}_i^\top \mathbf{X}_i \mathbf{X}_i^\top \mathbf{z}_i \quad (5)$$

$$= \sum_{t=1}^{T_i} z_t \mathbf{x}_t^\top \mathbf{x}_t - \frac{1}{s} \mathbf{z}_i^\top \mathbf{X}_i \mathbf{X}_i^\top \mathbf{z}_i \quad (6)$$

(2) and (3) are derived by an identity  $Tr(\sum_i A_i) = \sum_i Tr(A_i)$ , and  $Tr(\mathbf{a}\mathbf{a}^\top) = \mathbf{a}^\top \mathbf{a}$ . To derive (5), we utilize the definition  $\mathbf{v}_i = \frac{1}{s} \mathbf{X}_i^\top \mathbf{z}_i$  and constraint  $\|\mathbf{z}_i\|_0 = s$ .

CVPR  
2161

CVPR  
2161

## 5.2. Trace of within-class variance

$$Tr(S_{(k)}^W) = Tr\left(\sum_{i \in L_{(k)}} s(\mathbf{v}_i - \boldsymbol{\mu}_k)(\mathbf{v}_i - \boldsymbol{\mu}_k)^\top\right) \quad (7)$$

$$= \sum_{i \in L_{(k)}} s(\mathbf{v}_i - \boldsymbol{\mu}_k)^\top (\mathbf{v}_i - \boldsymbol{\mu}_k) \quad (8)$$

$$= s \sum_{i \in L(k)} \mathbf{v}_i^\top \mathbf{v}_i - 2s(\sum \mathbf{v}_i)^\top \boldsymbol{\mu}_k + n_k s \boldsymbol{\mu}_k^\top \boldsymbol{\mu}_k \quad (9)$$

$$= \frac{1}{s} \sum_{i \in L_{(k)}} \mathbf{z}_i^\top \mathbf{X}_i \mathbf{X}_i^\top \mathbf{z}_i - \frac{2}{n_k s} \hat{\mathbf{z}}_{(k)}^\top \hat{\mathbf{X}}_{(k)} \hat{\mathbf{X}}_{(k)}^\top \hat{\mathbf{z}}_{(k)} + \frac{1}{n_k s} \hat{\mathbf{z}}_{(k)}^\top \hat{\mathbf{X}}_{(k)} \hat{\mathbf{X}}_{(k)}^\top \hat{\mathbf{z}}_{(k)} \quad (10)$$

389

390

$$= \frac{1}{s} \sum_{i \in \mathcal{I}} \mathbf{z}_i^\top \mathbf{X}_i \mathbf{X}_i^\top \mathbf{z}_i - \frac{1}{n_k s} \hat{\mathbf{z}}_{(k)}^\top \hat{\mathbf{X}}_{(k)} \hat{\mathbf{X}}_{(k)}^\top \hat{\mathbf{z}}_{(k)} \quad (11)$$

### 5.3. Trace of between-class variance

$$Tr(S^B) = Tr\left(\sum_{k=1}^K n_k s(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})^\top\right)$$

$$= \sum_{k=1}^K n_k s(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})^\top (\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}}) \quad (12)$$

$$= s \sum_{k=1}^K n_k \boldsymbol{\mu}_k^\top \boldsymbol{\mu}_k - 2s \bar{\boldsymbol{\mu}}^\top \left( \sum_{k=1}^K n_k \boldsymbol{\mu}_k \right) + N s \bar{\boldsymbol{\mu}}^\top \bar{\boldsymbol{\mu}} \quad (13)$$

$$= \frac{1}{s} \sum_{k=1}^K \frac{1}{n_k} \hat{\mathbf{z}}_{(k)}^\top \hat{\mathbf{X}}_{(k)} \hat{\mathbf{X}}_{(k)}^\top \hat{\mathbf{z}}_{(k)} - \frac{2}{Ns} \hat{\mathbf{z}}^\top \hat{\mathbf{X}} \hat{\mathbf{X}}^\top \hat{\mathbf{z}} + \frac{1}{Ns} \hat{\mathbf{z}}^\top \hat{\mathbf{X}} \hat{\mathbf{X}}^\top \hat{\mathbf{z}} \quad (14)$$

$$= \hat{\mathbf{z}}^\top \left( \frac{1}{s} \oplus \sum_{k=1}^K \frac{1}{n_k} \hat{\mathbf{X}}_{(k)} \hat{\mathbf{X}}_{(k)}^\top \right) \hat{\mathbf{z}} - \frac{1}{Ns} \hat{\mathbf{z}}^\top \hat{\mathbf{X}} \hat{\mathbf{X}}^\top \hat{\mathbf{z}} \quad (15)$$

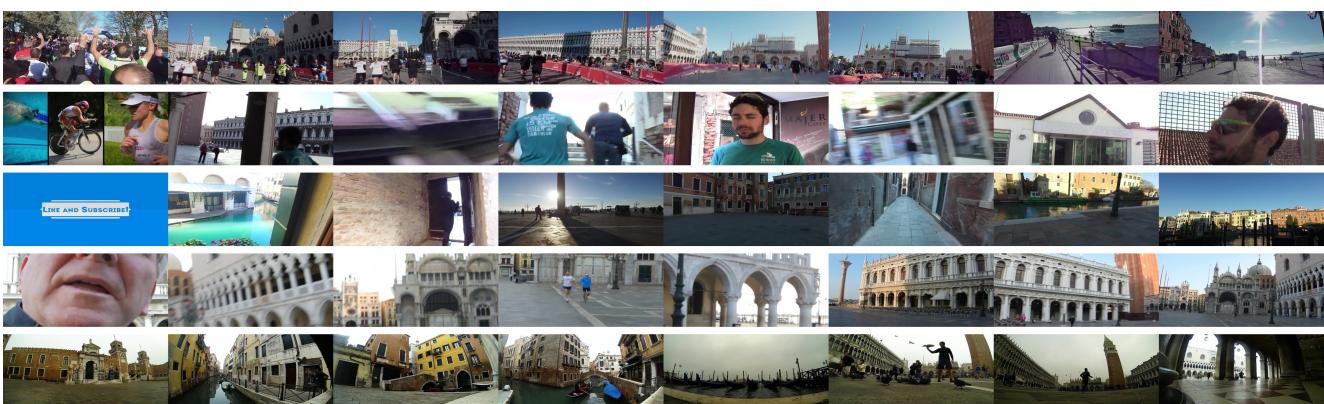
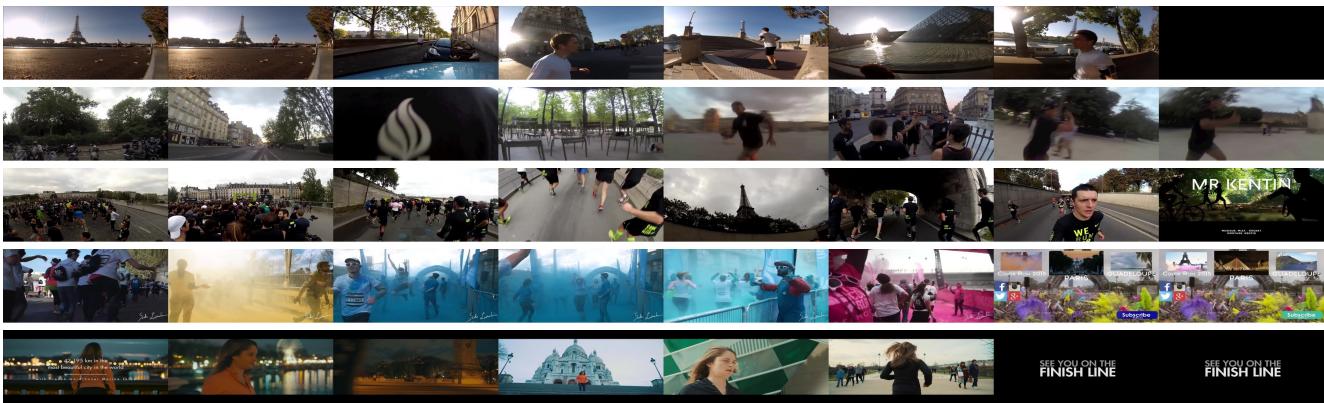
$$= \hat{\mathbf{z}}^\top (C - A)\hat{\mathbf{z}} \quad (16)$$

$\mathbf{v}_k - \frac{1}{n_k s} \mathbf{A}(k) \mathbf{z}(k)$  and  $\boldsymbol{\mu} = \frac{1}{Ns} \mathbf{A}^\top \mathbf{z}$  are used for (14). 415

(13) is derived  $\sum_{k=1}^K n_k = N$ .  $\boldsymbol{\mu}_k = \frac{1}{n_k s} \hat{\mathbf{X}}_{(k)}^\top \hat{\mathbf{z}}_{(k)}$  and  $\bar{\boldsymbol{\mu}} = \frac{1}{Ns} \hat{\mathbf{X}}^\top \hat{\mathbf{z}}$  are used for (14).

## **6. Examples of dataset**

We show randomly selected frames of videos of our dataset in the following figures. The order of figure corresponds to the ones written in the Table. 1 in the main paper, namely, in the order of **TG**, **OG1**, **OG2**, and from top-row to bottom-row. Each row of figures corresponds to one video.

(a) Randomly selected frames from videos belonging to class run venice (**TG**). Each row corresponds to one video.(b) Randomly selected frames from videos belonging to class run paris (**OG1**). Each row corresponds to one video.(c) Randomly selected frames from videos belonging to class shopping venice (**OG2**). Each row corresponds to one video.

A collage of 24 small video frames arranged in a grid. The frames depict various cycling scenes, likely from a GoPro camera. Some frames show cyclists on a paved path, while others show them on a sandy beach. One frame shows a child wearing a pink unicorn helmet. Another frame shows a person performing a handstand on a rocky shore. The frames are numbered on the left side from 540 to 554 and on the right side from 594 to 608.

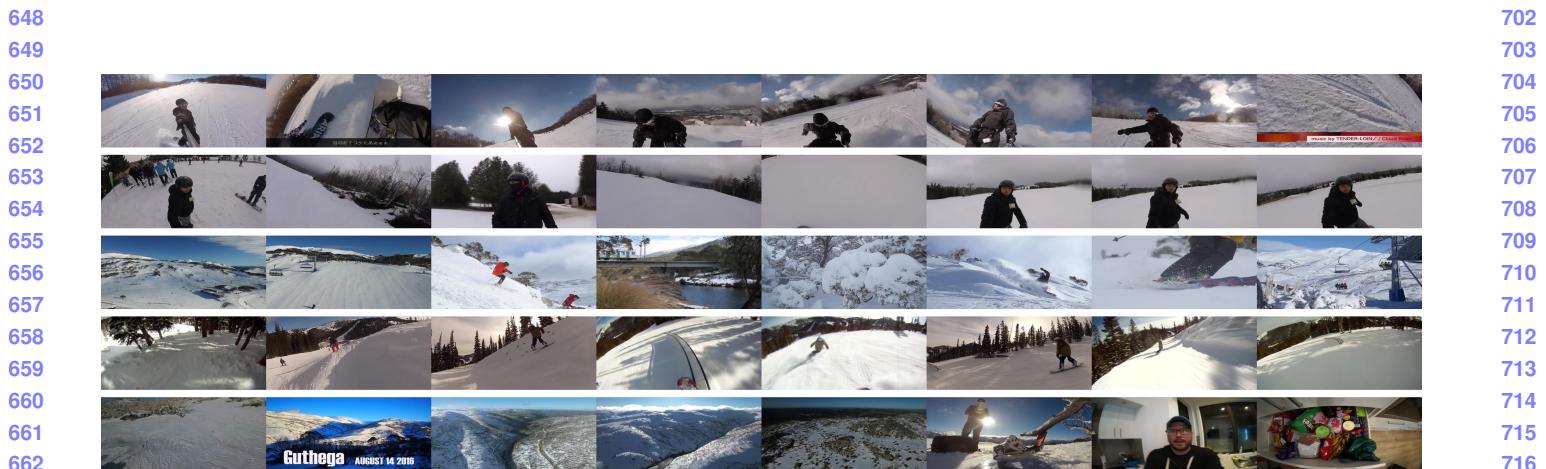
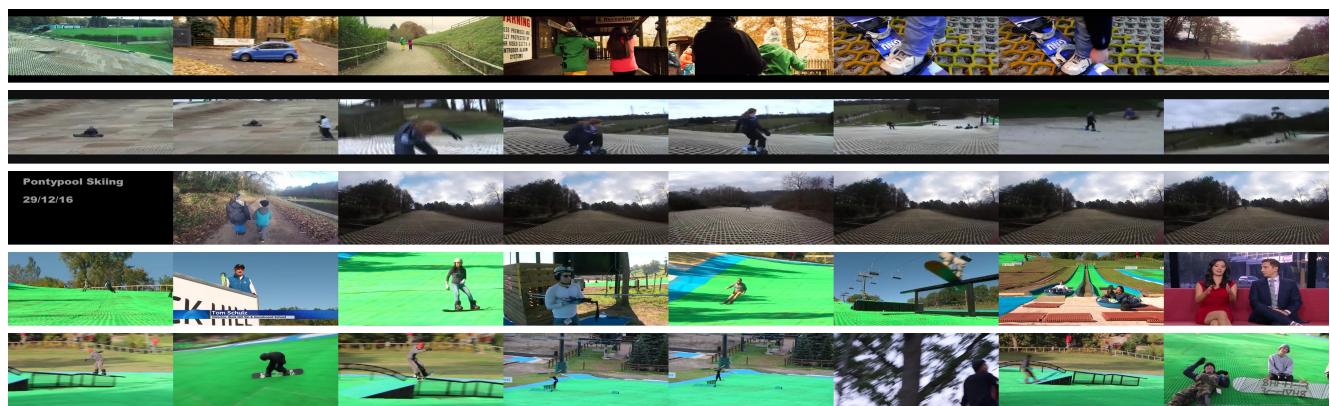
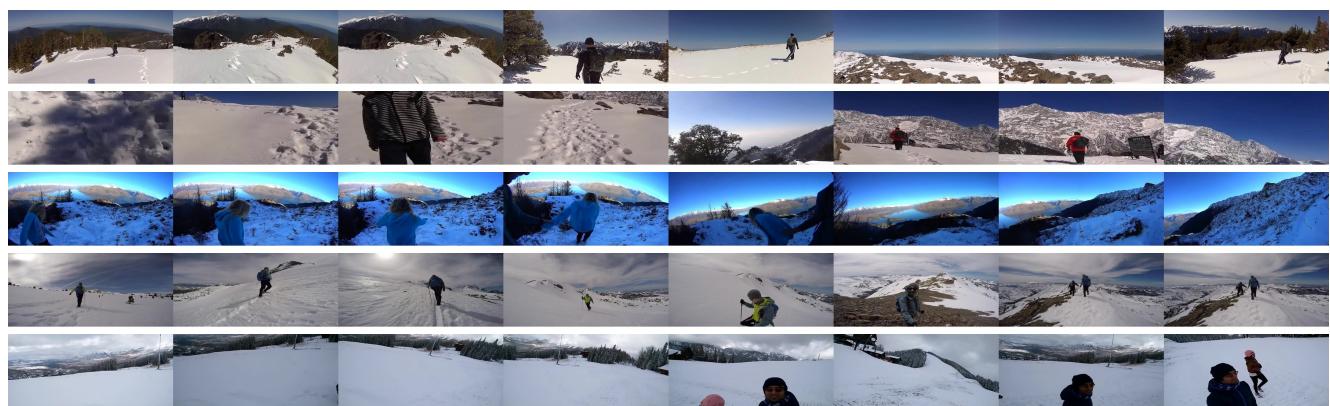
(a) Randomly selected frames from videos belonging to class ride bike beach (**TG**). Each row corresponds to one video.

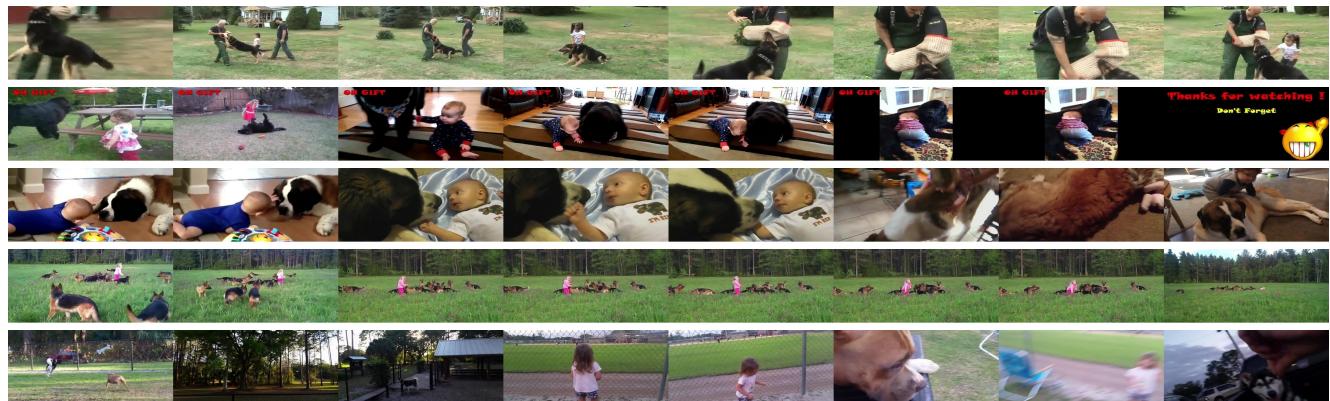
A grid of 15 images from a GoPro camera, arranged in three rows of five. The images show various cycling scenes: 1. A person walking on a city street. 2. A person walking on a sidewalk next to a yellow fence. 3. A person riding a bicycle on a red bike lane. 4. A person riding a bicycle on a city street. 5. A person riding a bicycle in a tunnel. 6. A person riding a bicycle on a city street. 7. A group of pigeons on the ground. 8. A person riding a bicycle from a first-person perspective. 9. A person riding a bicycle on a city street. 10. A person riding a bicycle on a city street. 11. A person riding a bicycle on a city street. 12. A person riding a bicycle on a city street. 13. A person riding a bicycle on a city street. 14. A person riding a bicycle on a city street. 15. A person riding a bicycle on a city street.

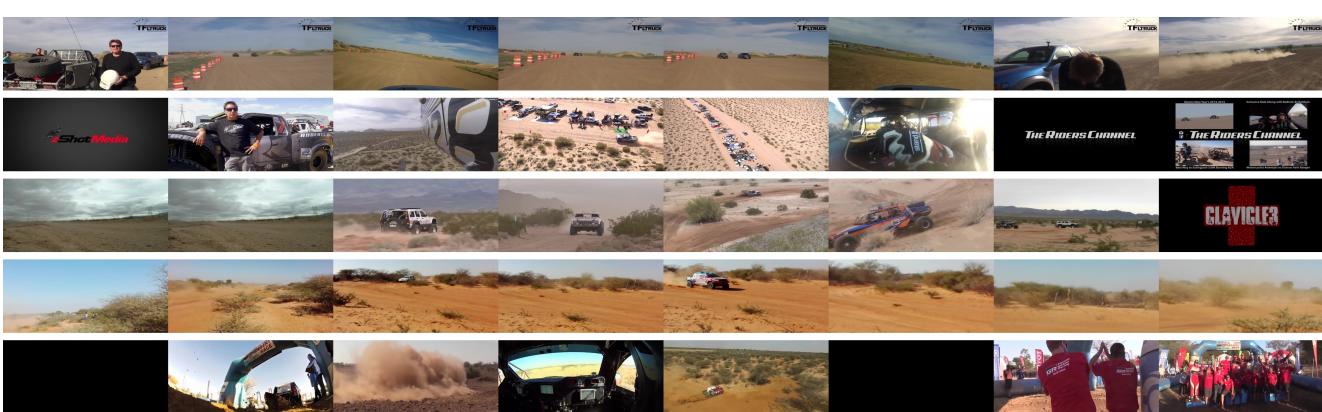
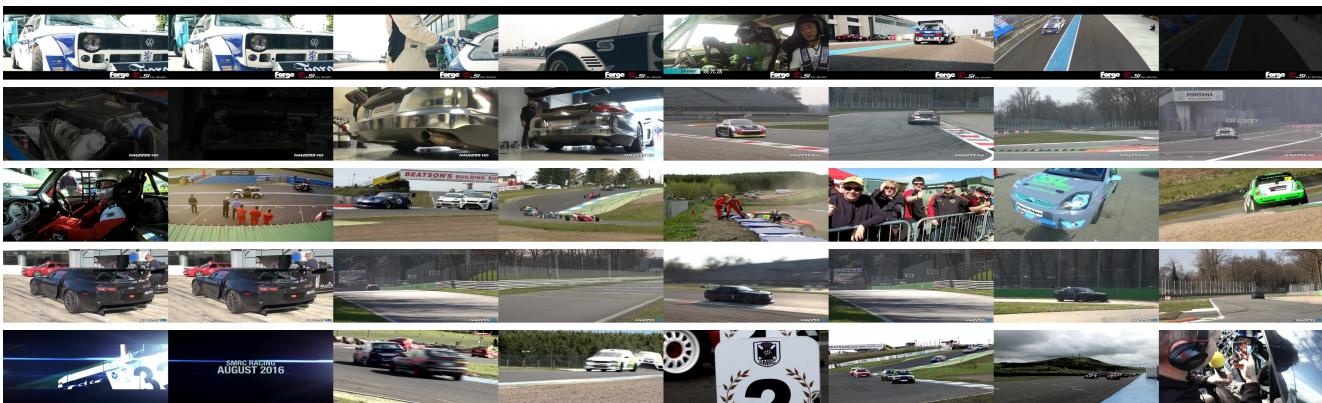
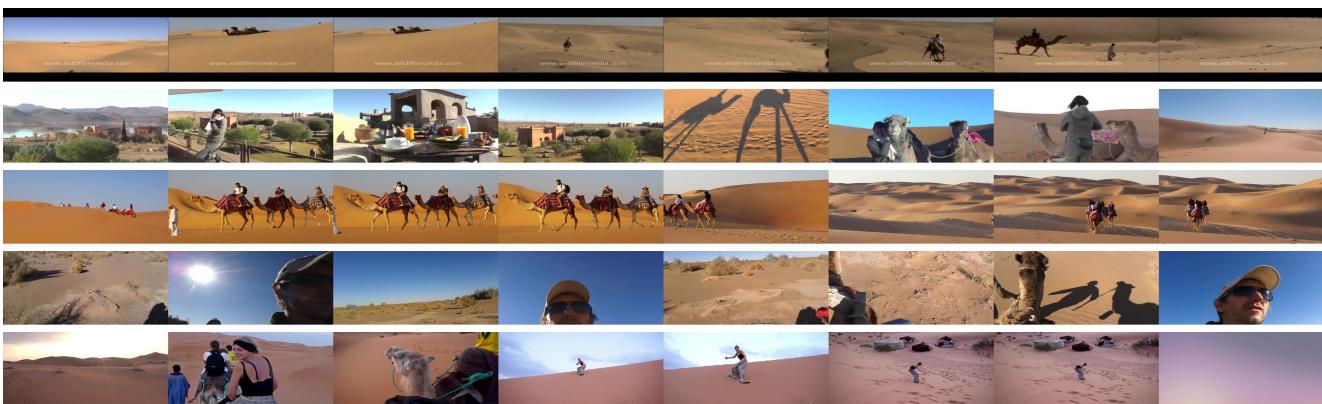
(b) Randomly selected frames from videos belonging to class ride bike city (**OG1**). Each row corresponds to one video.

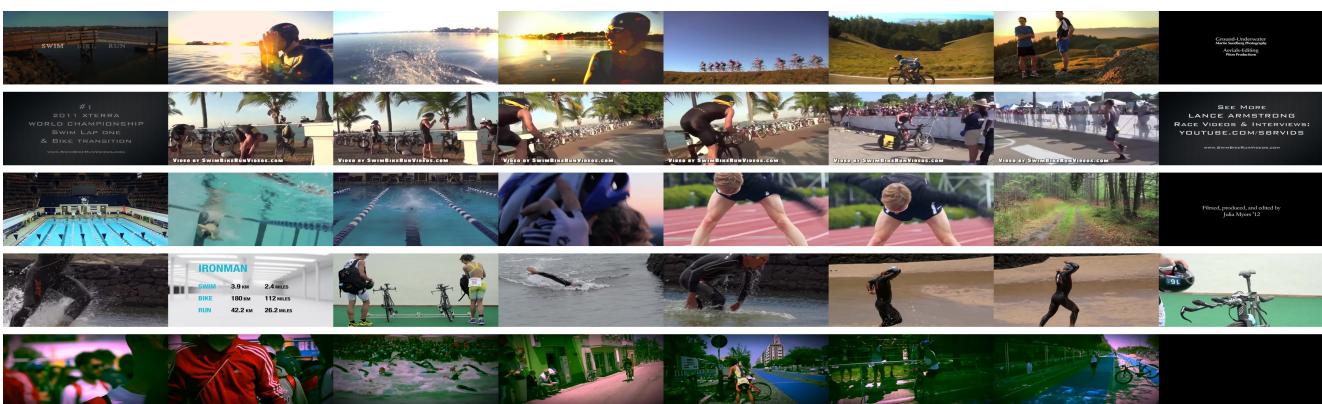
A grid of 40 images arranged in a 5x8 pattern. The images depict various scenes related to surfing and beach life, including surfers in action, people on the beach, and surfboards. Some images have captions or credits overlaid on them, such as 'Photo: Michael Shain - Life is a Beach Photography Studio' and '© 2013 Michael Shain'. The images are numbered on the left side from 574 to 587, and on the right side from 629 to 641.

(c) Randomly selected frames from videos belonging to class surf beach (**OG2**). Each row corresponds to one video.

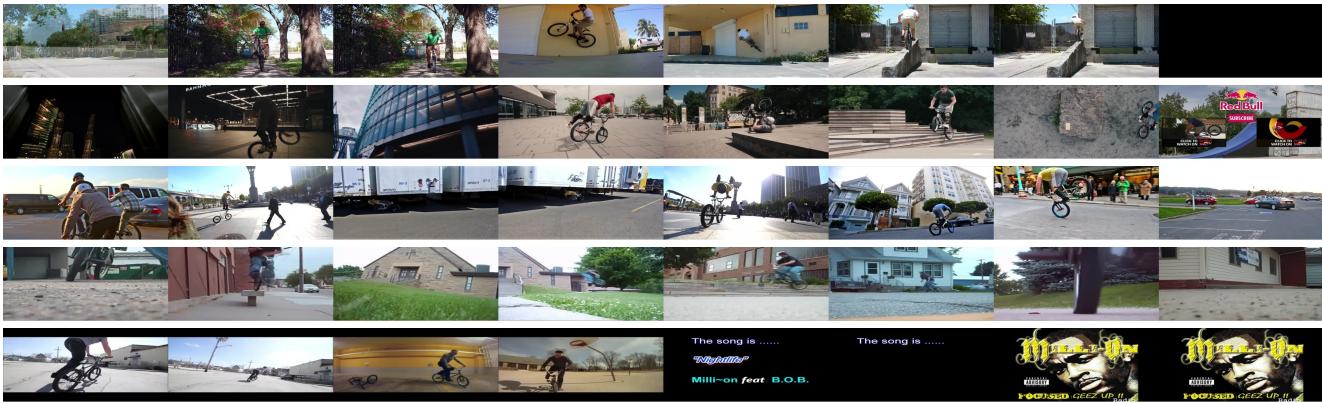
(a) Randomly selected frames from videos belonging to class boarding snow mountain (**TG**). Each row corresponds to one video.(b) Randomly selected frames from videos belonging to class boarding dry sloop (**OG1**). Each row corresponds to one video.(c) Randomly selected frames from videos belonging to class hiking snow mountain (**OG2**). Each row corresponds to one video.

(a) Randomly selected frames from videos belonging to class dog chase sheep (**TG**). Each row corresponds to one video.(b) Randomly selected frames from videos belonging to class dog play with kids (**OG1**). Each row corresponds to one video.(c) Randomly selected frames from videos belonging to class sheep graze grass (**OG2**). Each row corresponds to one video.

(a) Randomly selected frames from videos belonging to class racing desert (**TG**). Each row corresponds to one video.(b) Randomly selected frames from videos belonging to class racing circuit (**OG1**). Each row corresponds to one video.(c) Randomly selected frames from videos belonging to class riding camel desert (**OG2**). Each row corresponds to one video.



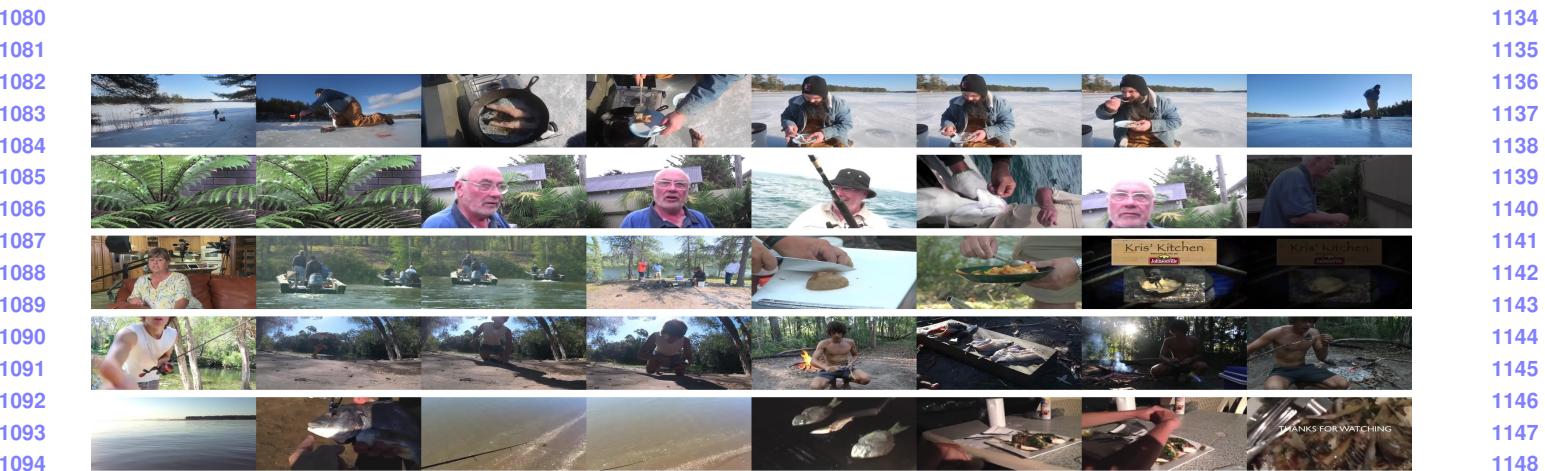
(a) Randomly selected frames from videos belonging to class swim riding bike (TG). Each row corresponds to one video.



(b) Randomly selected frames from videos belonging to class riding bike trick (OG1). Each row corresponds to one video.



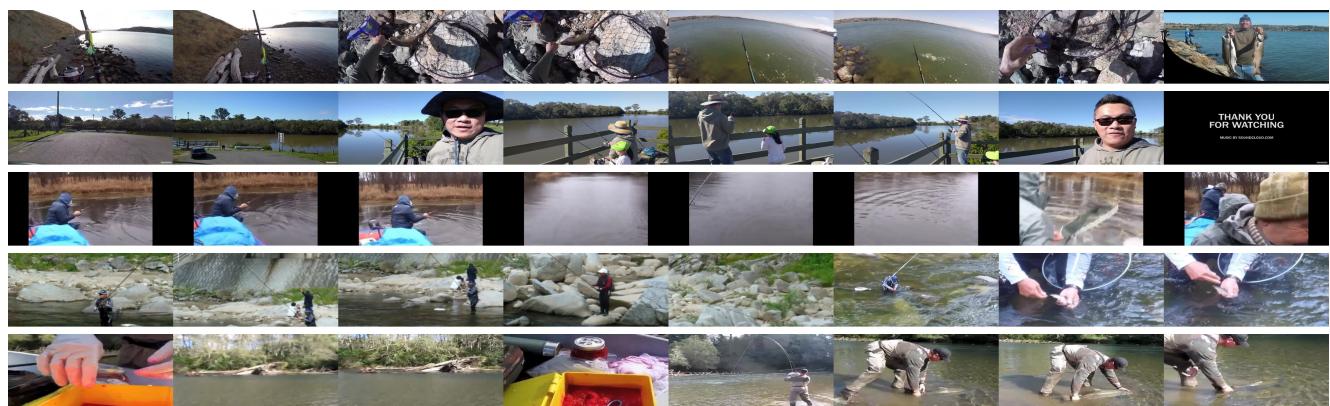
(c) Randomly selected frames from videos belonging to class swim dive (OG2). Each row corresponds to one video.



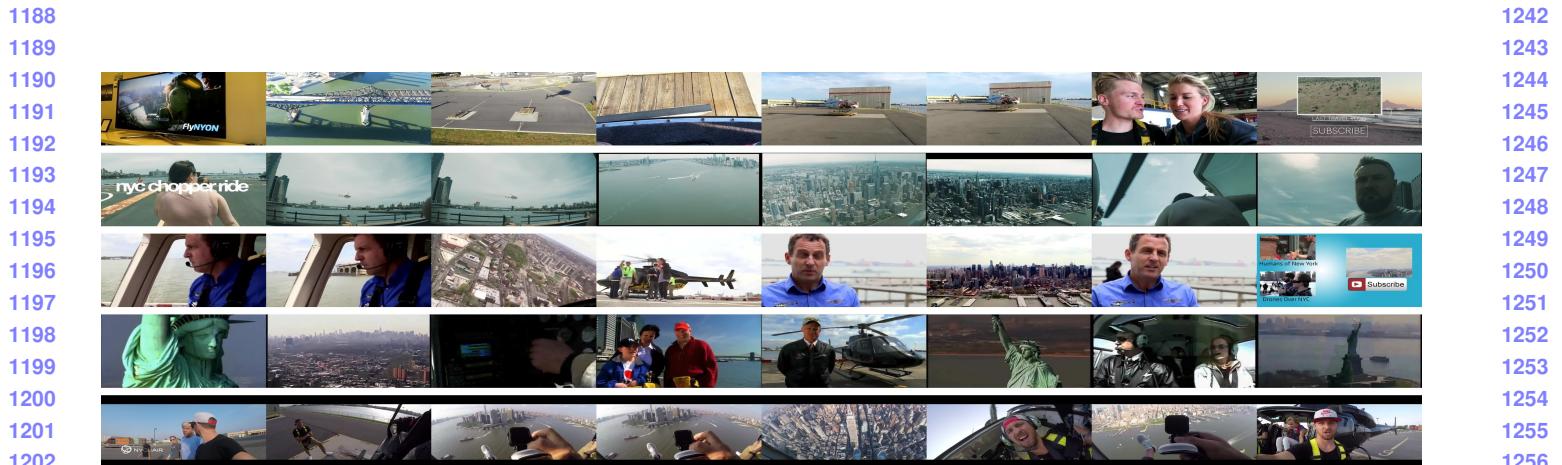
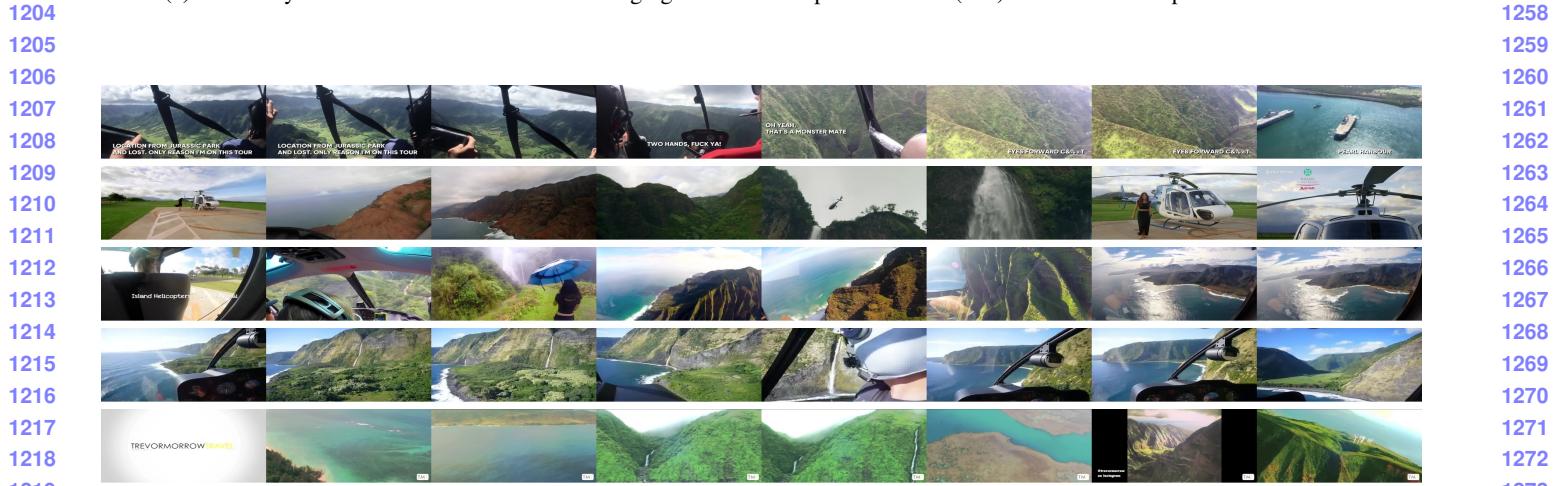
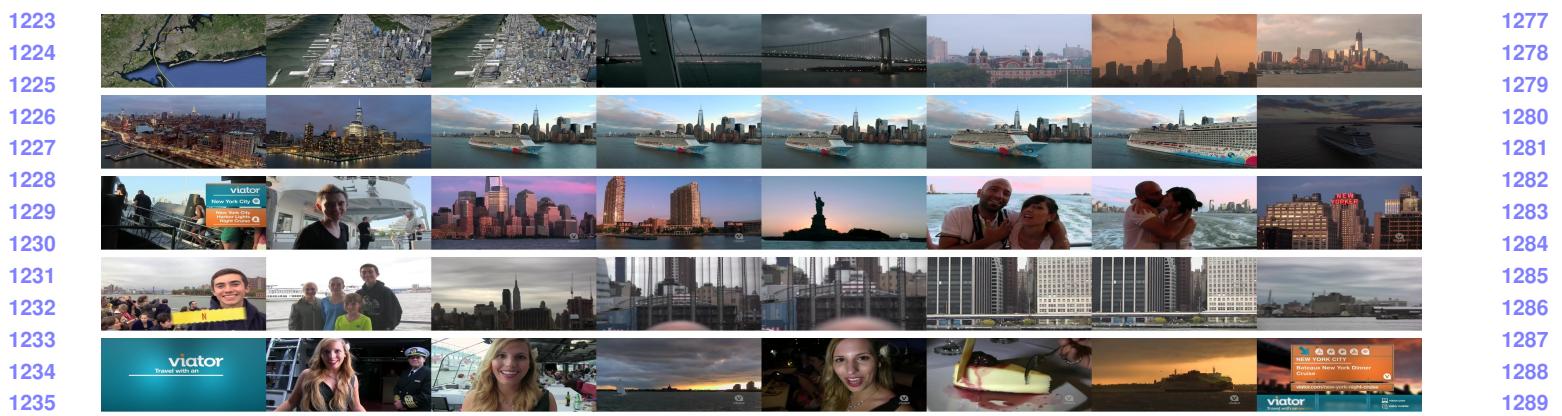
(a) Randomly selected frames from videos belonging to class fishing cook fish (TG). Each row corresponds to one video.

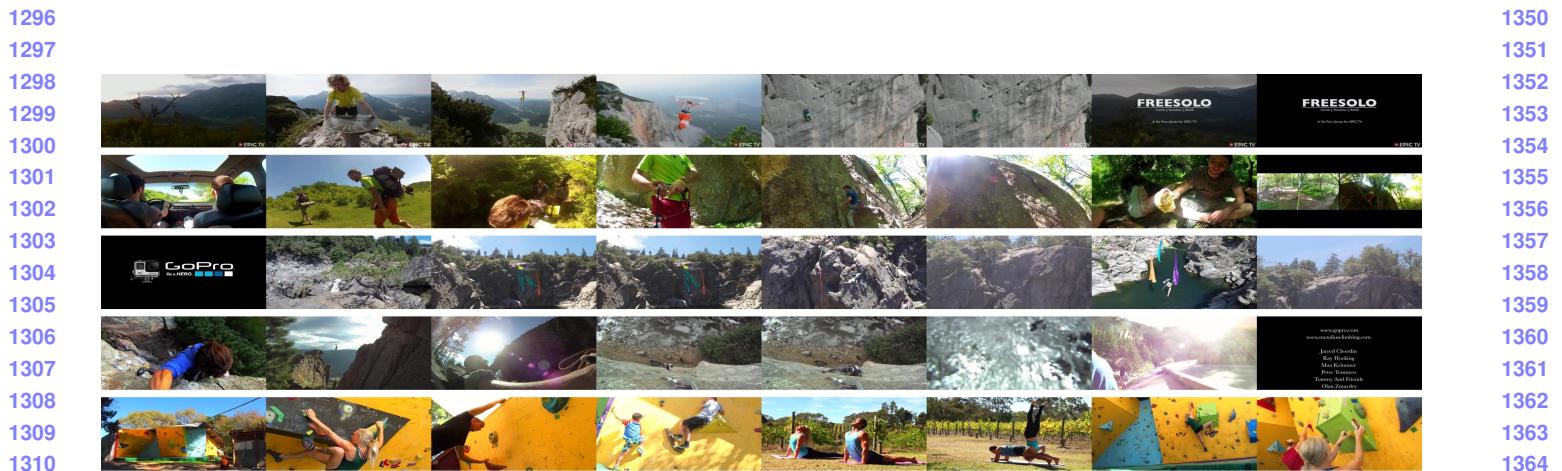


(b) Randomly selected frames from videos belonging to class cook fish village (OG1). Each row corresponds to one video.

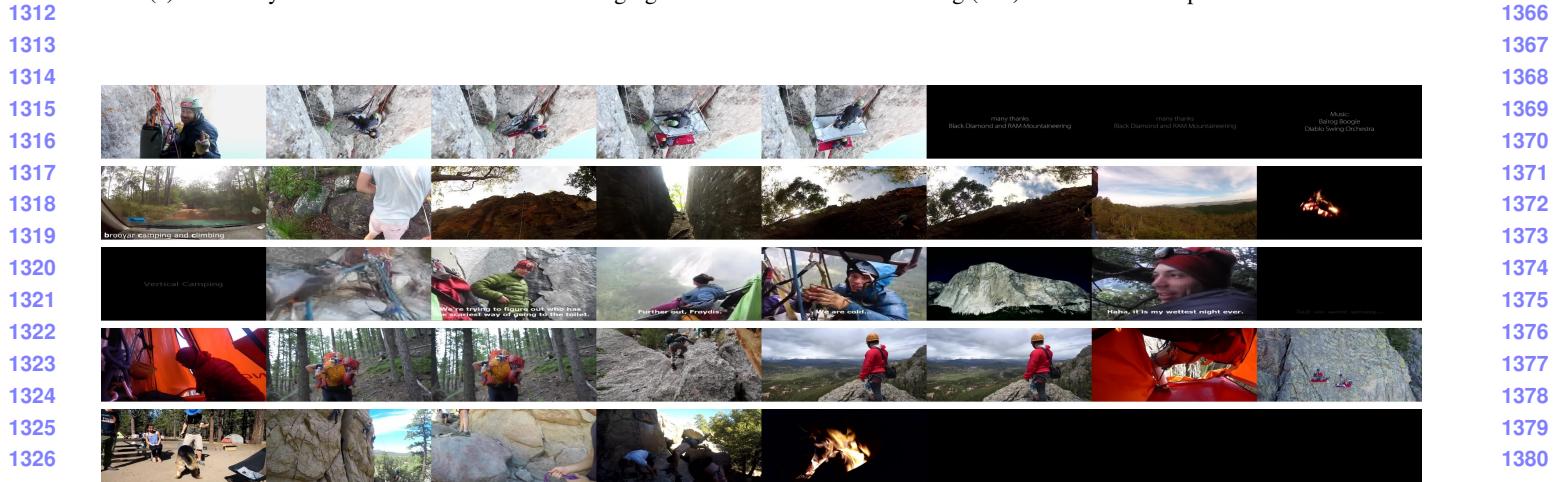


(c) Randomly selected frames from videos belonging to class fishing river (OG2). Each row corresponds to one video.

(a) Randomly selected frames from videos belonging to class helicopter NewYork (**TG**). Each row corresponds to one video.(b) Randomly selected frames from videos belonging to class helicopter Hawaii (**OG1**). Each row corresponds to one video.(c) Randomly selected frames from videos belonging to class NewYork cruise (**OG2**). Each row corresponds to one video.



(a) Randomly selected frames from videos belonging to class slackline rock climbing (TG). Each row corresponds to one video.



(b) Randomly selected frames from videos belonging to class rock climbing camping (OG1). Each row corresponds to one video.



(c) Randomly selected frames from videos belonging to class slcakline juggling (OG2). Each row corresponds to one video.



1419 (a) Randomly selected frames from videos belonging to class ride horse safari (**TG**). Each row corresponds to one video.  
1420  
1421  
1422



1441 (b) Randomly selected frames from videos belonging to class ride horse mountain (**OG1**). Each row corresponds to one video.  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451



1452 (c) Randomly selected frames from videos belonging to class ride vehicle safari (**OG2**). Each row corresponds to one video.  
1453  
1454  
1455  
1456  
1457

1512	<b>References</b>	1566
1513		1567
1514	[1] W.-S. Chu, Y. Song, and A. Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In <i>CVPR</i> , 2015. 1, 2, 3	1568
1515	[2] E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In <i>CVPR</i> , 2012. 2	1569
1516		1570
1517	[3] A. Kulesza, B. Taskar, et al. Determinantal point processes for machine learning. <i>Foundations and Trends® in Machine Learning</i> , 5(2–3):123–286, 2012. 1	1571
1518		1572
1519	[4] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In <i>CVPR</i> , 2013. 1	1573
1520	[5] R. Panda, A. Das, Z. Wu, J. Ernst, and A. K. Roy-Chowdhury. Weakly supervised summarization of web videos. In <i>ICCV</i> , 2017. 2	1574
1521	[6] R. Panda and A. K. Roy-Chowdhury. Collaborative summarization of topic-related videos. In <i>CVPR</i> , 2017. 1, 2, 3	1575
1522		1576
1523		1577
1524		1578
1525		1579
1526		1580
1527		1581
1528		1582
1529		1583
1530		1584
1531		1585
1532		1586
1533		1587
1534		1588
1535		1589
1536		1590
1537		1591
1538		1592
1539		1593
1540		1594
1541		1595
1542		1596
1543		1597
1544		1598
1545		1599
1546		1600
1547		1601
1548		1602
1549		1603
1550		1604
1551		1605
1552		1606
1553		1607
1554		1608
1555		1609
1556		1610
1557		1611
1558		1612
1559		1613
1560		1614
1561		1615
1562		1616
1563		1617
1564		1618
1565		1619