Viewpoint-aware Video Summarization (Supplemental Material)

A. 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 ϵ 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 $k(\mathbf{x}, \mathbf{x}) = \mathbf{1}$ 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.

B. Further results of user study

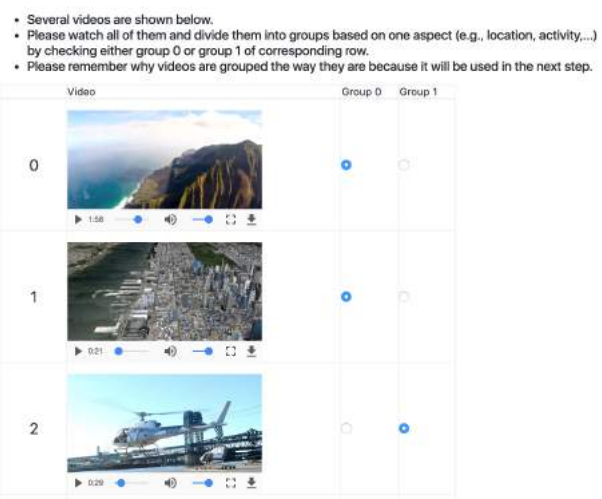
In the main paper, we fixed the *viewpoint*, and we compared the generated summaries with the ones created based on one explicit concept, which can be expressed with a few words, due to the difficulty of quantitative evaluation. We also conducted user study that measures the ability to estimate underlying *viewpoint* with weaker constraint using the same dataset. For this purpose, we developed AMT-like web page as shown in Fig. 1a and Fig. 1b.

Firstly, four videos were randomly picked from each of **TG**, **RG1**, **RG2**, and they were shown to the subjects. Subjects were asked to split them into two groups based on one criterion 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 criterion that was used to divide videos into groups was selected. (It was allowed to choose multiple summaries. Moreover, if there were not appropriate one, subjects do not need to choose anything.) For each task, five workers were assigned.

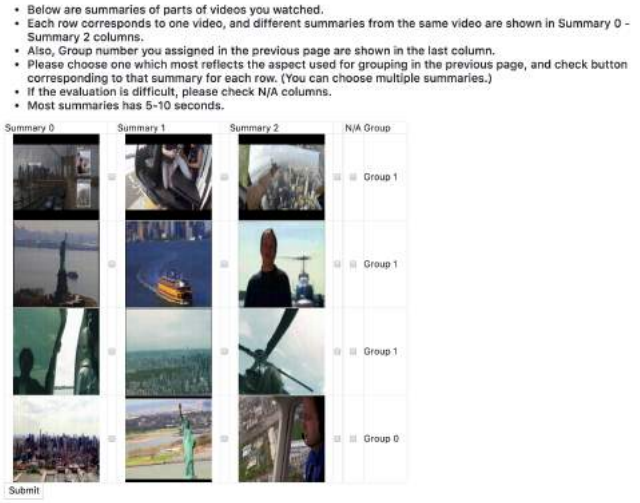
Table 1: The ratio that the summary generated from each method were selected. N/A means no method were selected.

	N/A	MBF [1]	CVS [6]	ours
score	0.09	0.37	0.38	**0.50**

We show the number that each method were selected divided by the number of videos in the Table 1, and the score of our method is better than the others in it. This result indicates that our method can generate the summary that explains the criteria of grouping when the *viewpoint* changes person to person.



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



(b) The screenshot image of the web page used for the evaluation of summaries.

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

C. Further results of quantitative experiments

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

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

	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	0.393	0.295	0.337	0.280	0.335	**0.434**	0.400	0.289	0.349
CS	0.344	0.344	0.326	0.333	0.319	0.304	0.330	0.384	**0.450**	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	**0.404**	0.381	0.358	0.387	0.374	0.376	0.408	0.380	0.382
WSVS [5]	0.358	0.303	0.356	0.353	0.318	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	**0.404**	0.393	0.366	0.423	0.338	0.540	0.412	0.386	0.387	0.375	0.402
ours (feature learning)	0.395	**0.407**	0.335	**0.430**	**0.363**	**0.545**	**0.423**	0.375	0.399	**0.393**	**0.406**

D. Detailed result of topic selection task

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 summary generated by our algorithm is correctly answered with higher probability than other methods, which demonstrates the ability to recover the criteria of grouping. The performance of MBF was near random rate (0.5), and worse than that in several groups. We conjecture the reason attributes to the fact that MBF uses only two videos to find the visual co-occurrence. If the feature representation of shots which is representative to topics are similar each other, it may fail to find the common pattern within the group.

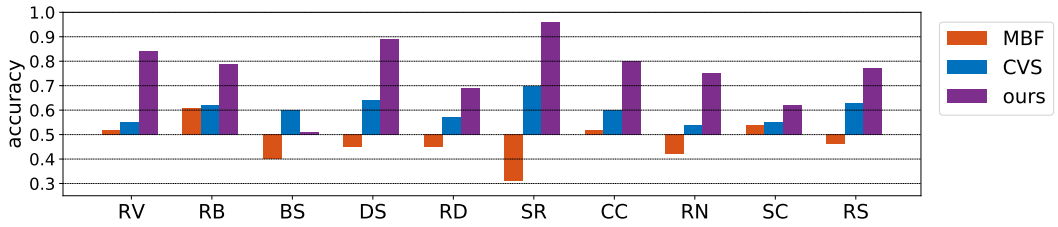


Figure 2: Per-group accuracy of topic selection task. Each bar corresponds to the each method, namely, MBF [1] (orange), CVS [6] (blue), and ours (purple). Please note 0.5 (random rate) are set to the center of this graph. For referring to the abbreviated names of groups, please see the Table 1 in the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

E. Additional Analysis

Applicability for long videos: To investigate the applicability of the proposed method to long videos, 2 out of 5 videos in each group were expanded to 5 times longer by synthesizing it with randomly selected clips in other irrelevant videos and set their scores to 0. The top-5 mAP of MBF, CSV, and ours got 0.217, 0.221, and 0.275 respectively. Results showed the applicability of proposed algorithm for long videos.

Comparison with human performance: We also compared the performance with that of the summary created by human. Treating a summary for one user as a prediction, we computed mAP in the same way with the main experiment, and we regarded the human performance by averaging them. The average score of the human summary was 0.456, and 0.498 respectively. The performance of our method was approximately 80% compared with it.

Computation time: Average computation time per video of MBF, CVS, ours, and ours (feature learning) are 0.02(s), 36.82(s), 42.82(s), and 3562.34(s) with 1 CPU (Intel Xeon, 2.60GHz) and 2 GPUs (Tesla K40).

Ablation study: The top-5 mAP when dropping λ_1, λ_2, λ_3, and nothing are 0.370, 0.365, 0.336, and 0.379, which reveals the importance of discriminativeness.

Choosing hyper-parameters and their sensitivity: Fixing λ_3 to 1.0, and empirically setting λ_1 to 0.05, we changed λ_2 in [0.0, 1.0] at 0.1 interval. and we found performance is not sensitive to λ_2 unless it reaches to 0.0 or 1.0. For fair comparison, we showed the performance of best parameter ($\lambda_2 = 0.1$) in the same way as other methods.

F. Detailed derivation of equations

F.1. Trace of inner-video variance

$$
\begin{aligned}
\mathrm{Tr}(S_i^V) &= \mathrm{Tr}(\sum_{t=1}^{T_i} z_t(\mathbf{x}_t - \mathbf{v}_i)(\mathbf{x}_t - \mathbf{v}_i)^\top) && (1)\\
&= \sum_{t=1}^{T_i} \mathrm{Tr}(z_t(\mathbf{x}_t - \mathbf{v}_i)(\mathbf{x}_t - \mathbf{v}_i)^\top) && (2)\\
&= \sum_{t=1}^{T_i} z_t(\mathbf{x}_t - \mathbf{v}_i)^\top(\mathbf{x}_t - \mathbf{v}_i) && (3)\\
&= \sum_{t=1}^{T_i} z_t\mathbf{x}_t^\top\mathbf{x}_t - 2\mathbf{v}_i^\top\sum_{t=1}^{T_i} z_t\mathbf{x}_t + \sum_{t=1}^{T_i} z_t\mathbf{v}_i^\top\mathbf{v}_i && (4)\\
&= \sum_{t=1}^{T_i} z_t\mathbf{x}_t^\top\mathbf{x}_t - \frac{2}{s}\mathbf{z}_i^\top\mathbf{X}_i\mathbf{X}_i^\top\mathbf{z}_i + \frac{1}{s}\mathbf{z}_i^\top\mathbf{X}_i\mathbf{X}_i^\top\mathbf{z}_i && (5)\\
&= \sum_{t=1}^{T_i} z_t\mathbf{x}_t^\top\mathbf{x}_t - \frac{1}{s}\mathbf{z}_i^\top\mathbf{X}_i\mathbf{X}_i^\top\mathbf{z}_i && (6)
\end{aligned}
$$

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

F.2. Trace of within-class variance

$$
\begin{aligned}
\mathrm{Tr}(S_{(k)}^W) &= \mathrm{Tr}(\sum_{i\in L_{(k)}} s(\mathbf{v}_i - \boldsymbol{\mu}_k)(\mathbf{v}_i - \boldsymbol{\mu}_k)^\top) && (7)\\
&= \sum_{i\in L_{(k)}} s(\mathbf{v}_i - \boldsymbol{\mu}_k)^\top(\mathbf{v}_i - \boldsymbol{\mu}_k) && (8)\\
&= s\sum_{i\in L_{(k)}} \mathbf{v}_i^\top\mathbf{v}_i - 2s(\sum \mathbf{v}_i)^\top\boldsymbol{\mu}_k + n_k s\boldsymbol{\mu}_k^\top\boldsymbol{\mu}_k && (9)\\
&= \frac{1}{s}\sum_{i\in L_{(k)}} \mathbf{z}_i^\top\mathbf{X}_i\mathbf{X}_i^\top\mathbf{z}_i - \frac{2}{n_k s}\hat{\mathbf{z}}_{(k)}^\top\hat{\mathbf{X}}_{(k)}\hat{\mathbf{X}}_{(k)}^\top\hat{\mathbf{z}}_{(k)} + \frac{1}{n_k s}\hat{\mathbf{z}}_{(k)}^\top\hat{\mathbf{X}}_{(k)}\hat{\mathbf{X}}_{(k)}^\top\hat{\mathbf{z}}_{(k)} && (10)\\
&= \frac{1}{s}\sum_{i\in L_{(k)}} \mathbf{z}_i^\top\mathbf{X}_i\mathbf{X}_i^\top\mathbf{z}_i - \frac{1}{n_k s}\hat{\mathbf{z}}_{(k)}^\top\hat{\mathbf{X}}_{(k)}\hat{\mathbf{X}}_{(k)}^\top\hat{\mathbf{z}}_{(k)} && (11)
\end{aligned}
$$

F.3. Trace of between-class variance

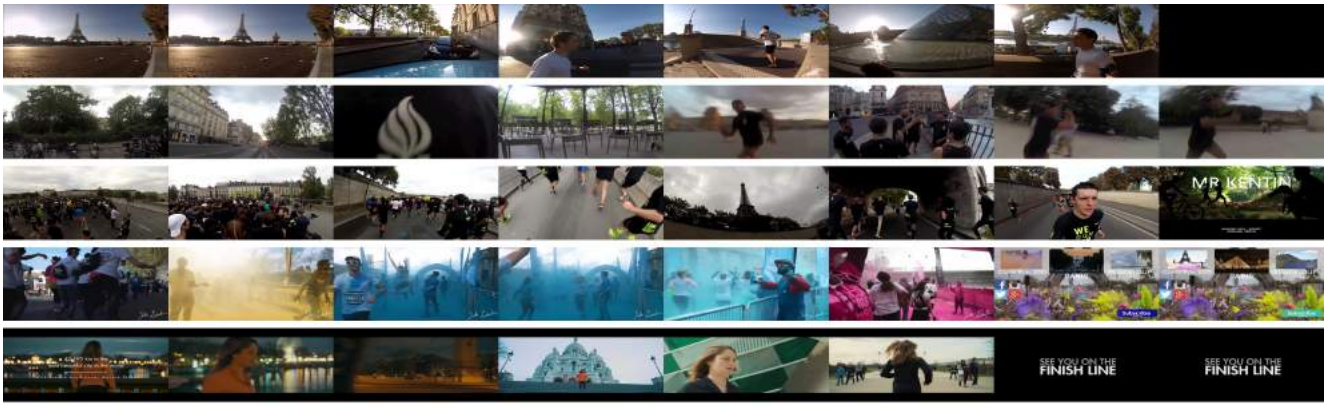
$$
\begin{aligned}
\mathrm{Tr}(S^B) &= \mathrm{Tr}(\sum_{k=1}^{K} n_k s(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})^\top) \\
&= \sum_{k=1}^{K} n_k s(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})^\top(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}}) && (12) \\
&= s\sum_{k=1}^{K} n_k \boldsymbol{\mu}_k^\top \boldsymbol{\mu}_k - 2s\bar{\boldsymbol{\mu}}^\top(\sum_{k=1}^{K} n_k \boldsymbol{\mu}_k) + Ns\bar{\boldsymbol{\mu}}^\top\bar{\boldsymbol{\mu}} && (13) \\
&= \frac{1}{s}\sum_{k=1}^{K}\frac{1}{n_k}\hat{\mathbf{z}}_{(k)}^\top\hat{\mathbf{X}}_{(k)}\hat{\mathbf{X}}_{(k)}^\top\hat{\mathbf{z}}_{(k)} - \frac{2}{Ns}\hat{\mathbf{z}}^\top\hat{\mathbf{X}}\hat{\mathbf{X}}^\top\hat{\mathbf{z}} + \frac{1}{Ns}\hat{\mathbf{z}}^\top\hat{\mathbf{X}}\hat{\mathbf{X}}^\top\hat{\mathbf{z}} && (14) \\
&= \hat{\mathbf{z}}^\top(\frac{1}{s}\oplus\sum_{k=1}^{K}\frac{1}{n_k}\hat{\mathbf{X}}_{(k)}\hat{\mathbf{X}}_{(k)}^\top)\hat{\mathbf{z}} - \frac{1}{Ns}\hat{\mathbf{z}}^\top\hat{\mathbf{X}}\hat{\mathbf{X}}^\top\hat{\mathbf{z}} && (15) \\
&= \hat{\mathbf{z}}^\top(C - A)\hat{\mathbf{z}} && (16)
\end{aligned}
$$

(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).

G. Examples of dataset

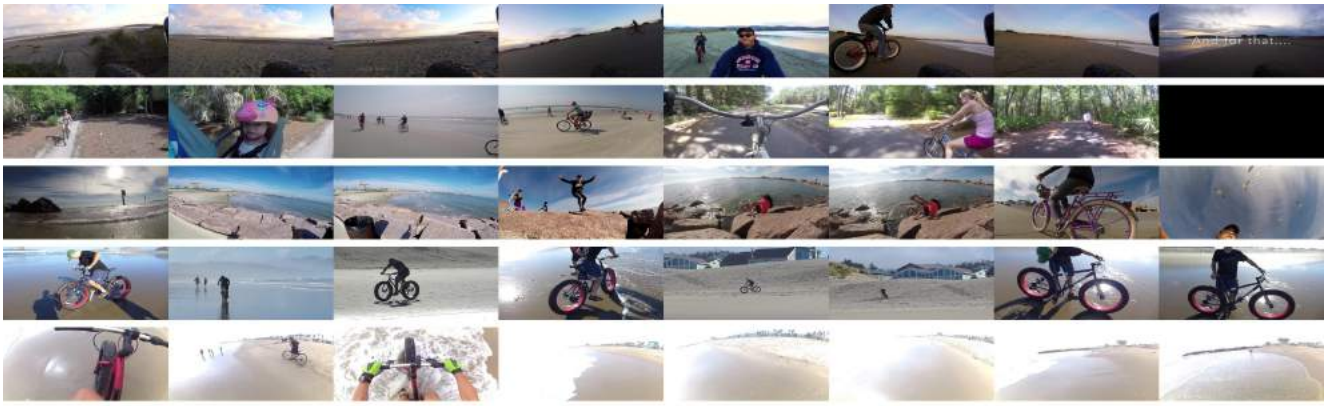
We show randomly selected frames of videos of our dataset in the following figures. The order of figure corresponds to the ones written in the Table. 1 in the main paper, namely, in the order of **TG**, **RG1**, **RG2**, and from top-row to bottom-row. Each row of figures corresponds to one video.

(a) Randomly selected frames from videos belonging to class run venice (**TG**). Each row corresponds to one video.

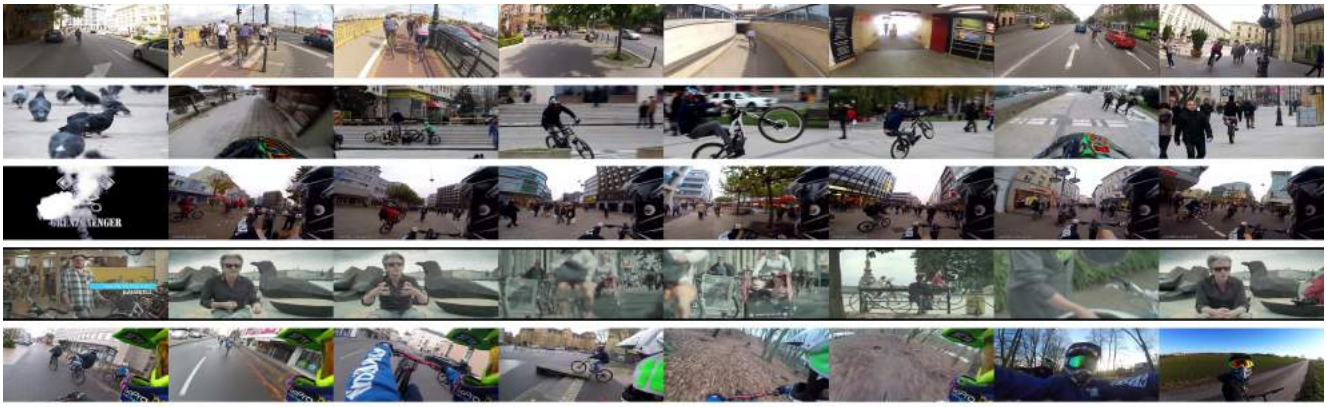


(b) Randomly selected frames from videos belonging to class run paris (**OG1**). Each row corresponds to one video.

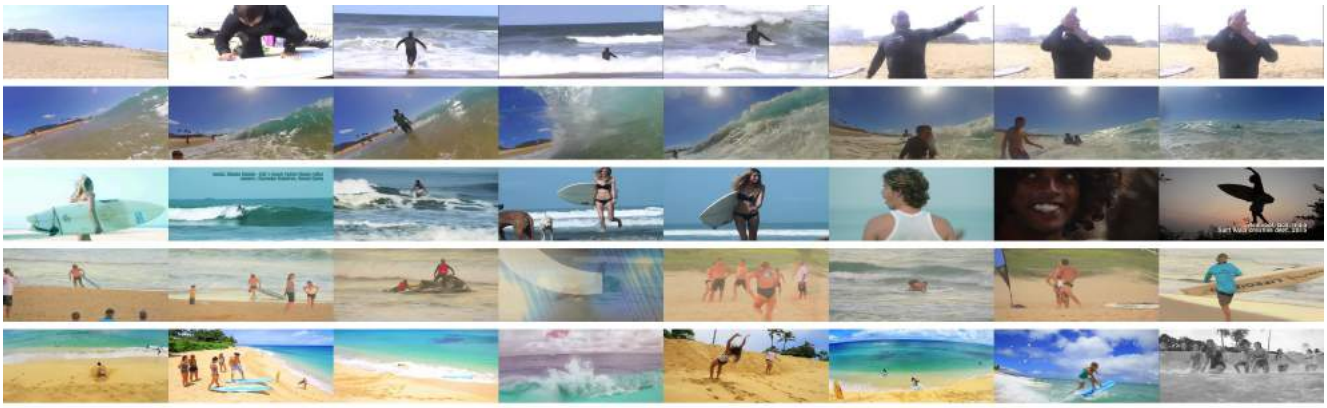
(c) Randomly selected frames from videos belonging to class shopping venice (**OG2**). Each row corresponds to one video.



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



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



(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.

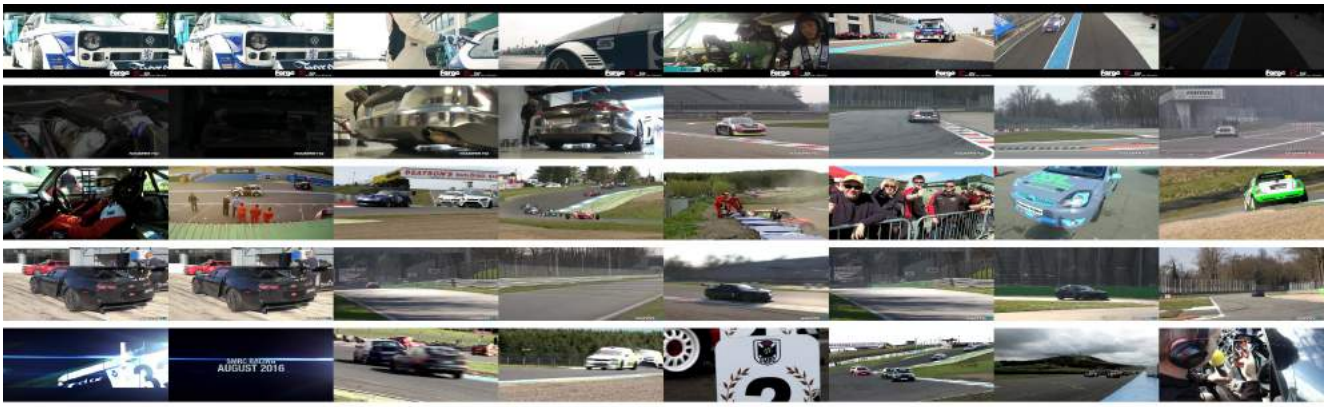
(c) Randomly selected frames from videos belonging to class hiking snow mountain (**OG2**). Each row corresponds to one video.

(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.

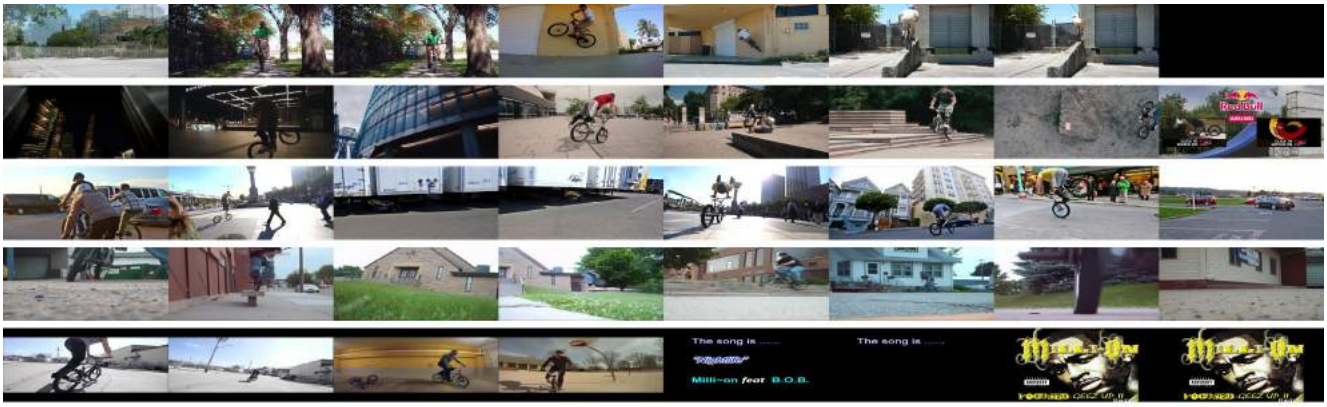
(a) Randomly selected frames from videos belonging to class racing desert (**TG**). Each row corresponds to one video.



(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.



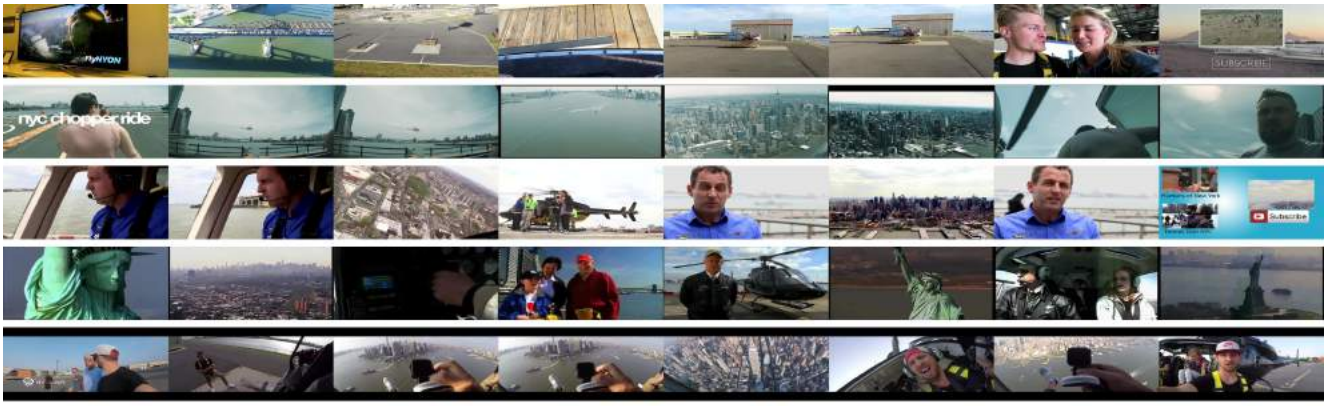
(b) Randomly selected frames from videos belonging to class riding bike trick (**OG1**). Each row corresponds to one video.

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

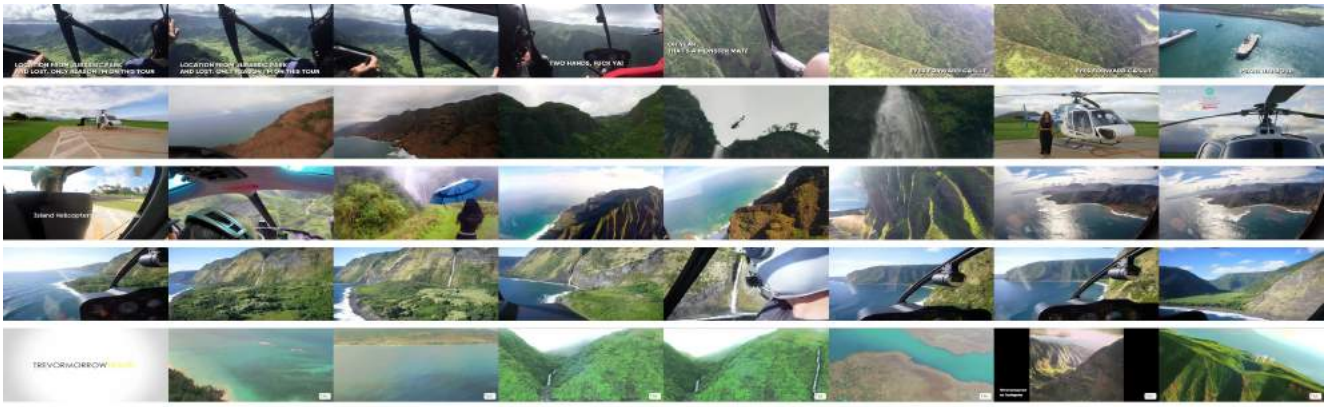
(a) Randomly selected frames from videos belonging to class fishing cook fish (**TG**). Each row corresponds to one video.

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

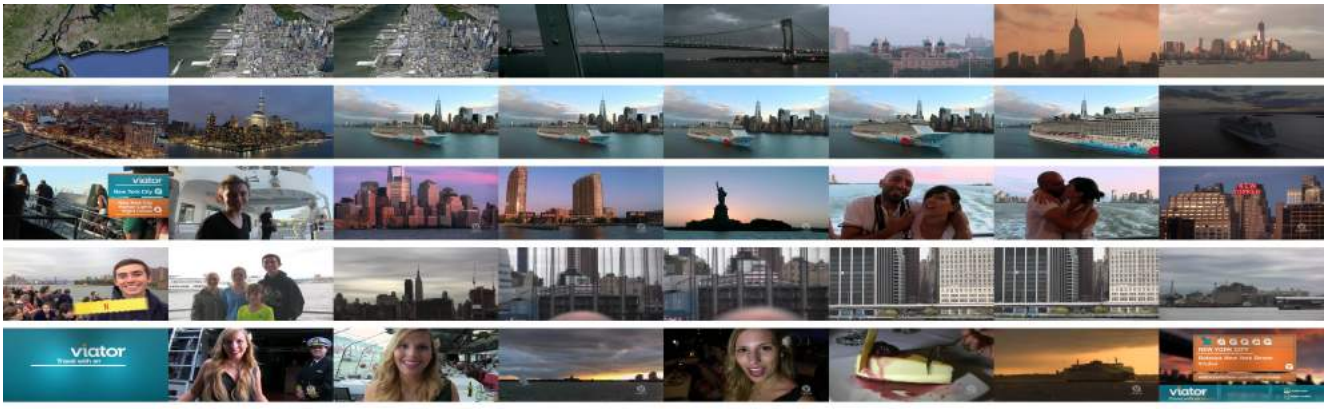
(c) Randomly selected frames from videos belonging to class fishing river (**OG2**). Each row corresponds to one video.



(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 cruse (**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 jaggling (**OG2**). Each row corresponds to one video.

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

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

(c) Randomly selected frames from videos belonging to class ride vehicle safari (**OG2**). Each row corresponds to one video.

References

[1] W.-S. Chu, Y. Song, and A. Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *CVPR*, 2015. 1, 2, 3

[2] E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *CVPR*, 2012. 2

[3] A. Kulesza, B. Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012. 1

[4] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *CVPR*, 2013. 1

[5] R. Panda, A. Das, Z. Wu, J. Ernst, and A. K. Roy-Chowdhury. Weakly supervised summarization of web videos. In *ICCV*, 2017. 2

[6] R. Panda and A. K. Roy-Chowdhury. Collaborative summarization of topic-related videos. In *CVPR*, 2017. 1, 2, 3