FFNet: Video Fast-Forwarding via Reinforcement Learning (Supplementary Material)

Shuyue Lan [1] Rameswar Panda [2] Qi Zhu [1] Amit K. Roy-Chowdhury [2]
[1]Northwestern University. {slan@u., qzhu@}northwestern.edu
[2]University of California, Riverside. {rpand002@, amitrc@ece.}ucr.edu

1. Detailed information on the experimental datasets

1.1. Tour20 Dataset

To the best of the authors' knowledge, Tour20 dataset [5] is the largest publicly available summarization dataset, which contains 140 videos about 20 tourist attractions selected from the Tripadvisor travelers choice landmarks 2015 list. Three user summaries are offered with the segmentation files. The descriptive statistics of Tour20 dataset is shown in Table. 1. In Fig. 1 and Fig. 2, we show topic-wise example images from Tour20 dataset.

Tourist Attractions	**# Videos**	**Length**	**# Frames**	**# Segments**
Angkor Wat, Cambodia (AW)	7	26m 57s	44,410	803
Machu Picchu, Peru (MP)	7	26m 15s	43,125	914
Taj Mahal, India (TM)	7	22m 21s	36,554	705
Basilica of the Sagrada Familia, Spain (BF)	6	23m 30s	22,641	400
St. Peter's Basilica, Italy (SB)	5	14m 39s	23,777	406
Milan Cathedral, Italy (MC)	10	24m 18s	37,749	768
Alcatraz, United States (AT)	6	05m 22s	09,733	223
Golden Gate Bridge, United States (GB)	6	19m 21s	33,063	521
Eiffel Tower, Paris (ET)	8	16m 10s	26,071	495
Notre Dame Cathedral, Paris (NC)	8	26m 49s	44,583	862
The Alhambra, Spain (TA)	6	21m 20s	38,087	779
Hagia Sophia Museum, Turkey (HM)	6	24m 27s	38,608	853
Charles Bridge, Prague (CB)	6	27m 33s	48,395	769
Great Wall at Mutiantu, Beijing (GM)	5	13m 16s	22,117	477
Burj Khalifa, Dubai (BK)	9	23m 21s	40,557	809
Wat Pho, Bangkok (WP)	5	11m 48s	20,461	382
Chichen Itza, Mexico (CI)	8	16m 51s	28,737	545
Sydney Opera House, Sydney (SH)	10	25m 55s	49,735	695
Petronas Twin Towers, Malaysia (PT)	9	18m 32s	30,009	470
Panama Canal, Panama (PC)	6	17m 33s	31,625	623
Total	**140**	**6h 46m 18s**	**669,497**	**12,499**

Table 1. Descriptive statistics of Tour20 dataset.

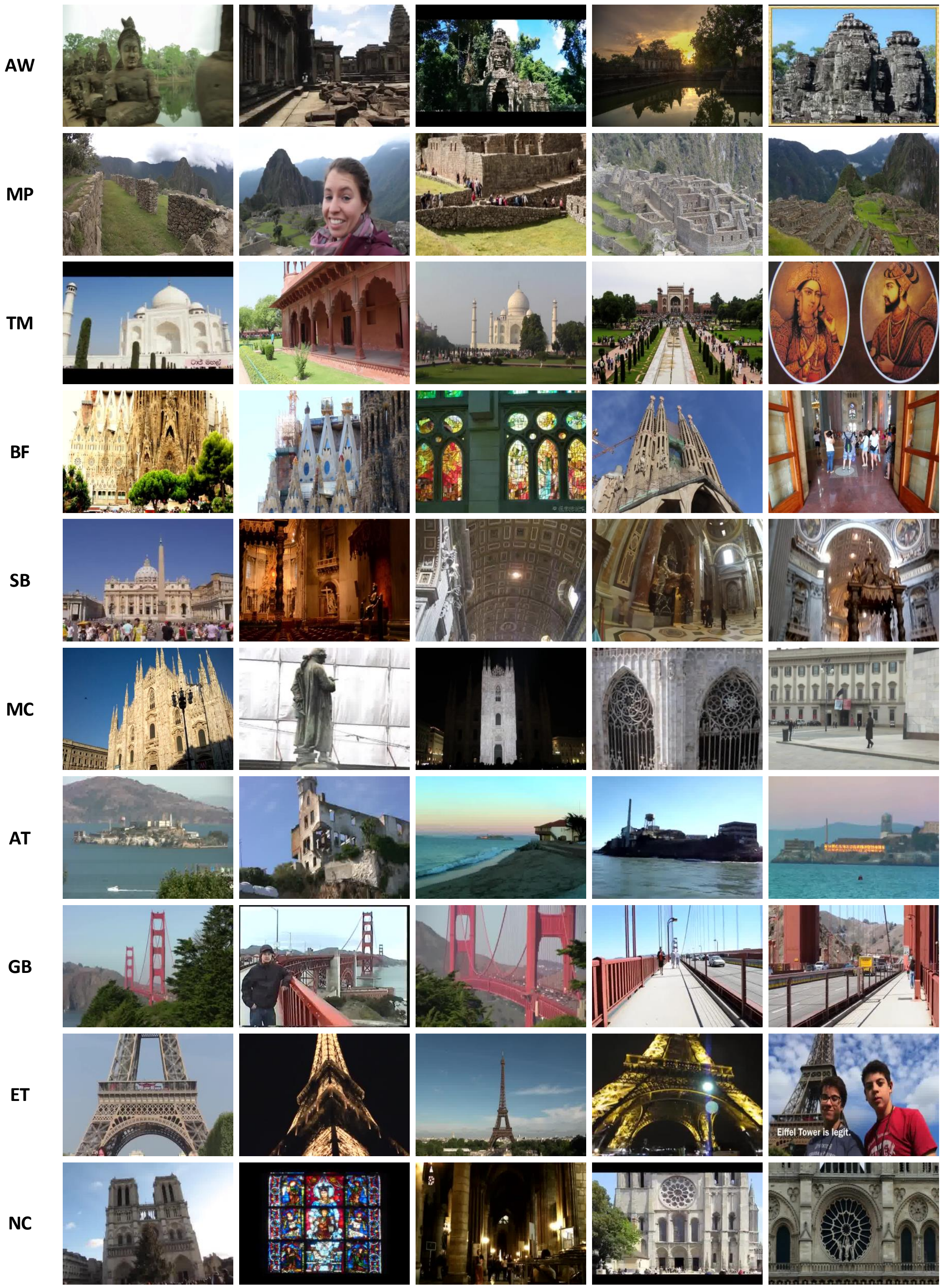

Figure 1. **Topic-wise example images from Tour20 dataset - 1.**

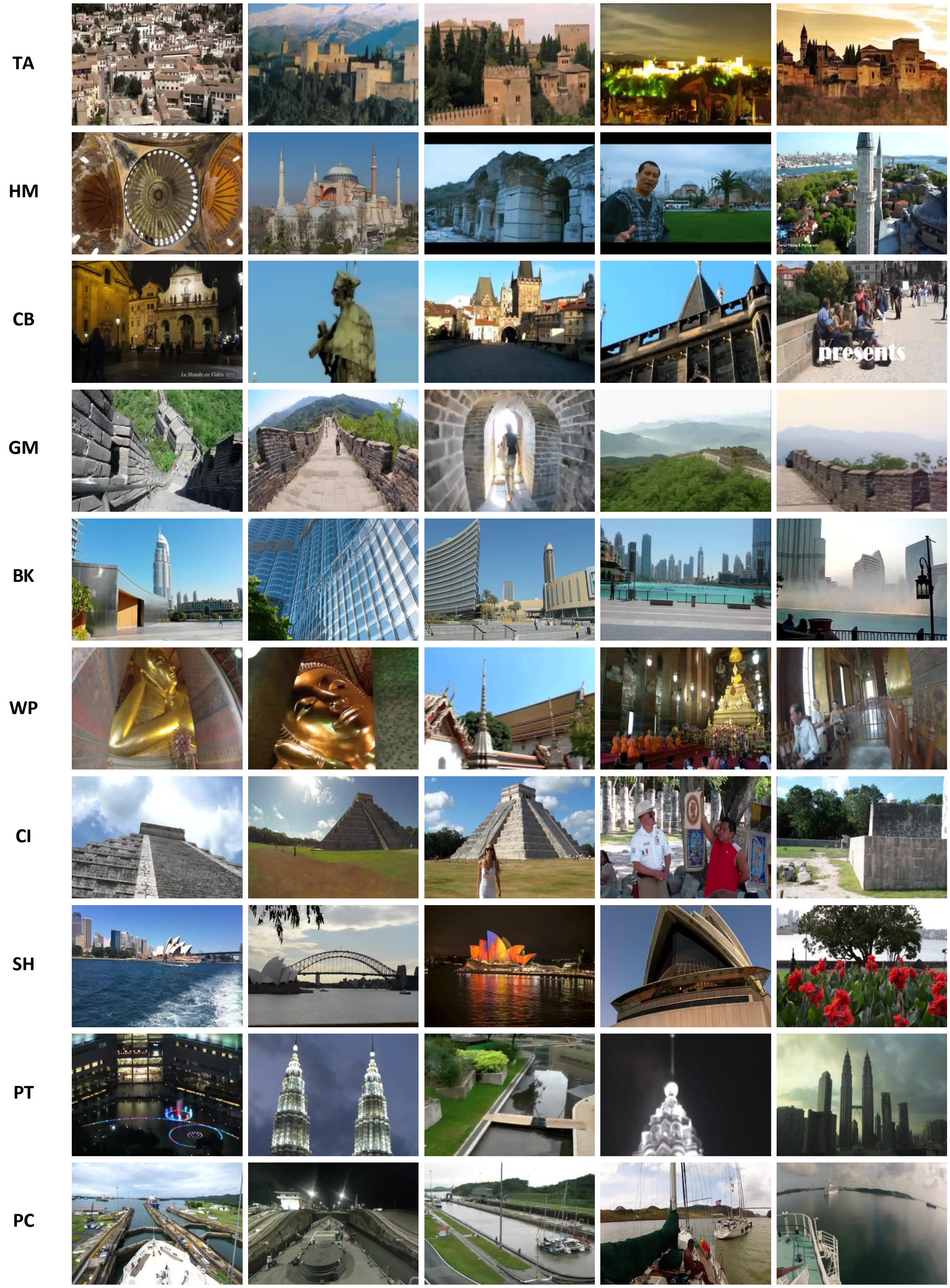

Figure 2. **Topic-wise example images from Tour20 dataset - 2.**

1.2. TVSum Dataset

TVSum [6] dataset consists of 50 videos downloaded from YouTube in 10 categories, as defined in the TRECVid Multimedia Event Detection task. Frame-level importance scores are offered from 20 labelers. The descriptive statistics of TVSum dataset is shown in Table. 2. In Fig. 3, we show topic-wise example images from TVSum dataset.

Video Topics	# Videos	Length	# Frames	# Segments
Changing Vehicle Tire (VT)	5	25m 25s	39,841	911
Getting Vehicle Unstuck (VU)	5	19m 28s	35,014	841
Grooming an Animal (GA)	5	18m 07s	30,920	767
Making Sandwich (MS)	5	24m 58s	37,095	770
ParKour (PK)	5	24m 50s	41,634	993
PaRade (PR)	5	25m 03s	44,042	715
Flash Mob Gathering (FM)	5	18m 37s	30,747	618
Bee Keeping (BK)	5	17m 30s	30,489	678
Attempting Bike Tricks (BT)	5	14m 39s	25,747	523
Dog Show (DS)	5	20m 59s	36,827	754
Total	**50**	**3h 29m 42s**	**352,356**	**7570**

Table 2. Descriptive statistics of TVSum dataset.

2. Implementation details on the baseline methods for comparison

As described in Section 4.1 (Experimantal Setup) of the main paper, we compare our approach with several baseline methods that fall into 2 main categories: (1) offline approaches including SMRS [2], SC [7] and MH [3]; and (2) online methods including LL [8] and OK [1]. For all the algorithms, we generate a subset of video frames that have the same average length as in ground truth to make a fair comparison. Alexnet [4] fc7 features (4096-dimensional) is used to represent each video frame for all methods except MH, as MH is already integrated into Microsoft Hyperlapse and we use it directly.

2.1. Offline methods

- **SMRS:** Sparse Modeling Representative Selection (SMRS) uses the entire video as the dictionary and finds the representative frames based on the zero patterns of the coding vector. Mathematically, it solves the following optimization problem to get a selection matrix $Z \in \mathbb{R}^{n \times n}$:

$$\min_{z} ||Y - YZ||_F^2 + \lambda ||Z||_{2,1} \tag{1}$$

where $Y \in \mathbb{R}^{d \times n}$ is the video feature matrix in which each column represent a frame. $||Z||_{2,1} = \sum_{i=1}^{n} ||Zi,.||_2$ denotes the $L_{2,1}$-norm and $||Zi,.||_2$ is the L_2-norm of the i-th row of Z. $\lambda > 0$ is a regularization parameter that controls the level of sparsity in the reconstruction. Eqn.(1) is solved

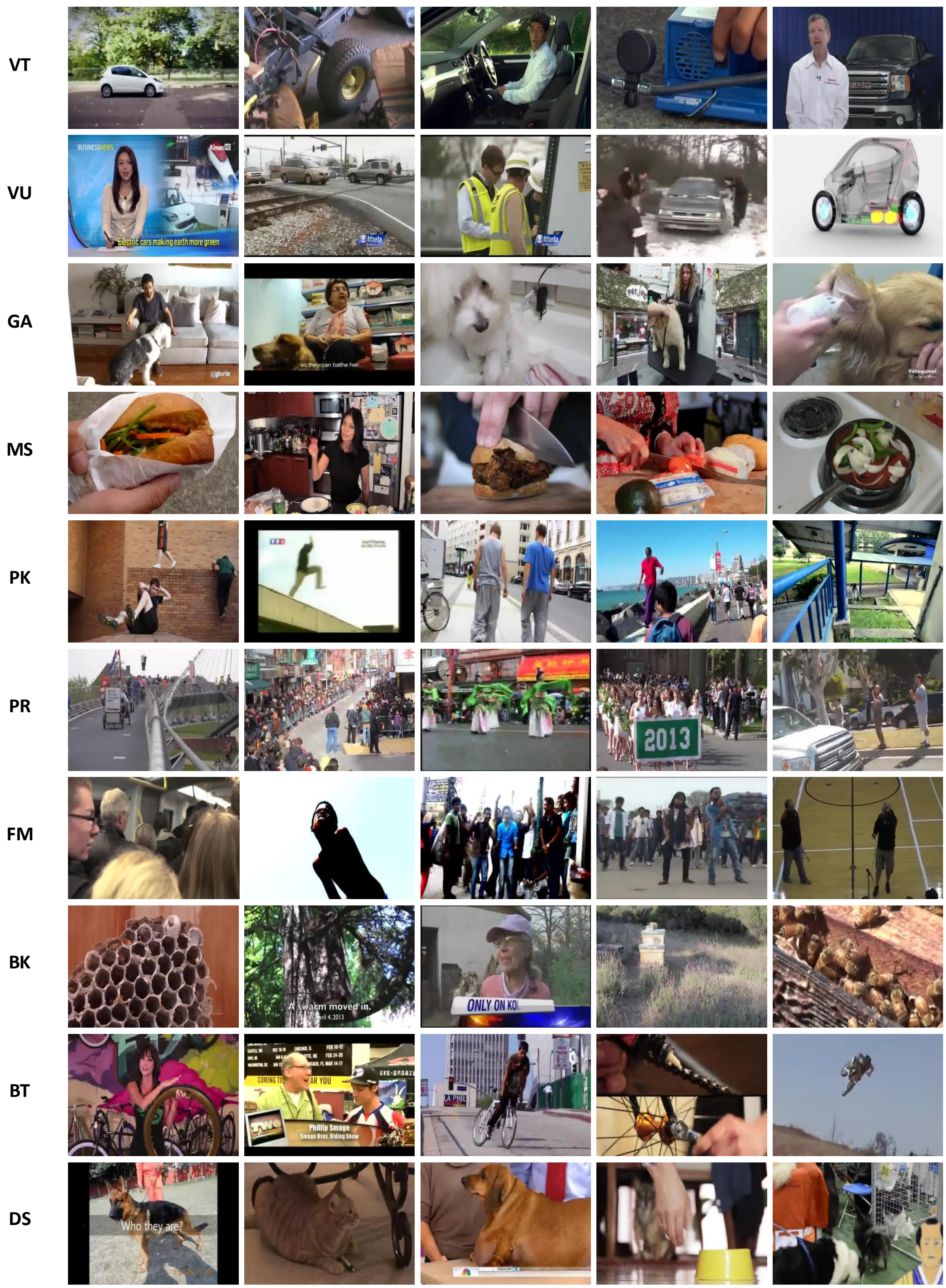

Figure 3. **Topic-wise example images from TVSum dataset.**

by Alternating Direction of Method of Multipliers (ADMM) and select the representative frames that have $||Zi,.||_2 \neq 0$.

- **SC:** Spectral Clustering (SC) is a classical clustering based method that clusters all the frames in a video to several clusters. In the experiments, we set the number of clusters to 20 and select the same number of frames inside each class, which are closest to the corresponding centroid.

- **MH:** Microsoft Hyperlapse (MH) is a well-developed software tool for generating smooth and stabilized time lapses from first-person videos, which provide multiple speed-up options. We use the desktop version to generate the videos with a $4\times$ speed-up factor.

2.2. Online methods

- **LL:** LiveLight (LL) is an online vdieo summarization method working in a quasi real-time fashion. It uses an online variant of sparse coding to generate a video skim over time, by computing the redundancy with an online updating dictionary of segments. Due to the unavailability of the source code, we implemented it with a dictionary initialized as the first 10% of segments in a video.

- **OK:** Online Kmeans (OK) is a classical clustering based method working in an online update phase. In the experiments, the number of clusters is set to 20, and key frames are selected as the frames that are closest to the centroid in every cluster.

3. Additional experimental results

As shown in Section 4.2 (Coverage Evaluation) of the main paper, we evaluate our approach qualitatively with other baseline methods. Here we show more result examples for illustrating the effectiveness of our method.

For the following Fig. 4 - Fig. **??**, the frames on top represent segments in our FFNet fast-forwarding results. The rows below illustrate (in order) the selected portions from FFNet, ground truth (GT), LiveLight (LL), Microsoft Hyperlapse (MH), Spectral Clustering (SC), Online Kmeans (OK), and Sparse Modeling Representative Selection (SMRS). The X-axis is the frame index over time. Figures are best viewed in color.

3.1. More qualitative examples on the Tour20 dataset

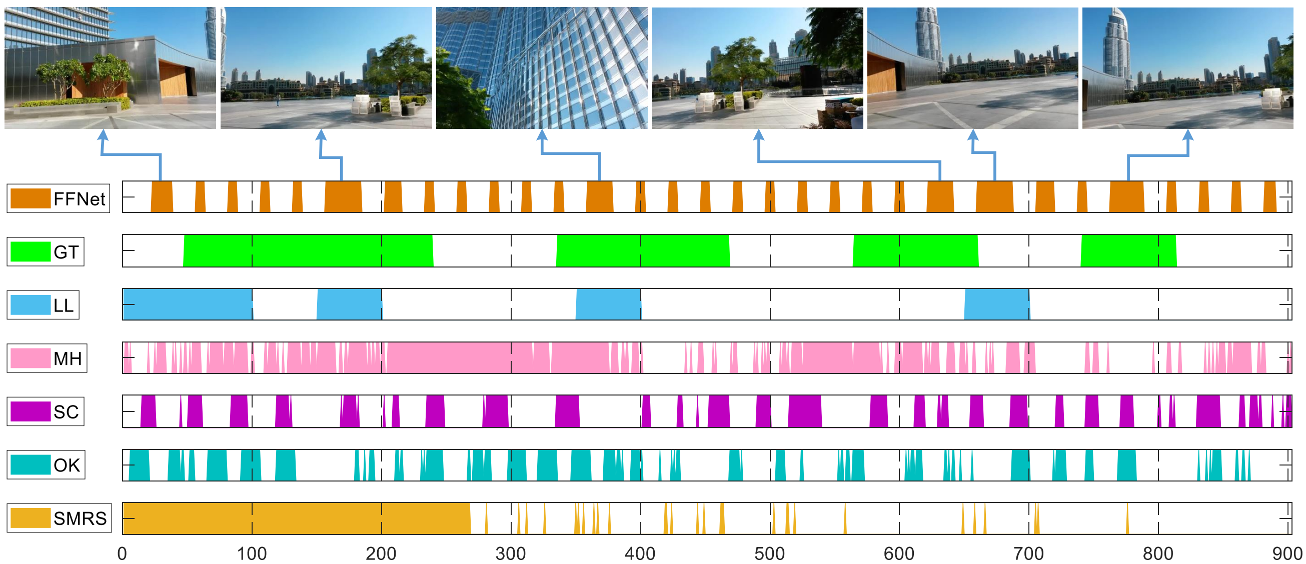

Figure 4. **Qualitative example of a video taken in Burj Khalifa in Tour20 dataset.** This is a short video with 900 frames, capturing the Burj Khalifa building and the surroundings. Our fast-forwarded result captures the zoom-in and zoom-out views of the building and also the square near it.

Figure 5. **Qualitative example of a video taken in Charles Bridge in Tour20 dataset.** In this video, a man playing accordion takes a large portion at the beginning, and then the camera turns to the surroundings near Charles Bridge. Our fast-forwarded result captures different playing actions of the man and various scenes.

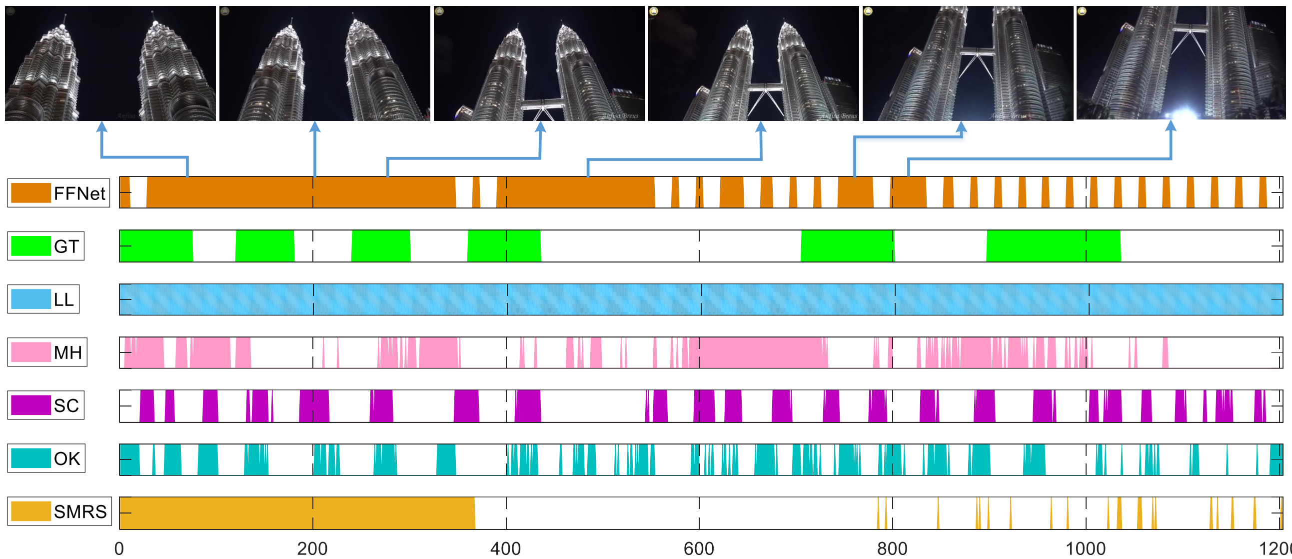

Figure 6. **Qualitative example of a video taken in Petronas Twin Towers in Tour20 dataset.** This is a short video capturing the Petronas Twin Towers from top to bottom. Our FFNet fast-forwarded results captures it from top to bottom too, with much fewer frames. Notice that LL is unable to make a summary as the scene is highly dynamic.

3.2. More qualitative examples on the TVSum dataset

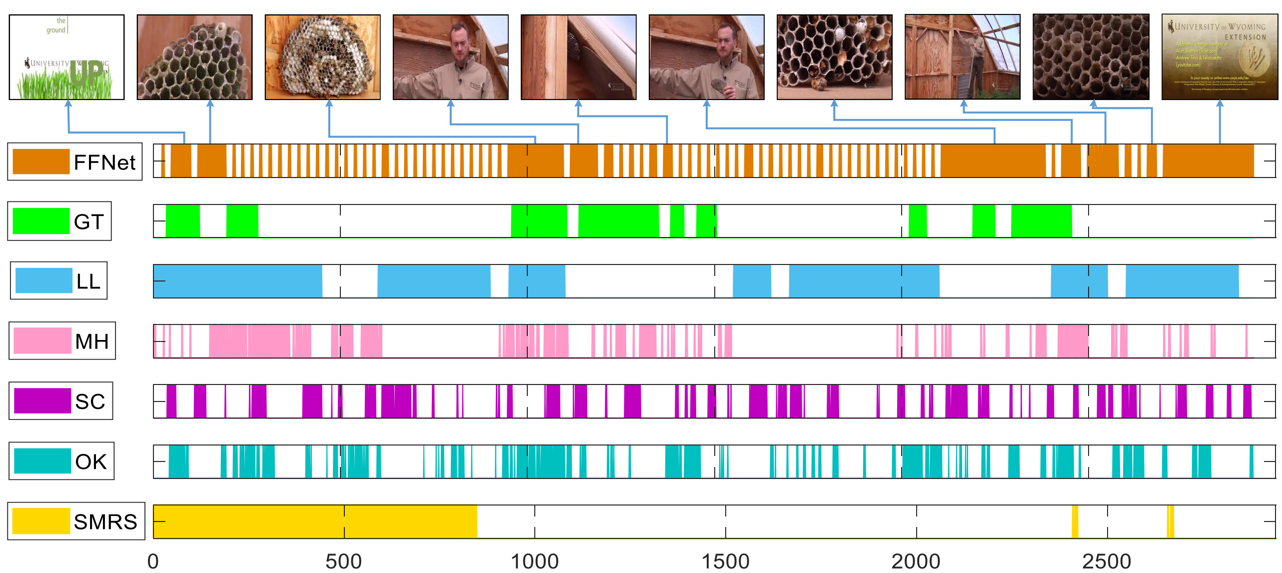

Figure 7. **Qualitative example for a bee-keeping tutorial of TVSum dataset.** FFNet results include multiple segments about the detail of honeycombs and their positions, as well as most of the human speaking portions (which are important in a tutorial).

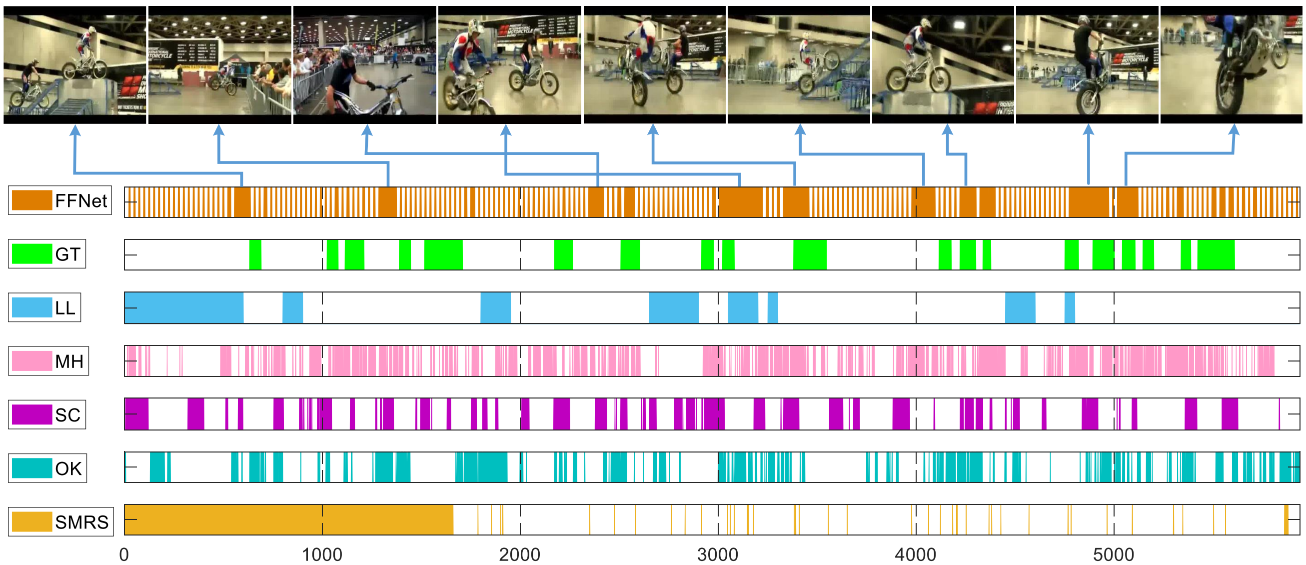

Figure 8. **Qualitative example for a video of attempting bike tricks in TVSum dataset.** Our FFNet fast-forwarded result shows multiple motions of attempting bike tricks with motorbicycles in different shapes of land.

Figure 9. **Qualitative example for a video of changing vehicle tire of TVSum dataset.** Our FFNet fast-forwarded result tells a complete story of changing vehicle tire: (1) find abnormal tire pressure when driving; (2) get off and check; (3) find the problem on tire; (4) get tools from the trunk; (5) shut down the car; (6) fix the problem; (7) get back to road.

4. Detailed information on the user study

As described in Section 4.3 (User Study) of the main paper, we perform an user study involving four human subjects to get a subjective evaluation. Here, we describe the entire study setup:

(a) We first chose a random subset of videos from each dataset, and ran every method on them. All the resulting videos were tuned to the same frame rate as the original video, i.e., 30 fps.

(b) Human subjects were asked to rate the overall quality of each fast-forwarded video by assigning a rating from 1 to 10, where 1 corresponded to "The selected frames are not at all informative in covering the important content from the original video" and 10 corresponded to "The selected frames are extremely informative in covering the important content from the original video" . For each video, the human rating is computed as the averaged rating from all human subjects.

(c) To perform a fair comparison, we provided all the fast-forwarded videos at a time, together with the original video, in a random order without revealing the identity of each method.

(d) It took roughly an hour for each participant to evaluate the results of both datasets, and 4 hours in total to complete the entire study.

References

[1] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[2] E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1600–1607. IEEE, 2012.

[3] N. Joshi, W. Kienzle, M. Toelle, M. Uyttendaele, and M. F. Cohen. Real-time hyperlapse creation via optimal frame selection. *ACM Transactions on Graphics (TOG)*, 34(4):63, 2015.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[5] R. Panda, N. C. Mithun, and A. Roy-Chowdhury. Diversity-aware multi-video summarization. *IEEE Transactions on Image Processing*, 2017.

[6] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5179–5187, 2015.

[7] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[8] B. Zhao and E. P. Xing. Quasi real-time summarization for consumer videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2513–2520, 2014.