

Comprehensive Video Understanding: Video summarization with content-based video recommender design

Yudong Jiang¹, Kaixu Cui¹, Bo Peng^{2‡}, Changliang Xu¹

¹Xinhua Zhiyun Technology Co., Ltd.

²Dept. of Statistics, Columbia University in the City of New York

{jiangyudong, cuikaixu, xuchangliang}@shuwen.com, bp2548@columbia.edu

Abstract

Video summarization aims to extract keyframes/shots from a long video. Previous methods mainly take diversity and representativeness of generated summaries as prior knowledge in algorithm design. In this paper, we formulate video summarization as a content-based recommender problem, which should distill the most useful content from a long video for users who suffer from information overload. A scalable deep neural network is proposed on predicting if one video segment is a useful segment for users by explicitly modelling both segment and video. Moreover, we accomplish scene and action recognition in untrimmed videos in order to find more correlations among different aspects of video understanding tasks. Also, our paper will discuss the effect of audio and visual features in summarization task. We also extend our work by data augmentation and multi-task learning for preventing the model from early-stage overfitting. The final results of our model win the first place in ICCV 2019 CoView Workshop Challenge Track.

1. Introduction

As information overload becomes a more and more serious topic in modern society, many efficient tools have been designed to overcome “information anxiety”. Videos, the fastest-growing information carrier, will account for more than 80% of all Internet traffic by 2020 [3]. Video summarization aims to address this problem by extracting keyframes/shots from a long video, which contains the most useful content for people. It serves as a vital way of comprehensively understanding video data in research area while saving viewers’ time on information acquisition. This technology also gains the attention of both the academic world and the industry. Adobe has already presented video

[‡]Bo Peng is an intern in Xinhua Zhiyun Technology Co., Ltd. when doing this work.

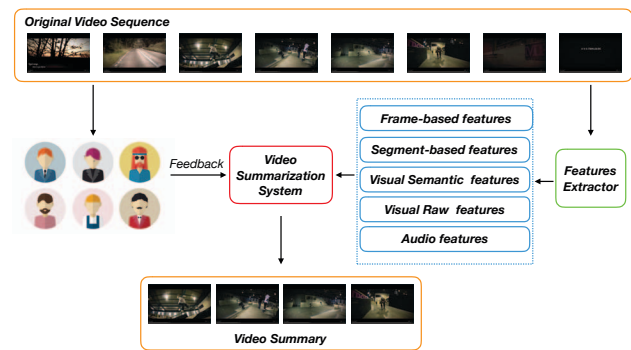


Figure 1. The whole picture of our video summarization system. Feature extractor can be a model trained on different video understanding tasks. Summarization system use viewers’ feedback and features from different source to generate the video summary. Users’ feedback is treated as supervised signals but not a necessary condition in the prediction process.

summarization feature in their products for video editing. Some cloud computing companies, like Microsoft Azure, AliCloud, provide this function as an online service.

One of the main challenges in video summarization is subjectivity, different people may have different selections on key shots for a same video. We investigate two top referenced dataset SumMe [6] and TVSum [15], which both provide manually curated consistency benchmark, with F1 score of 0.31 for SumMe, F1 score of 0.36 for TVSum. Based on the assumptions above, we don’t take diversity or representativeness and other subjective attributes into consideration, instead, we focus on solely the key shots and the full video. The mission of our algorithm is to find the most attractive video segments for a group of annotators/users when they watch a long video.

Content-based recommender modelling is one of the most important technology in recommendation system, especially for alleviating cold start problem, which recommend items with similar content to what users like. It is

usually formalized as a similarity learning problem [13]. A content-based video recommender deep model learns compact representation of videos and constructs a bridge between users’ feedback and video semantics information. We develop and deploy the HighlightNet to learn annotators’ preferences, the model is supervised by the segment’s importance feedbacks. It can easily combine different inputs, such as raw features, high-level features, audio features and vision feature together under this framework. The holistic picture of our work is shown in figure 1.

Currently, as video summarization problem appeals to a lot of researchers, many state-of-the-art approaches have been presented in solving this problem [3, 24–26]. Generally, they treat video summarization as sequence-to-sequence learning problem. Since RNN and its variants LSTM, GRU are very efficient in modelling long time dependencies under encoder-decoder architectures, many works in machine translation, image/video captioning, reading comprehension adopt these technologies. In this paper, GRU modelling is adopted in order to consider both the isolated segment, and its roles in the whole video.

Video summarization, as one of the comprehensive video understanding tasks, is extremely difficult and requires a large amount of data in deep learning architecture. However, the collection of such summarization labels is time-consuming and labor-intensive, resulting in an insufficient dataset. Since supervised learning can easily overfit on small data, many methods have been explored on alleviating this issue. Many state-of-the-art videos summarization work address this problem by unsupervised learning, semi-supervised learning, or multi-task learning [25, 26, 28]. We explore the self-supervised learning on video sequence modelling and joint training with segments importance score prediction. Another way to confront overfitting is to shuffle information flows which potentially improves the model’s generalization ability. The contribution of our work:

[1] **We unify various inputs on different semantic levels to one framework by formatting video summarisation into a recommender problem.**

[2] **We develop an algorithm that models independent segments and segment sequence(whole video).**

[3] **We extend the summarisation framework with self-supervised learning and data augmentation to deal with lack of labelled data.**

2. Related Work

Video classification, as a fundamental task for video understanding, has been studied seriously. Many high-quality datasets [1, 10, 11, 16] are published which drives the research to a higher level. Supervised learning brings the performance of algorithms nearly to human performance on large-scale video classification tasks [4, 7, 9, 18–21].

A significant number of deep learning based frameworks have been explored recently in solving video summarization [3, 22, 24, 26, 28]. K. Zhang et al. creatively applied LSTM in supervised video sequence labelling to model video temporal information with good performance [24]. Jiri Fajtl et al. introduced self-attention instead of RNN to enhance computation efficiency [3]. K. Zhou et al. showed that fully unsupervised learning can outperform many supervised methods by considering diversity and representativeness in reinforcement learning-based framework [28]. Y. Zhang et al. introduced adversarial loss to video summarization which learns a dilated temporal relational generator and a discriminator with three-player loss [26].

Researchers believe learning good visual representation can help deep neural network improve both fitting and generalization ability, especially in classification task [5, 12, 14, 17]. Some video summarization work adopts unsupervised learning as an auxiliary task to help to improve the supervised learning system’s performance. K. Zhang et al. used “retrospective encoder” that embeds the predicted summary and original video into the same abstract semantic space with closer distance. The semi-supervised settings help increase F-score on TVSum dataset from 63.9% to 65.2% [25]. K. Zhou et al. on the other hand, extending their reinforcement learning based framework to semi-supervised style elevated performance on both SumMe and TVSum comparing to purely supervised or unsupervised methods [28].

For the subjectivity of video summarization task, Kanehira et al. provided a solution by building a summary that depends on the particular aspect of a video the viewer focuses on [8]. Joonseok Lee et al. from Google published a content-only video recommendation system which we regarded as a work close to video summarization [13].

3. Approach

In this section, we detail our method on CoView 2019 comprehensive video understanding challenge track. First, we simply introduce our work on action and scene recognition in untrimmed video. Second, we formalize the summarization problem and present our solution based on deep neural network. Third, we talk about some important technology on how to prevent the DNN from early-stage overfitting.

3.1. Action and Scene Recognition in Untrimmed Video

We first adopt I3D with non-local blocks video classification [2, 21] in this subtask. A ResNet101 backbone is used for this framework. We tried two ways of training:

1. *Action and scene recognition task sharing the same backbone with two SoftMax loss branches for each classification tasks.*

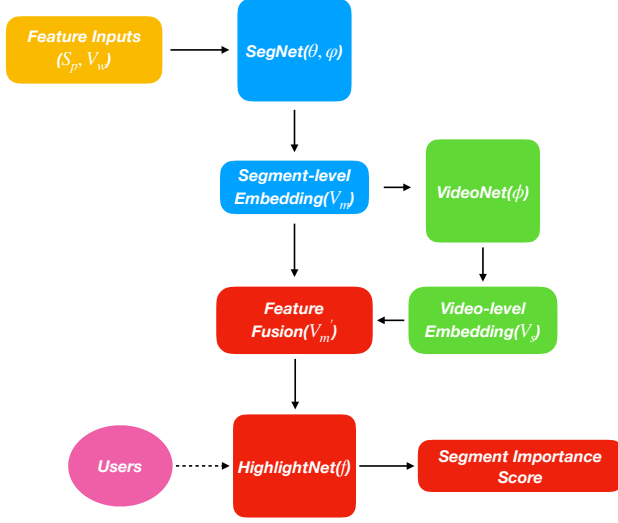


Figure 2. The Video Summarization Network structure. It contains three subnetworks including: SegNet, VideoNet, and HighlightNet.

2. We get one model each for action and scene recognition.

Since the training data just comes from only 1000 videos, the scene data for training is too small and lacks variation for recognition. There are many public image scene classification datasets such as Place360 [27] and SUN [23] for image scene recognition. We assume the video scene classification task can be trained by image classification without much information loss. We can easily make full use of public datasets to potentially improve models' robustness. We use average pooling operation on image classification predictions for final video level results. To capture the advantages of both video-based and image-based models, we ensemble the output scores of two models together with logistic regression.

3.2. Video Summarization Framework

We take the video summarization task as a content-based recommender problem. Let

$$V_m = [s_1, s_2, s_3, \dots, s_{T_N}]^T \in \mathbb{R}^{T_N \times sd},$$

be the feature matrix for a video m with T_N segments in sd dimension.

For each segment we have a mean importance score a_i as users' feedback, can be written by

$$A_m = [a_1, a_2, a_3, \dots, a_{T_N}]^T \in \mathbb{R}^{T_N \times 1}.$$

Our goal is to select segments from the video by Top-k highest importance prediction score as summarization set $\{S\}_k$. So the problem can be defined as finding a ranking function

$f(s)$ that can predict the segment importance score in video segments sequence

$$\hat{A}_m = f(V_m).$$

A loss function can be defined by mean-square error(MSE),

$$L(A_m, \hat{A}_m) = \frac{1}{H \times T_N} \sum_{m=1, i=1}^{H, T_N} ((A_m^i - f(V_m^i))^2),$$

where H is the number of videos in training set, our algorithm is to find optimal function f that minimizes the overall loss:

$$\hat{f} = \underset{R(f)}{\operatorname{argmin}} L(f),$$

in the prediction model space R .

Since s_i lacks information from the whole video sequence as the input of f , we set V_m' defined by

$$V_m' = V_m \oplus V_s,$$

\oplus is the operator that concatenate each row of V_m to V_s , V_s is a $T_N \times vd$ matrix that represent vd dimension whole video feature to each segment, which can also be taken as a learnable function of $V_s = \phi(V_m)$. In this work, we set the same value for each row of V_s .

We can also get sequence descriptors for one segment, like image frame, audio frame, a segment is defined by

$$S_p = [f_1, f_2, f_3, \dots, f_{T_G}]^T \in \mathbb{R}^{T_G \times fd}.$$

Another learnable function $V_r = \theta(S_p)$ is used to map frames sequence features from $T_G \times fd$ to $1 \times sfd$ space. We combine image based feature V_r and segment based feature V_w by learnable function φ to final segment-level feature

$$V_m = \varphi(V_r, V_w).$$

We can minimize f, ϕ, θ, φ in one framework by minimizing

$$\begin{aligned} L(f) &= L(f(V_m')) \\ &= L(f(V_m \oplus V_s)) \\ &= L(f(V_m \oplus \phi(V_m))) \\ &= L(f(\varphi(\theta(S_p), V_w) \oplus \phi(\varphi(\theta(S_p), V_w)))). \end{aligned}$$

We design a video summarization network which consists of three subnetworks, SegNet for function θ and φ , VideoNet for ϕ , and HighlightNet for f as shown in figure 2.

As we utilize ImageNet features for each sampled video frame, SegNet combines all frame features' sequence to a frame-based vector. SegNet also fuses frame-based feature and segment-level feature together. The SegNet is designed to process either 2D or 3D frame sequence. For 2D features,

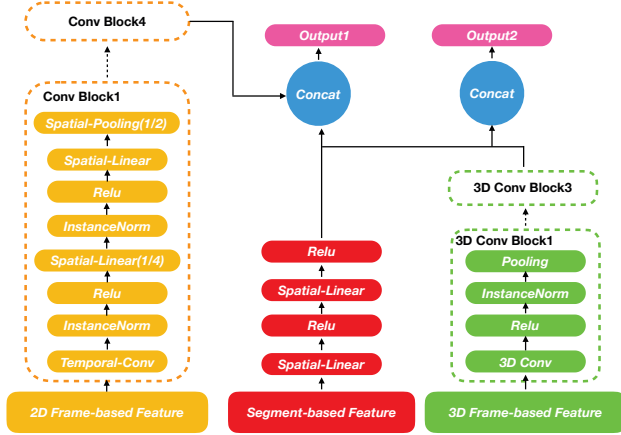


Figure 3. The SegNet structure details. The network is designed to process either 2D or 3D frame sequence. Only one output(output1 or output2) is taken as the segment embedding.

we adopt temporal-convolution and pooling to squeeze temporal dimension. And for 3D features, we use three 3D convolution blocks. The 2D convolution blocks adopt “bottleneck” structure for spatial fully connected layers (figure 3).

For each long video, the segments’ feature sequence generated by SegNet are taken as inputs to VideoNet for video-level modelling. VideoNet use Bi-GRU for sequence encoding. The final GRU unit hidden layer, which is a fixed-length context vector, is taken as the output of VideoNet representing the video-level feature.

Each segment-level feature concatenates with the video-level feature of the same long video as the final segment representation. The representation contains not only the independent video segment information, but also the whole video sequence information, which is passed to highlight subnetwork to predict the segment importance score. Figure 4 shows the structure of highlight subnetwork.

3.3. Multi-task Learning and Data Augmentation for Video Summarization Network

We consider an auxiliary self-supervised task to better modelling video sequence. Some segments are selected from one video by the fixed proportion before inputting to the network. We shuffle the selected segments and train the network to distinguish the odd-position segments as shown in figure 5. This operation assumes that a good video sequence encoder has the ability to model the right segments order. We control the difficulty of the task by only indicate the odd-position segments but not sort the shuffled sequence to right order. Parameters α and β are used for adjusting the shuffled segments’ proportion and the weight of self-supervised task in multi-task learning. The final loss can be

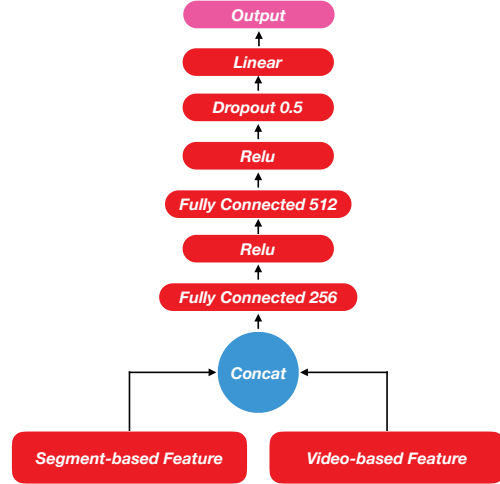


Figure 4. The HighlightNet structure details. The network is designed to predict the segment importance score by fusion features.

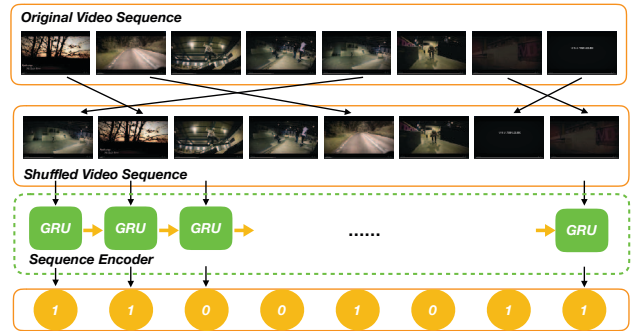


Figure 5. Self-taught learning for odd-position labelling. The input segment sequence are shuffled randomly by α ratio. The video encoder learn to find the odd-position segments in the sequence.

defined by:

$$L(f) = L(f(V_m')) + \beta * \delta(\text{shuffle}(\alpha, V_m)).$$

The advantages of applying multi-task learning are:

1. Learning several tasks simultaneously can suppress early-stage overfitting, by sharing the same representations.
2. The auxiliary task is helpful in providing more useful information.
3. Our method implicitly performs data augmentation since we shuffle the input video sequence, which may improve the robustness of the algorithm.
4. This method utilize unlabelled data for video sequence modelling.

A data augmentation method is proposed which only choose a portion of information from each segment when modelling the whole video each time. After modelling all portion of segments, we average the embeddings from different portion to one vector (figure 6).

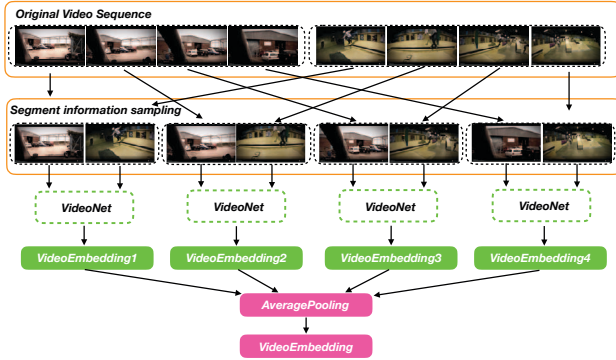


Figure 6. Data augmentation method. VideoNet only models a portion of each segment to video embedding. Multiple embeddings are combined by average pooling.

Method	Scene Top-5	Action Top-5
a (JT)	82.76	81.72
b (IT)	86.92	87.77
c (AED)	-	84.86
d (SED)	82.06	-

Table 1. Top-5 accuracy of different models on validation.

4. Experiments

In this section, we provide our comparison experiment on CoView 2019 dataset. The dataset consists of 1200 videos for training and 300 for testing. Videos are sampled from Youtube-8M, Dense-Captioning, and Video summarization dataset. Every video from the dataset is segmented into a set of 5-second-long segments and asked 20 users to annotate importance score and 99 action / 79 scene labels. The average of importance scores and the most voted action/scene labels are provided as ground-truth in segment-level. Before any experiments, we randomly split videos into training(1000) and validation (200) set based on original videos instead of five-seconds videos.

4.1. Scene and Action Classification in Untrimmed Video

For scene/action classification task, top-5 accuracy on validation results show in table 1. a. joint training I3D+non-local attention with two SoftMax branches(JT). b. independent I3D+non-local model(IT). c. independent I3D+non-local model with external action classification data(AED). d. ResNet-50 with external scene image classification data(SED).

We train I3D+non-local with the pre-trained model on Kinetics, and the ResNet-50 pre-trained on ImageNet. The results indicate: 1. Although scene/action classification task may have inner connection, both joint training and ensemble failed to improve the results. 2. Pure 2D convolution-

Feature Length	SummaryScore
Dim = 512	81.47
Dim = 256	81.81
Dim = 128	81.49
Dim = 64	81.19

Table 2. Summary score on different embedding dimension.

based network method is less accurate than 3D convolution-based network on both scene and action recognition task. 3. External data didn't help improve classification accuracy on validation set, but we still use it for potential model generalization enhancement.

4.2. Video Summarization

We investigate our summarization network on different parameter scales, visual and visual-audio feature inputs, single-task and multi-task learning, with and without data augmentation.

Evaluation Protocol. The video summarization evaluation metric computes the sum of importance scores of selected segments comparing with ground truth summary segments. The importance score metric is defined as

$$\frac{1}{N} \sum_{k=1}^N \frac{\sum_{i=1}^{N_s} ImportanceScore(k, Sub_i)}{\sum_{i=1}^{N_s} ImportanceScore(k, GT_i)},$$

where N is the number of test videos, N_s is the number of summarized segments. $ImportanceScore(k, GT_i)$ is the importance score from the i -th segment of the ground truth summary of the k -th video, and $ImportanceScore(k, Sub_i)$ is the importance score of the i -th segment of the submitted summary for the k -th video. Importance Score is shared in both ground truth and submitted summary. N_s is set to 6 for all videos, the top N_s importance segments are ground truth summary. We randomly choose N_s segments for 10 times as the baseline of summary score, the mean value is 74.92%.

Embedding Dimension. We explore different embedding dimension for the network by regularizing the same length of segNet and videoNet outputs. For each segment, we sample 16 frames for image-based extractor and 32 frames for video-based extractor. We use GoogleNet trained on ImageNet to extract image-based feature and I3D+non-local blocks trained on Kinetics to extract video-based feature. Table 2 shows our results for different feature length from 512 to 64. The 256-feature length is slightly better.

Semantic Features Combination. Image-based feature is extracted by image classifier. We select three types of image classifier: 1. ResNet-50 trained on CoView2019 scene classification task(R50_s). 2. ResNet-152 trained on ImageNet(R152_i). 3. GoogleNet trained on ImageNet(G_i). Video-based feature is extracted by video clas-

Feature Type	SummaryScore
G_i+INK	81.81
G_i+INA	81.06
R50_s+INA	80.83
R152_i+INA	80.99
R50_s+INK	81.45
R152_i+INK	81.45

Table 3. Summary score with different visual semantic feature combination.

Method	SummaryScore
a. (sup_no_prt)	80.69
b. (sup_with_prt)	80.98
c. (multi_with_prt)	81.64

Table 4. Summary score with Multi-Task learning.

sifier. We choose two video classifier: 1. I3D+non-local model trained on Kinetics(INK). 2. I3D+non-local model trained on CoView2019 action recognition task(INA). Table 3 shows different visual feature combination on feature length 256. Our model didn’t benefit a lot from action and scene recognition task.

Multi-Task Learning. Since the self-taught learning increases the learning difficulty, we adopt 3-layers bi-GRU on video net and feature length equals to 512 with R50_s+INK features. Dropout and bottleneck structure are removed from segNet. The self-taught learning is firstly trained for two days using videos on CoView2019 with $\alpha = 15\%$ shuffle ratio. We use the trained model, with 96.64% accuracy on self-taught task and 78.6% recall on odd order segments, as pre-trained model in multi-task learning setting. In joint training, we set $\alpha = 0.02$, $\beta = 1$, and $\alpha = 0$ when evaluate on validation. Table 4 shows the comparison results, we evaluate three models: a. the baseline supervised learning without pre-trained model(sup_no_prt). b. supervised learning with self-taught pre-trained model(sup_with_prt). c. multi-task learning with self-taught pre-trained model(multi_with_prt). The results shows that multi-task setting is better than the baseline model with small margin.

Data augmentation. The basic network setting is the same as R50_s+INK in ”Semantic Features Combination” experiment, we get 0.29% improvement from 81.45% to 81.74%.

Audio feature. We extract audio features MFCC, Chroma-gram for each segment by using python audio processing package LibROSA. We concatenate the audio features to 5450-dimension segment-level features before inputting to SegNet. The audio fusion model builds bases on Multi-Task setting in ”Multi-Task Learning” experiment. The result decrease from 81.64% to 80.42%, that may due

to increased model complexity caused by larger feature dimension without important information.

Optical flow. Optical flow as a low-level feature describing video motion, is shown as complementary information to RGB in action recognition task. Optical flow feature sequence are obtained by using Gunnar Farneback’s algorithm to extract 16 frames for each segment. We modified SegNet in Multi-Task setting to 4 layers 3D convolution for processing optical flow features on temporal and spectral. Same as audio feature, this modification decrease summary score from 81.64% to 80.78%.

Model ensemble. We ensemble five models: base 256 feature length model(81.81%), audio feature model(80.42%), optical flow model(80.78%), multi-task model(81.64%), data augmentation model(81.74%) by linear regression. Then we do model selection according to linear regression weights which direct us choosing base 256-length model and data augmentation model ensemble result(81.86%) on validation as the final submission.

5. Conclusion and Future Work

In this paper, we propose a scalable deep neural network for video summarization in content-based recommender formulation, which predicts the segments importance score by considering both segment and video level information. Our work shows that data augmentation and Multi-Task learning is helpful in solving the limitation of dataset problem. To better understand the video content, we also perform action and scene recognition in untrimmed video with state-of-the-art video classification algorithm. We experiment with combinations of high-level visual semantic features, audio features and optical flow, and concluded that visual semantic features play the most important role in this summarization task.

There are some areas we can further explore to improve comprehensive video understanding in the future. First of all, we didn’t benefit from action and scene connection in both recognition and summarization task, which leaves us room for better utilize this prior knowledge. Second, as we re-formalize video summarization in recommender framework, some state-of-the-art recommendation technologies can be introduced such as Collaborative Filtering, Factorization Machine, Wide & Deep Learning and so on. Last but not the least, although CoView2019, to our knowledge, provides the largest public video summarization dataset, it is too small in the scale of deep learning. We need more, like users’ action when browsing video websites, video language description, and large-scale dataset to accomplish such complex task.

References

- [1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.
- [2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- [3] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention. In *ACCV Workshops*, 2018.
- [4] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *arXiv preprint arXiv:1812.03982*, 2018.
- [5] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3636–3645, 2017.
- [6] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer, 2014.
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [8] Atsushi Kanehira, Luc Van Gool, Yoshitaka Ushiku, and Tatsuya Harada. aware video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7435–7444, 2018.
- [9] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.
- [10] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.
- [11] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.
- [12] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017.
- [13] Joonseok Lee and Sami Abu-El-Haija. Large-scale content-only video recommendation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 987–995, 2017.
- [14] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016.
- [15] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187, 2015.
- [16] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- [17] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015.
- [18] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [19] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.
- [20] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.
- [21] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.
- [22] Huawei Wei, Bingbing Ni, Yichao Yan, Huanyu Yu, Xiaokang Yang, and Chen Yao. Video summarization via semantic attended networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [23] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010.
- [24] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *European conference on computer vision*, pages 766–782. Springer, 2016.
- [25] Ke Zhang, Kristen Grauman, and Fei Sha. Retrospective encoders for video summarization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 383–399, 2018.
- [26] Yujia Zhang, Michael Kampffmeyer, Xiaoguang Zhao, and Min Tan. Dtr-gan: Dilated temporal relational adversarial network for video summarization. In *Proceedings of the ACM Turing Celebration Conference-China*, page 89. ACM, 2019.
- [27] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database

for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

- [28] Kaiyang Zhou, Yu Qiao, and Tao Xiang. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.