

Activity Image-to-Video Retrieval by Disentangling Appearance and Motion

Liu Liu, Jiangtong Li, Li Niu, Ruicong Xu, Liqing Zhang

MoE Key Lab of Artificial Intelligence,
Department of Computer Science and Engineering,
Shanghai Jiao Tong University



Introduction



- Image-to-Video Retrieval:
 - retrieve relevant videos based on a query image
- Task Classification:
 - Instance-based Image-to-Video Retrieval (IIVR)
 - Activity-based Image-to-Video Retrieval (AIVR)



Fig 1.Examples for IIVR task

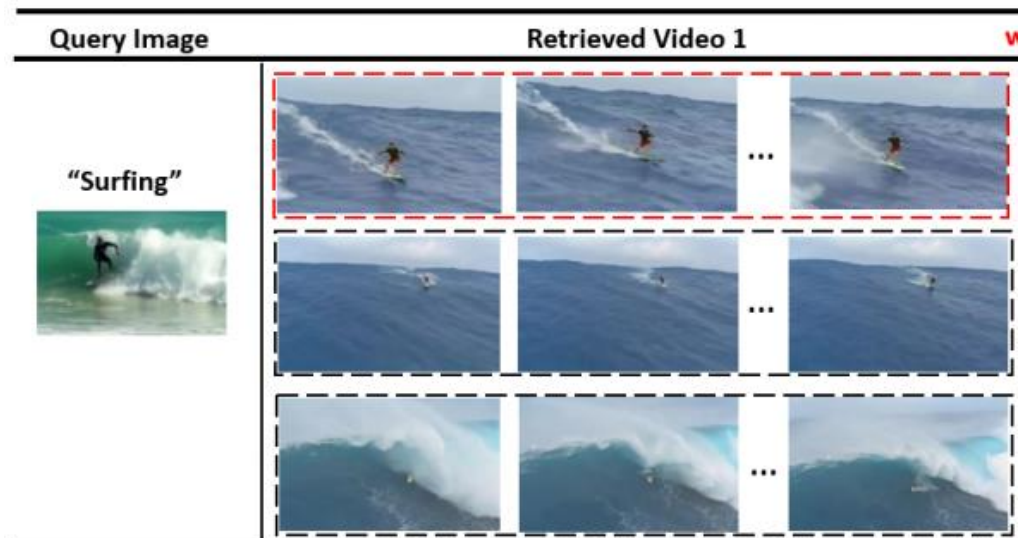


Fig 2.Examples for AIVR task

Introduction – IIVR & AIVR



- Existing IIVR research:
 - Early methods applied image retrieval methods for image-to-video retrieval,
 - Yu extracted object proposals from each frame and measured the similarity between the query object and the whole video through hamming distance,
 - Zhu and Wang introduced a large vocabulary quantization based Bag-of-Words to index videos.

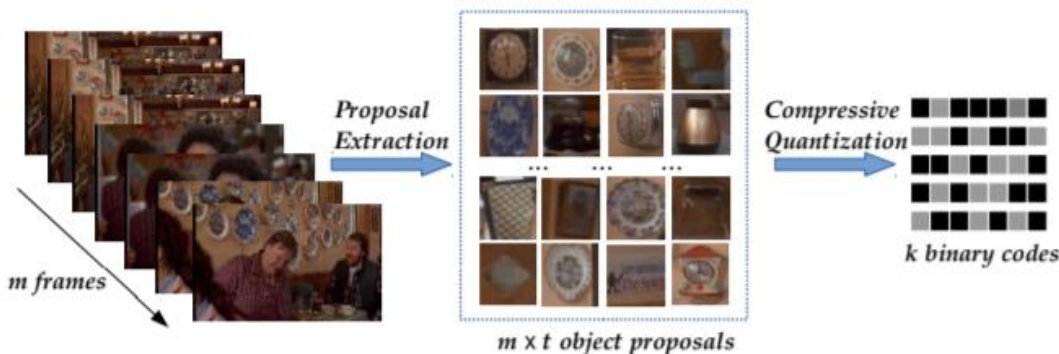


Fig 1. Overview of Yu's method

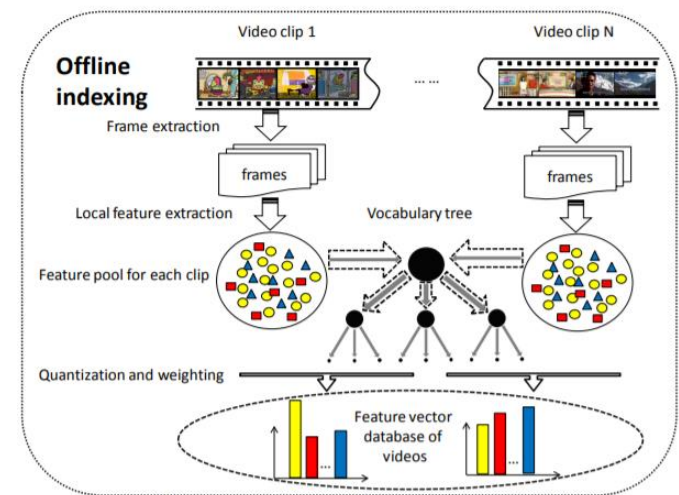


Fig 2. Overview of Zhu's method

Introduction – IIVR & AIVR



- Limitations of IIVR:
 - focus more on object detection and representation while ignoring the dynamic motion tendency of different objects in videos and images
- Existing AIVR research:
 - APIVR projected the image features and activity proposal-based video features into a joint space and employed Graph Multi-Instance Learning module to filter out the noisy proposals.

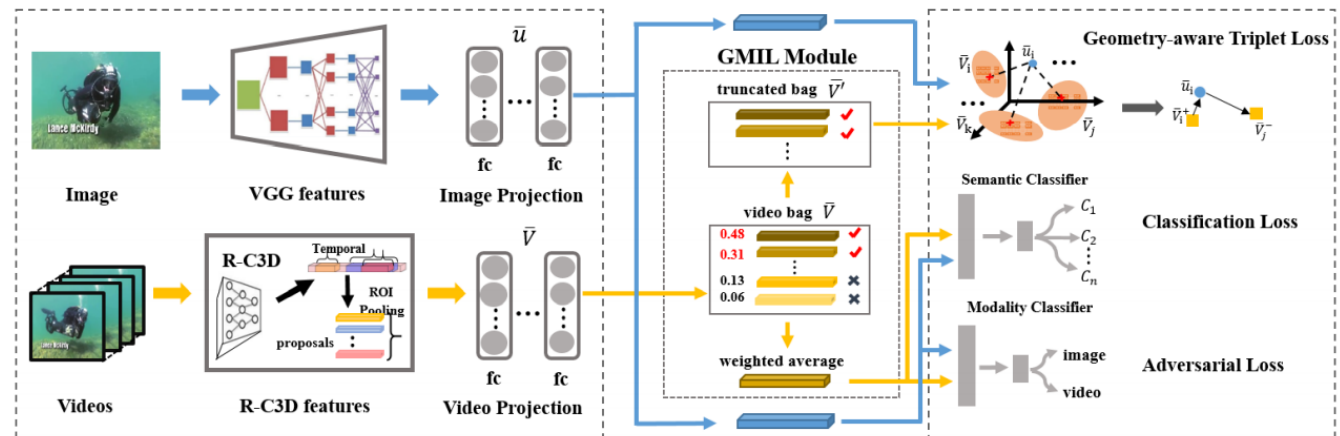
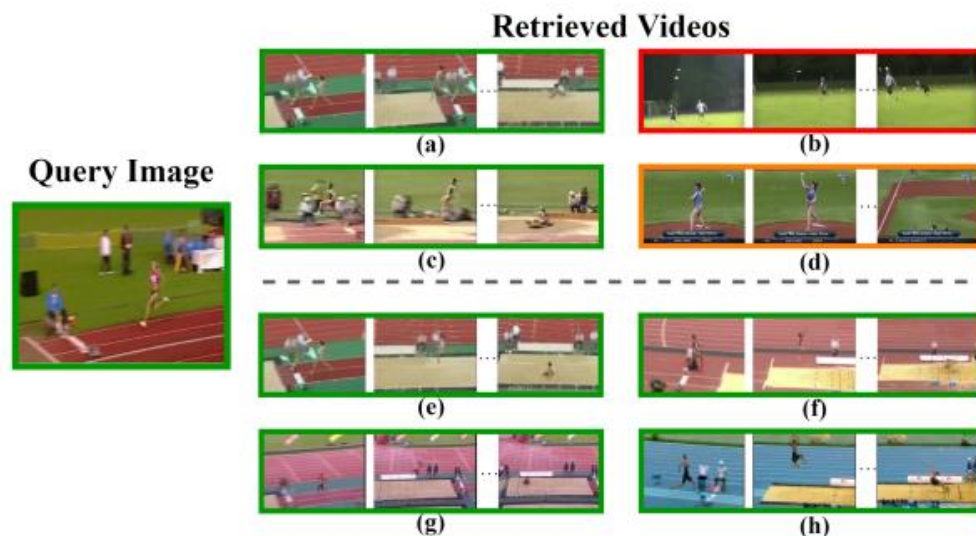


Fig 1. Overview of APIVR

Introduction - Motivation



- Defects of APIVR: ignore the asymmetric relationship between images and videos.
- Image: appearance information; Video: appearance + motion information
 - Appearance information: the shape, pose, texture, and color of objects, ...
 - Motion information: the trajectory of key points and variation of objects, ...



Method - Overview

- Problem Definition:
 - Given an image and a video, return the similarity between them
- Two Modules: Feature Disentanglement and Video Feature Reconstruction

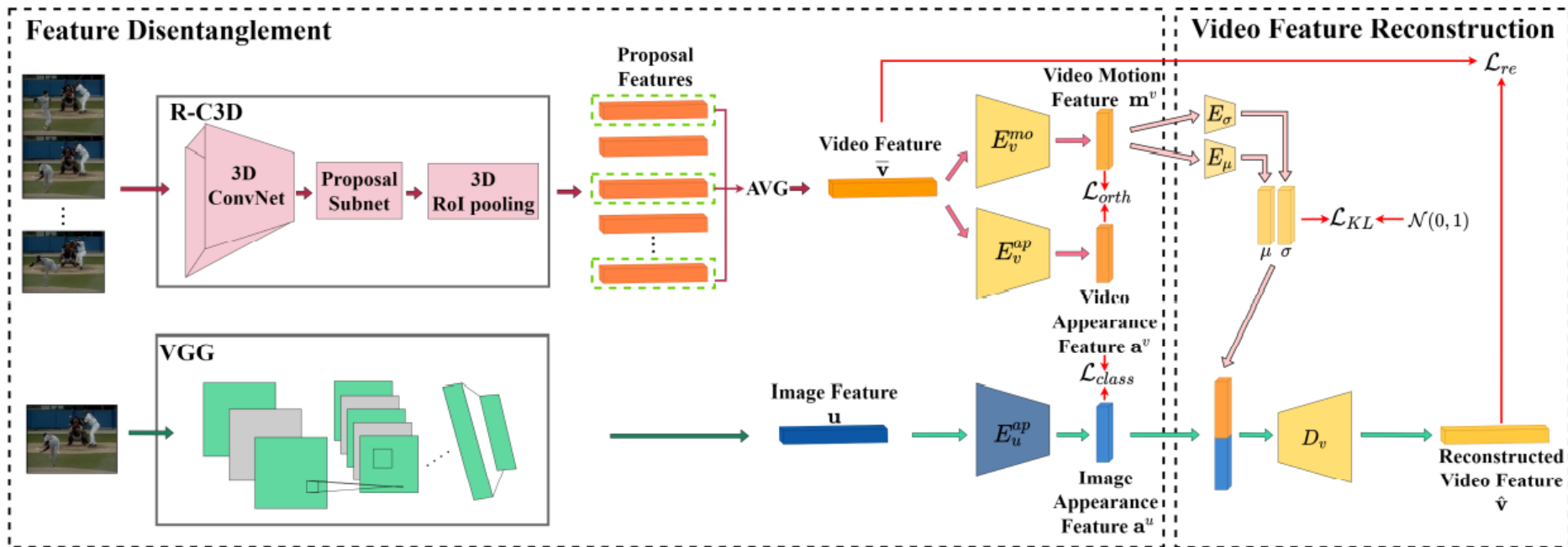
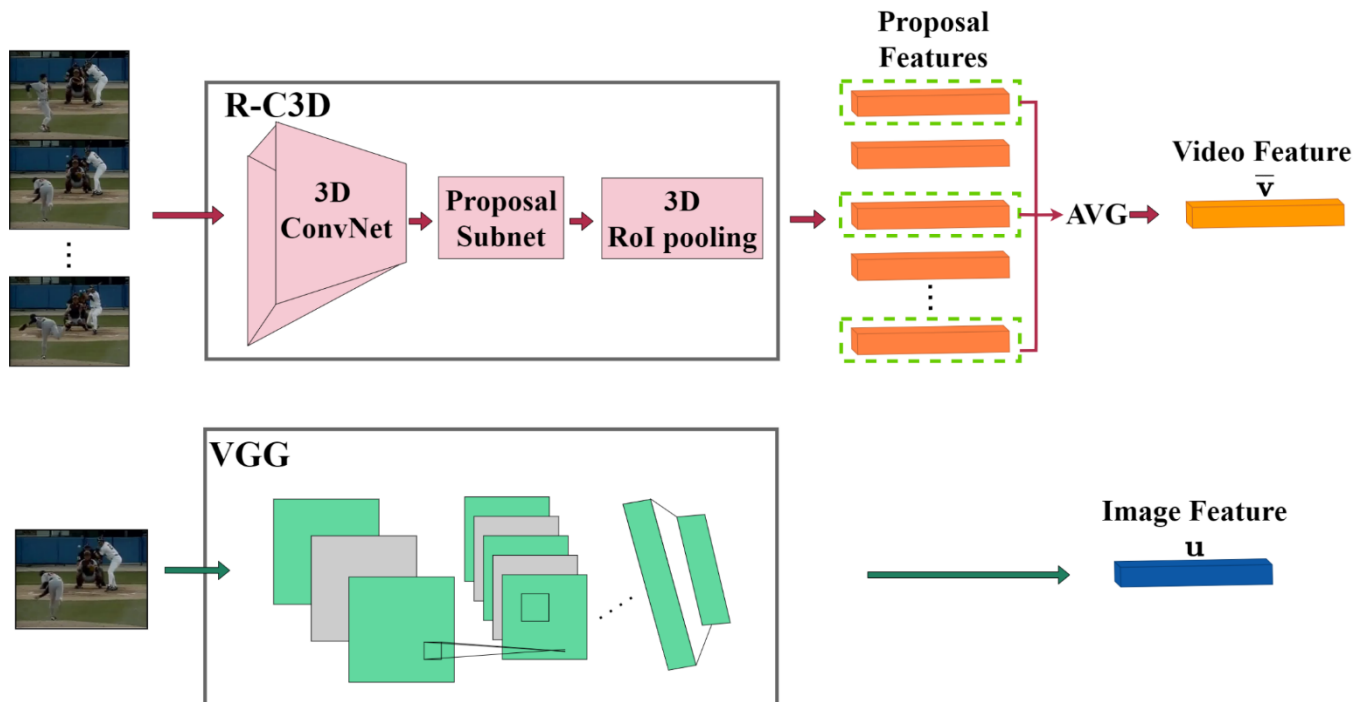


Fig 1. The flowchart of our method

Method - Feature Disentanglement



- Feature Extraction:
 - Video Clip: 1) apply a R-C3D model pretrained on the ActivityNet dataset; 2) Choose top k proposals with largest confident scores 3) Average;
 - Image: apply the VGG-16 pretrained on ImageNet;



Method - Feature Disentanglement

- Asymmetric Disentanglement:

- Video disentanglement:

$$\mathbf{m}^v = E_v^{mo}(\bar{\mathbf{v}}), \mathbf{a}^v = E_v^{ap}(\bar{\mathbf{v}})$$

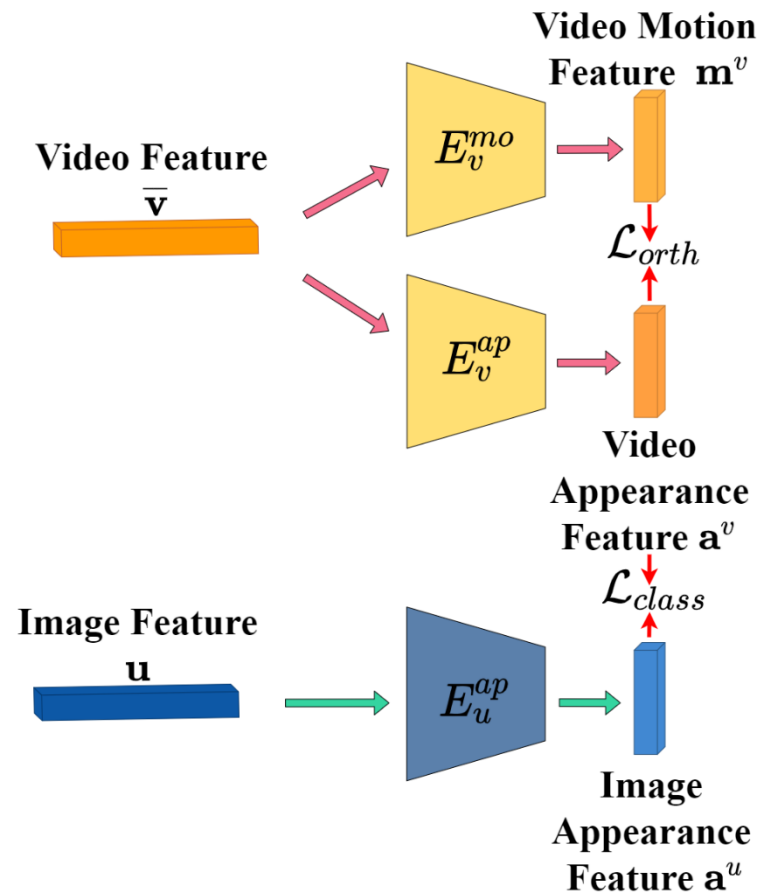
- Image: $\mathbf{u}^v = E_v^{ap}(\mathbf{u})$

- Orthogonal Loss:

$$\mathcal{L}_{orth} = \cos(\mathbf{m}^v, \mathbf{a}^v)$$

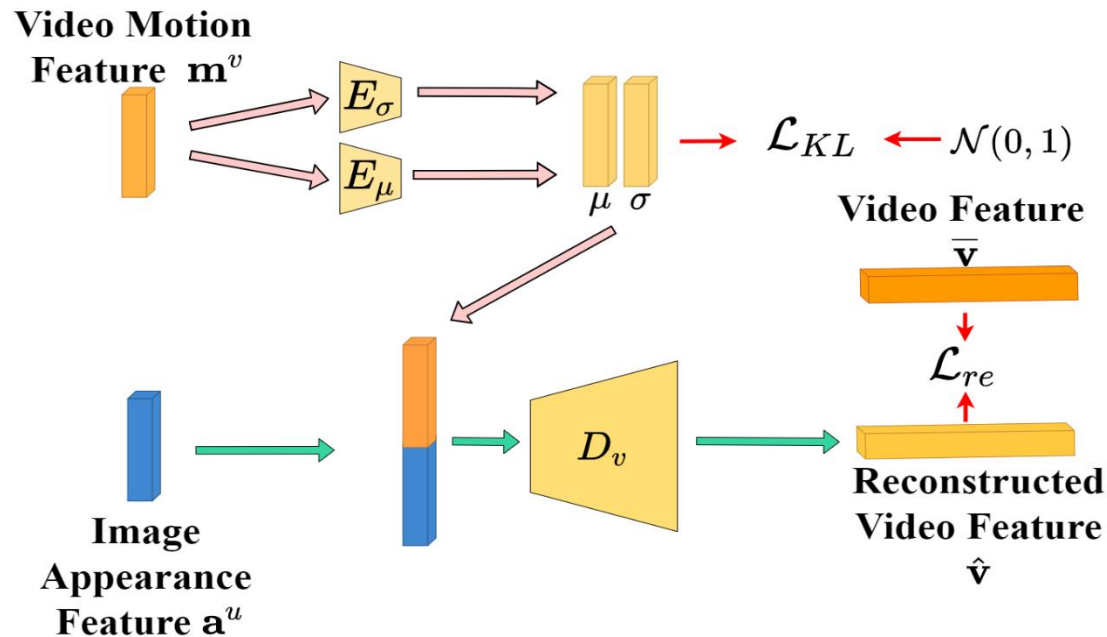
- Classification loss:

$$\mathcal{L}_{class} = -\log(p(\mathbf{a}^v)_y) - \log(p(\mathbf{a}^u)_y)$$



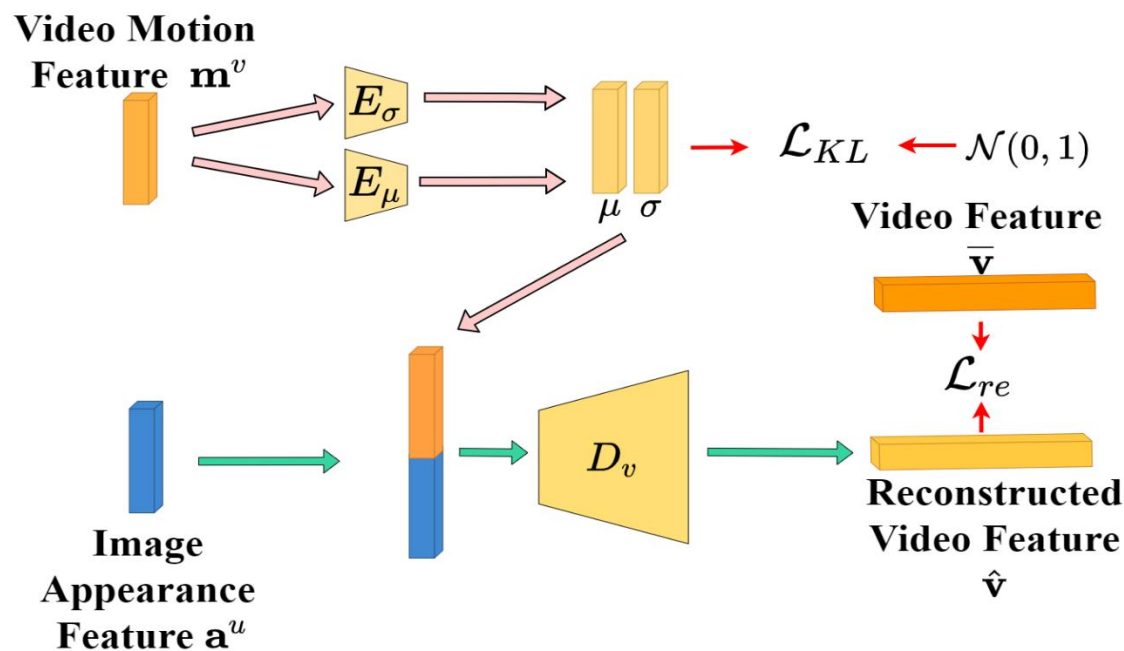
Method - Video Feature Reconstruction

- Since image-to-video translation is a multi-modal problem, inspired by VAE, we encode motion feature into motion uncertainty code \mathbf{z} ($p_{\theta}(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{1})$).
- Apply two encoders: $\mu = E_{\mu}(\mathbf{m}^v)$, $\sigma = E_{\sigma}(\mathbf{m}^v)$
- Kullback–Leibler divergence Loss: $\mathcal{L}_{KL} = \text{KL}(q_{\phi}(\mathbf{z} | \mathbf{m}^v) || p_{\theta}(\mathbf{z}))$



Method - Video Feature Reconstruction

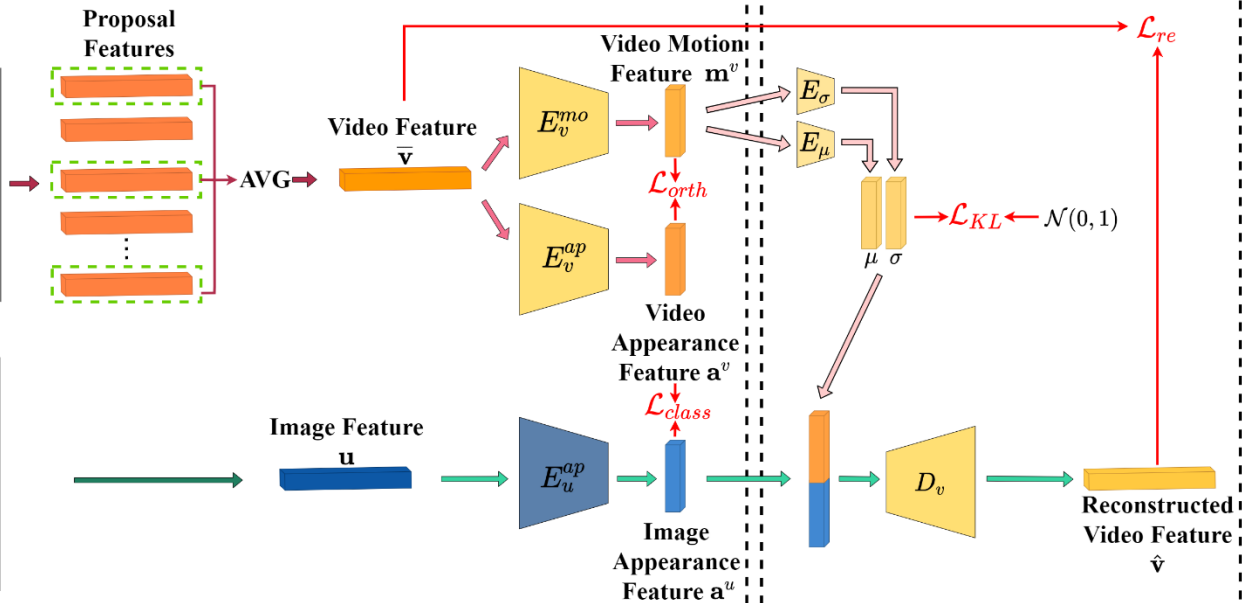
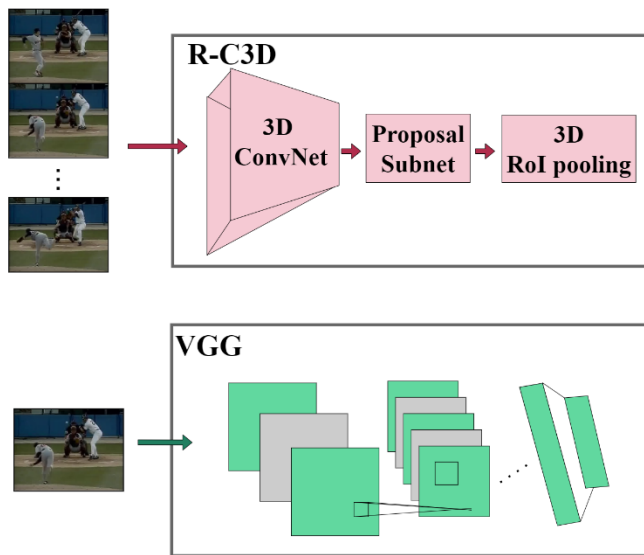
- Generate motion uncertainty code: $\mathbf{z} = \mu + \epsilon\sigma$
- Video Feature Reconstruction: $\hat{\mathbf{v}} = D_v([\mathbf{a}^u, \mathbf{z}])$
- Reconstruction Loss: $\mathcal{L}_{re} = \|\bar{\mathbf{v}} - \hat{\mathbf{v}}\|_2^2$



Method – Final Loss

- Final training loss: $\mathcal{L}_{total} = \lambda_o \mathcal{L}_{orth} + \mathcal{L}_{class} + \mathcal{L}_{KL} + \mathcal{L}_{re}$

Feature Disentanglement



Method – Retrieval



- Appearance Feature Space:

$$S_A = 1 - \cos(\mathbf{a}^u, \mathbf{a}^v)$$

- Video Feature Space:

- Sample motion uncertainty code from $\mathcal{N}(\mathbf{0}, \mathbf{1})$ for h times;
- Obtain h translated video features $\{\hat{\mathbf{v}}_i \mid i = 1 \dots h\}$
- Calculate similarity: $S_V = \min_{i=1}^h (1 - \cos(\bar{\mathbf{v}}, \hat{\mathbf{v}}_i))$

- Combination:

$$S_{all} = (1 - \lambda_v)S_A + \lambda_v S_V$$

- λ_v is a hyper-parameter to balance two feature spaces

Experiment – Dataset



- Follow the APIVR method to construct the dataset:
 - Divide long videos and select video clips
 - Randomly sample a frame as its paired image for each video clip
- ActivityNet:
 - 4727 validation videos from 200 activity categories
 - Final obtain: 4739 image-video pairs = 3790 training pairs + 949 test pairs
- THUMOS'14:
 - 200 validation videos and 213 test videos from 20 different sports activities
 - merge similar activity categories
 - Final obtain: 7028 image-video pairs = 5614 training pairs + 1414 test pairs

Experiment - Comparison with Other Methods

Method	ActivityNet				THUMOS'14			
	mAP@10	mAP@20	mAP@50	mAP@100	mAP@10	mAP@20	mAP@50	mAP@100
CMDN (Peng, Huang, and Qi 2016)	0.289	0.280	0.269	0.257	0.518	0.513	0.508	0.504
DSPE (Wang, Li, and Lazebnik 2016)	0.281	0.273	0.261	0.249	0.507	0.505	0.501	0.498
JFSSL (Wang et al. 2016)	0.277	0.268	0.256	0.244	0.476	0.473	0.469	0.465
ACMR (Wang et al. 2017)	0.294	0.288	0.273	0.259	0.526	0.522	0.514	0.505
CCL (Peng et al. 2018)	0.287	0.279	0.267	0.256	0.512	0.509	0.506	0.502
DSCMR (Zhen et al. 2019)	0.297	0.292	0.281	0.269	0.625	0.623	0.622	0.621
SDML (Hu et al. 2019)	0.304	0.301	0.289	0.279	0.648	0.647	0.646	0.645
BPBC (Xu et al. 2017)	0.295	0.287	0.275	0.258	0.514	0.511	0.507	0.503
APIVR (Xu et al. 2020)	0.308	0.298	0.283	0.269	0.655	0.653	0.651	0.649
MAP-IVR (Appearance)	0.304	0.297	0.284	0.273	0.643	0.641	0.637	0.635
MAP-IVR (Video)	0.323	0.313	0.296	0.282	0.691	0.689	0.682	0.677
MAP-IVR (Comb)	0.357	0.346	0.329	0.314	0.721	0.719	0.717	0.714

Table 1. Comparison with existing methods on ActivityNet and THUMOS'14.

Experiment - Ablation Study



	\mathcal{L}_{class}	\mathcal{L}_{KL}	\mathcal{L}_{re}	\mathcal{L}_{orth}	Comb	Ap	Vi
1	✓	✓	✓	✓	0.357	0.304	0.323
2	×	✓	✓	✓	0.296	—	0.296
3	✓	×	✓	✓	0.221	0.297	0.047
4	✓	✓	×	✓	0.299	0.299	—
5	✓	✓	✓	×	0.334	0.303	0.307
6	✓	×	×	×	0.285	0.285	—

Table 2. The ablation study of different loss terms. “Comb” represents the retrieval in the combination of two spaces; “Ap” and “Vi” represent the retrieval in appearance feature and video feature space, respectively. ✓ (*resp.*, ×) means adding (*resp.*, removing) this loss during training.

Experiment - Ablation Study



	\mathcal{L}_{class}	\mathcal{L}_{KL}	\mathcal{L}_{re}	\mathcal{L}_{orth}	Comb	Ap	Vi
1	✓	✓	✓	✓	0.357	0.304	0.323
2	×	✓	✓	✓	0.296	—	0.296
3	✓	×	✓	✓	0.221	0.297	0.047
4	✓	✓	×	✓	0.299	0.299	—
5	✓	✓	✓	×	0.334	0.303	0.307
6	✓	×	×	×	0.285	0.285	—

Table 2. The ablation study of different loss terms. “Comb” represents the retrieval in the combination of two spaces; “Ap” and “Vi” represent the retrieval in appearance feature and video feature space, respectively. ✓ (*resp.*, ×) means adding (*resp.*, removing) this loss during training.

Experiment - Ablation Study



	\mathcal{L}_{class}	\mathcal{L}_{KL}	\mathcal{L}_{re}	\mathcal{L}_{orth}	Comb	Ap	Vi
1	✓	✓	✓	✓	0.357	0.304	0.323
2	×	✓	✓	✓	0.296	—	0.296
3	✓	×	✓	✓	0.221	0.297	0.047
4	✓	✓	×	✓	0.299	0.299	—
5	✓	✓	✓	×	0.334	0.303	0.307
6	✓	×	×	×	0.285	0.285	—

Table 2. The ablation study of different loss terms. “Comb” represents the retrieval in the combination of two spaces; “Ap” and “Vi” represent the retrieval in appearance feature and video feature space, respectively. ✓ (*resp.*, ×) means adding (*resp.*, removing) this loss during training.

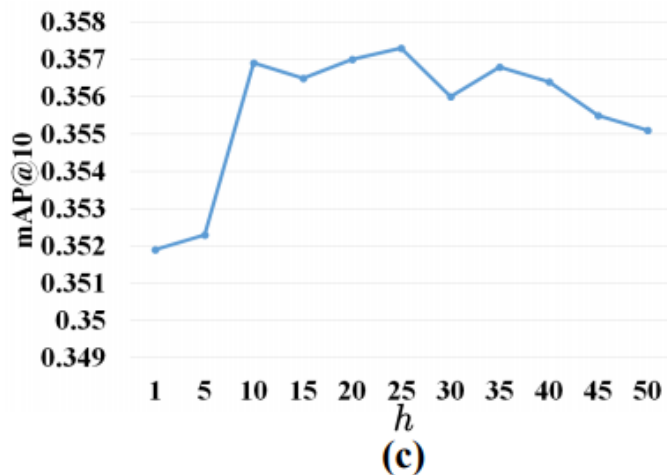
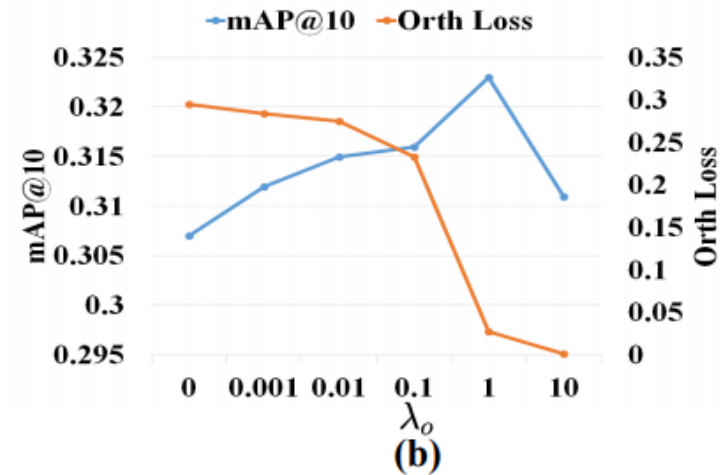
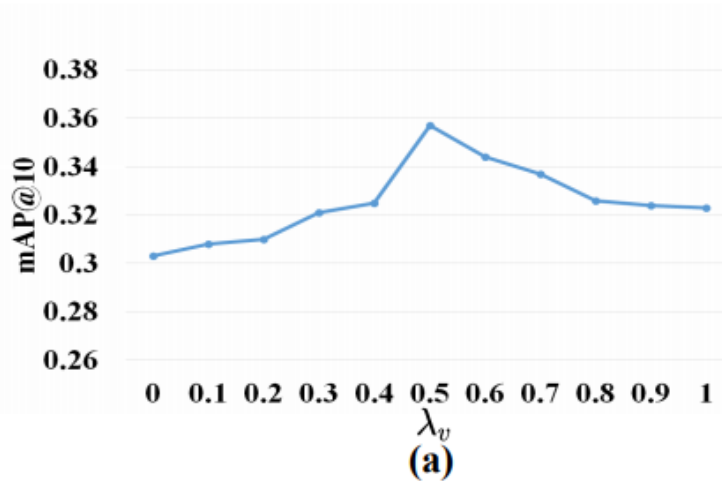
Experiment - Ablation Study



	\mathcal{L}_{class}	\mathcal{L}_{KL}	\mathcal{L}_{re}	\mathcal{L}_{orth}	Comb	Ap	Vi
1	✓	✓	✓	✓	0.357	0.304	0.323
2	×	✓	✓	✓	0.296	—	0.296
3	✓	×	✓	✓	0.221	0.297	0.047
4	✓	✓	×	✓	0.299	0.299	—
5	✓	✓	✓	×	0.334	0.303	0.307
6	✓	×	×	×	0.285	0.285	—

Table 2. The ablation study of different loss terms. “Comb” represents the retrieval in the combination of two spaces; “Ap” and “Vi” represent the retrieval in appearance feature and video feature space, respectively. ✓ (*resp.*, ×) means adding (*resp.*, removing) this loss during training.

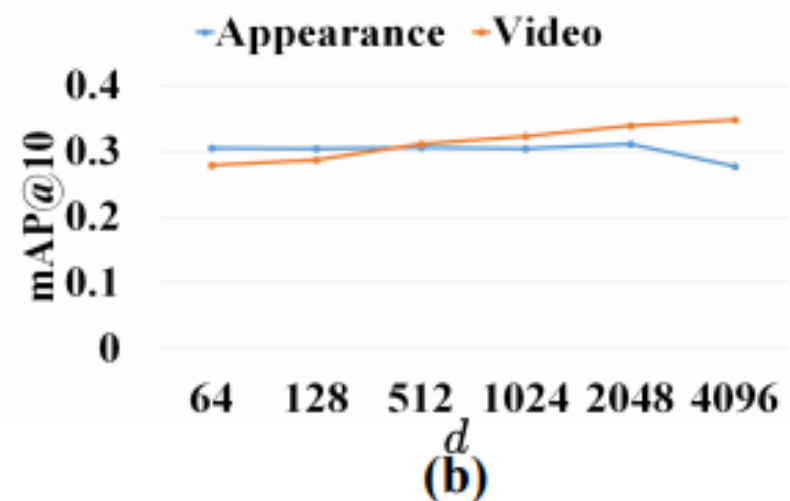
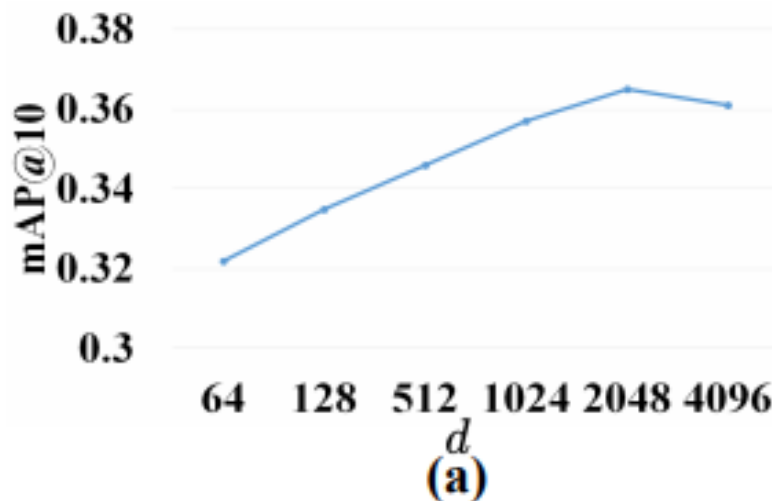
Experiment - Hyper-parameter Analysis



Analysis of different hyper-parameters

- (a) The retrieval results with various combination ratios of two spaces.
- (b) The retrieval results in video feature space and values of orthogonal loss with different λ_o .
- (c) The retrieval results based on the combination of two spaces with different h motion uncertain codes in the testing stage.



























Experiment - Effect of Appearance and Motion Feature Dimension



Analysis of appearance and motion feature dimension.

- (a) The retrieval results based on the combination of two spaces, where $\lambda_v = 0.5$.
- (b) The retrieval results in appearance feature space and video feature space, respectively.

Experiment - Visualization of Retrieved Videos

Retrieved Videos						
Query Image	Appearance Feature Space			Video Feature Space		
			...			
			...			
			...			
			...			
			...			

Thanks!

