Learn from Unlabeled Videos for Near-duplicate Video Retrieval

Xiangteng He[#1], Yulin Pan[#2], Mingqian Tang[2], Yiliang Lv[2] and Yuxin Peng[*1]

[1] Wangxuan Institute of Computer Technology, Peking University [2] Alibaba Group

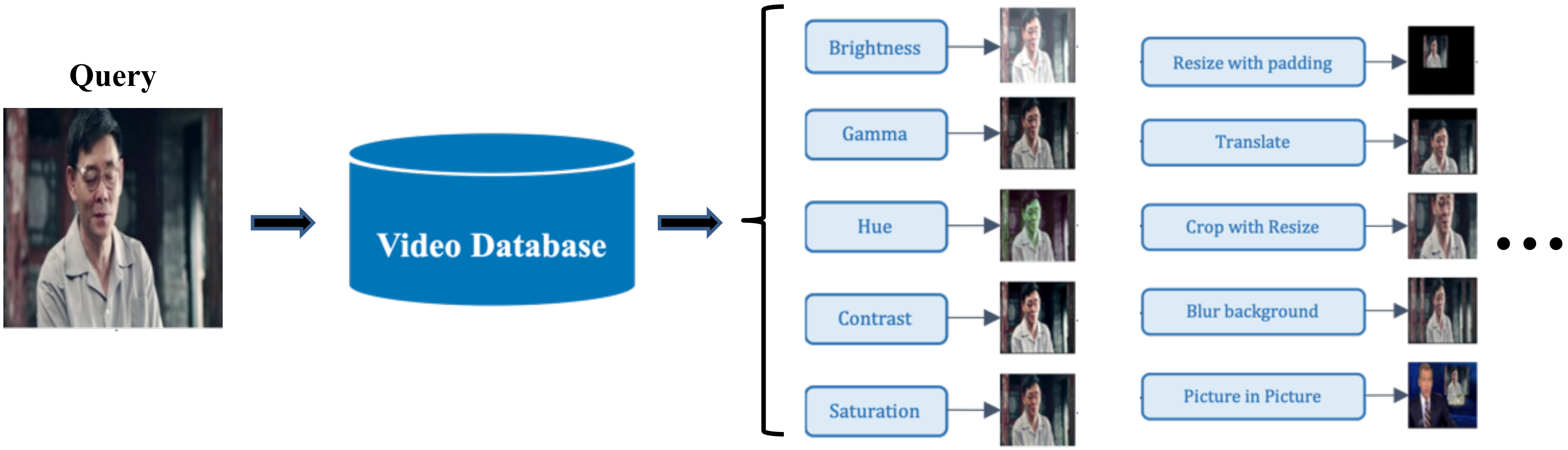
{hexiangteng, pengyuxin}@pku.edu.cn

Guide Line

- Introduction
- Our Approach
- Experiment
- Conclusion

Background

- **Near-Duplicate Video Retrieval (NDVR)**



- **Application Scenarios**

Video Tracing

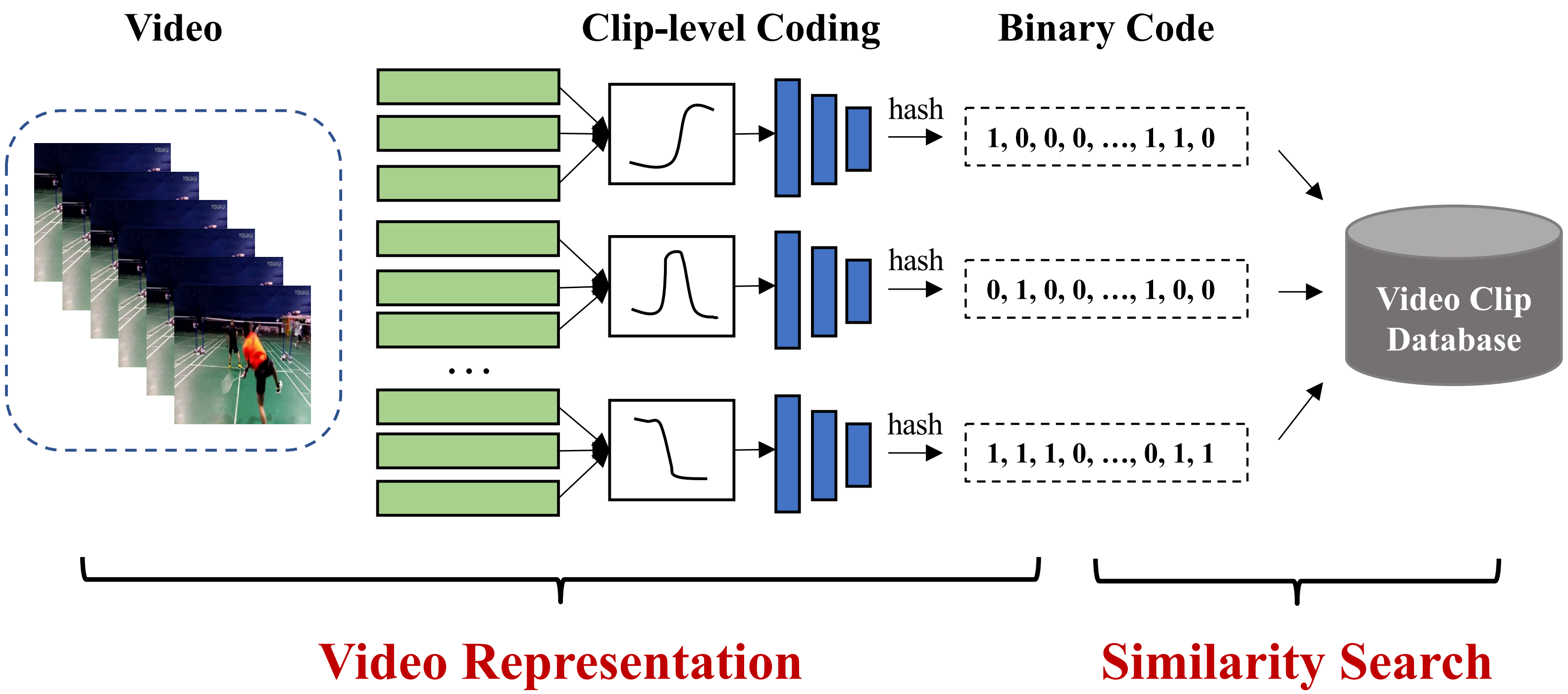
Copyright Protection

Material Search

Video Filtering

Background

- **To design an efficient and effective near-duplicate video retrieval system**
  - Video Representation
  - Similarity Search



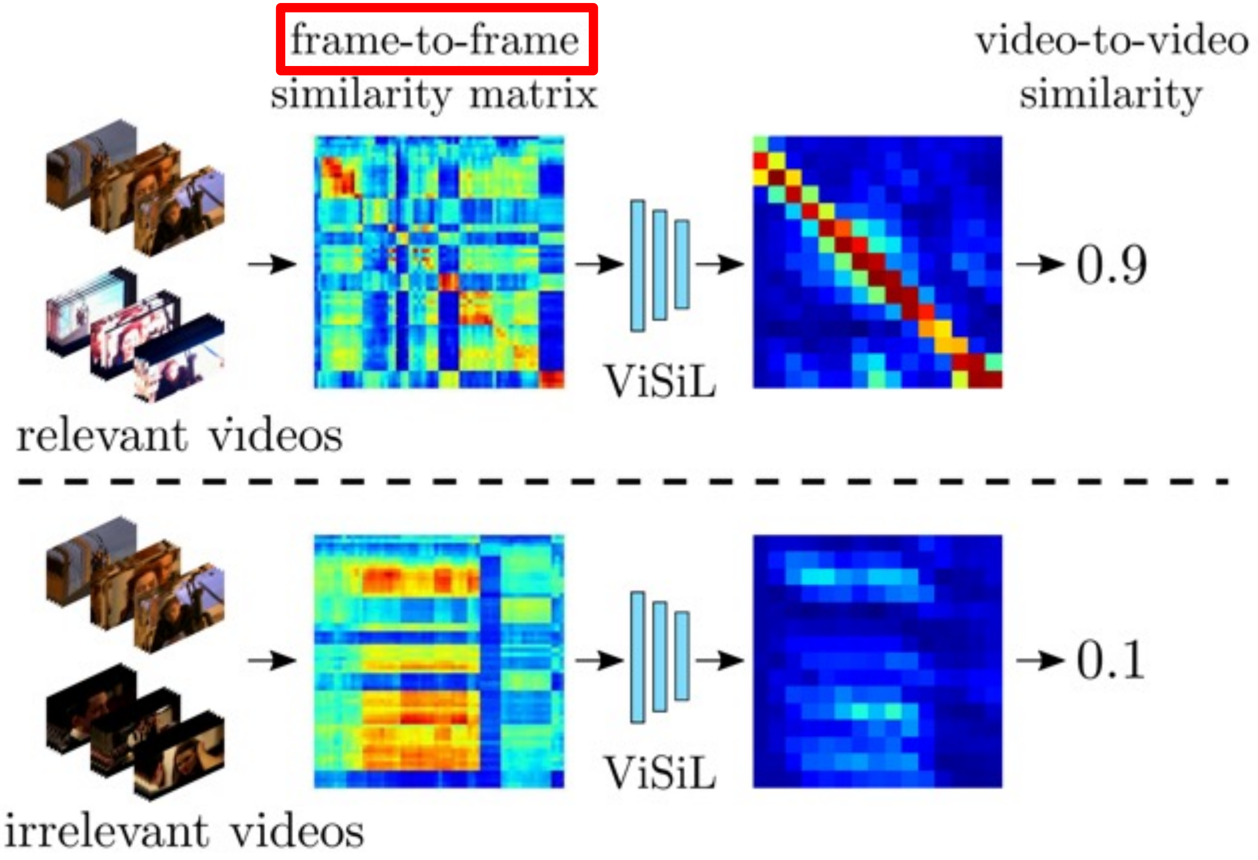
Related Work: Video Representation

- **A large amount of labeled videos are needed for the learning process.**

Jiang et al., VCDB: a large-scale database for partial copy detection in videos, ECCV 2014.

Related Work: Similarity Search

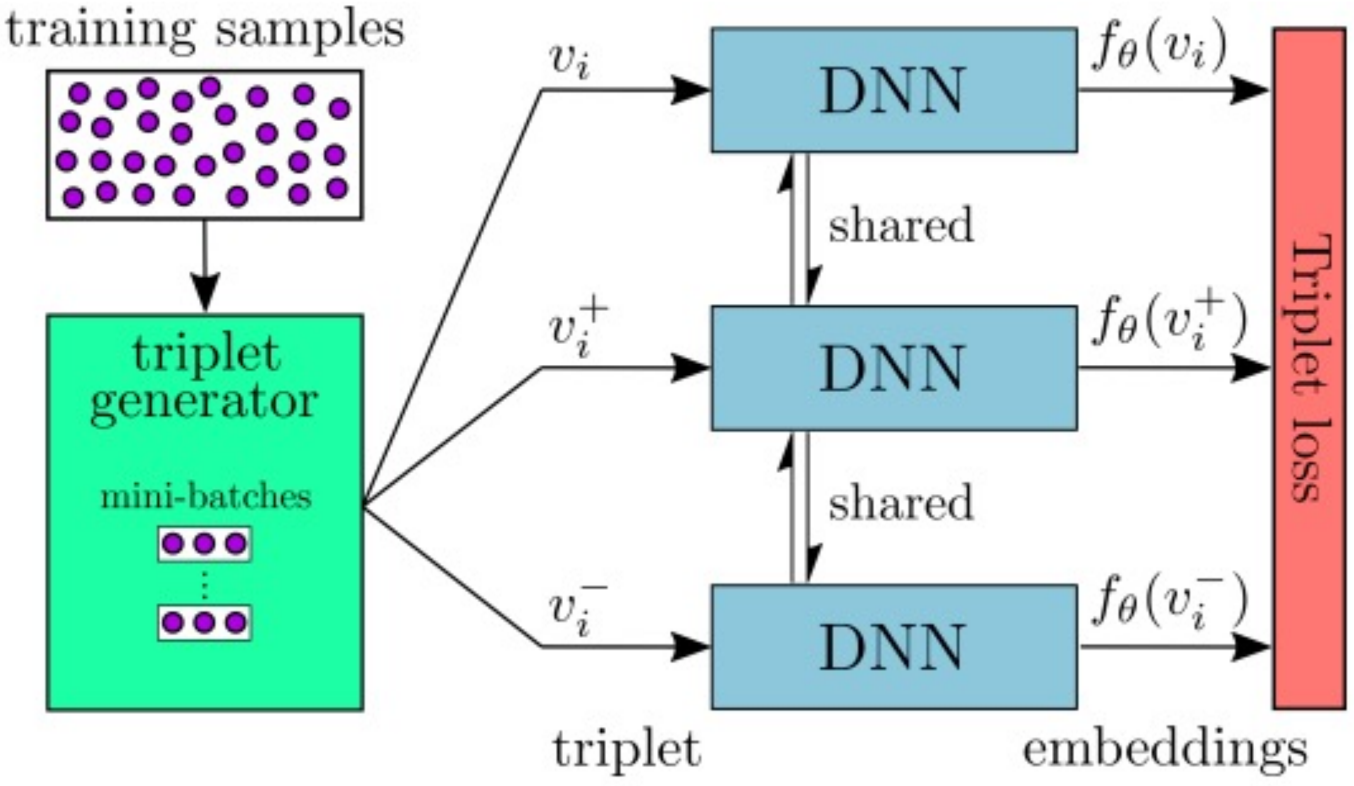
- **Based on** **frame-level** **features**
  - storage expensive and computationally expensive



Kordopatis-Zilos et al., ViSiL: Fine-grained Spatio-Temporal Video Similarity Learning, ICCV 2019.

Related Work: Similarity Search

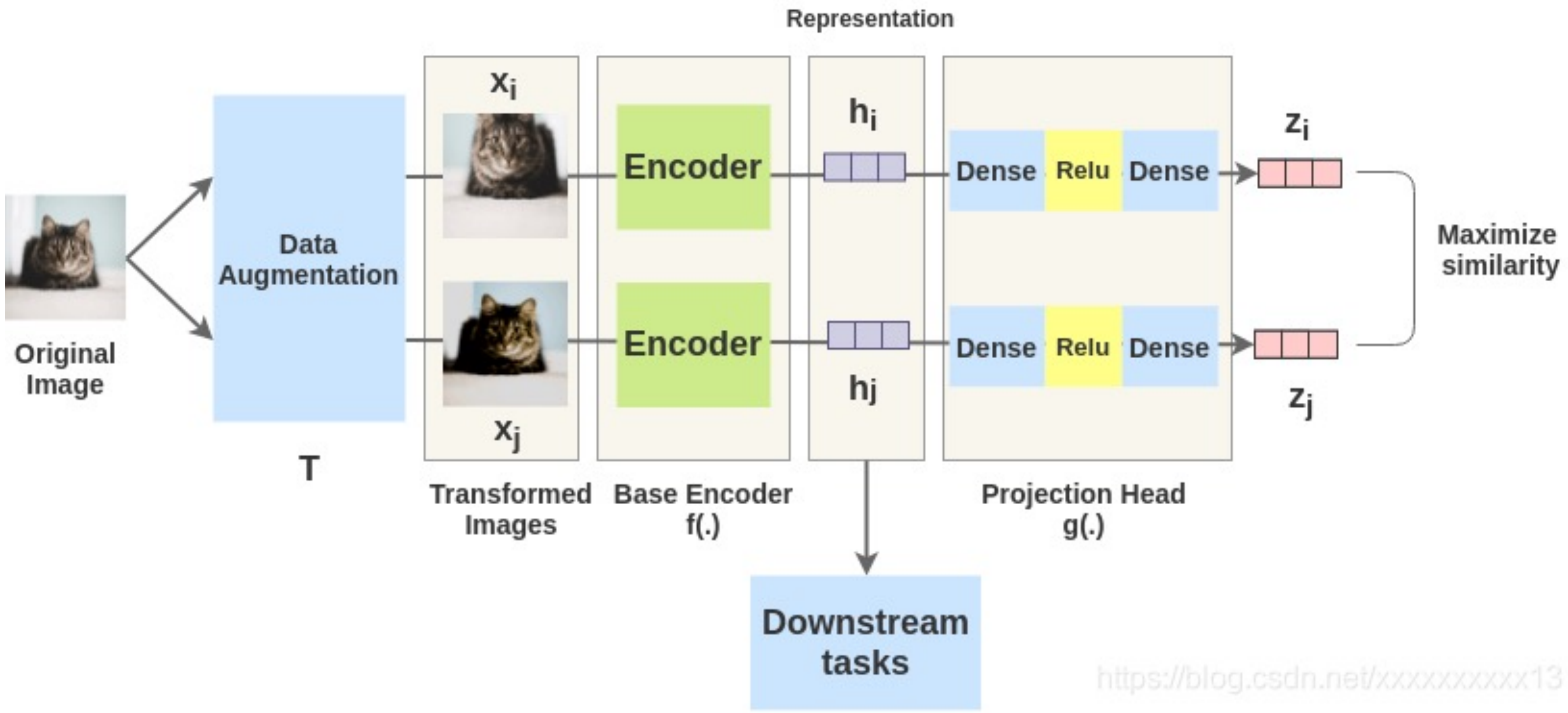
- **Based on video-level features**
  - insufficient to capture crucial details of individual videos



Kordopatis-Zilos et al., Near-Duplicate Video Retrieval with Deep Metric Learning, ICCVW 2017.

Motivation

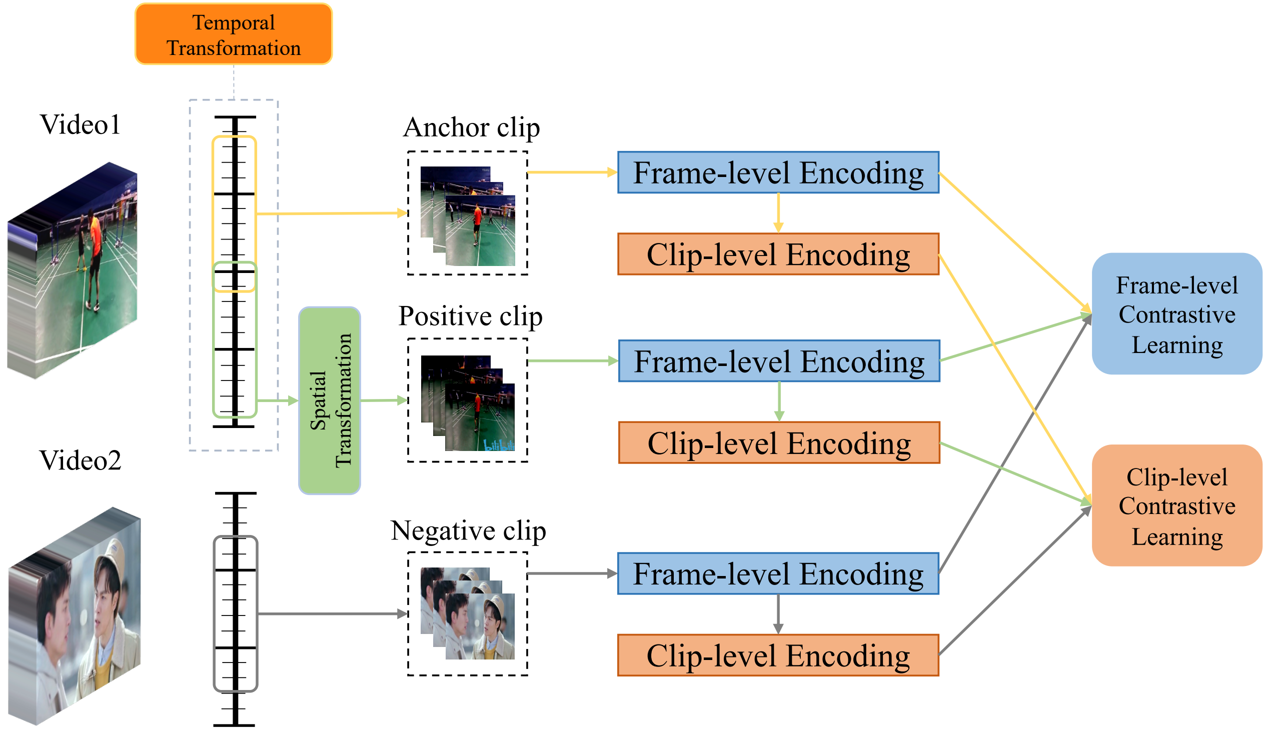
- **Contrastive learning**
  - learn visual representation from large amounts of unlabeled data



Chen et al., A Simple Framework for Contrastive Learning of Visual Representations, ICML 2020.

Our Contribution

- **We propose a video representation learning (VRL) approach**
  - **Frame-level encoding** is proposed to learn the frame-level feature with the pairs of the video frames and their transformations, thus **avoiding the high cost in manual annotation**
  - **Clip-level encoding** is proposed to aggregate frame-level features into clip-level, leading to significant **reduction in both storage space and search complexity**

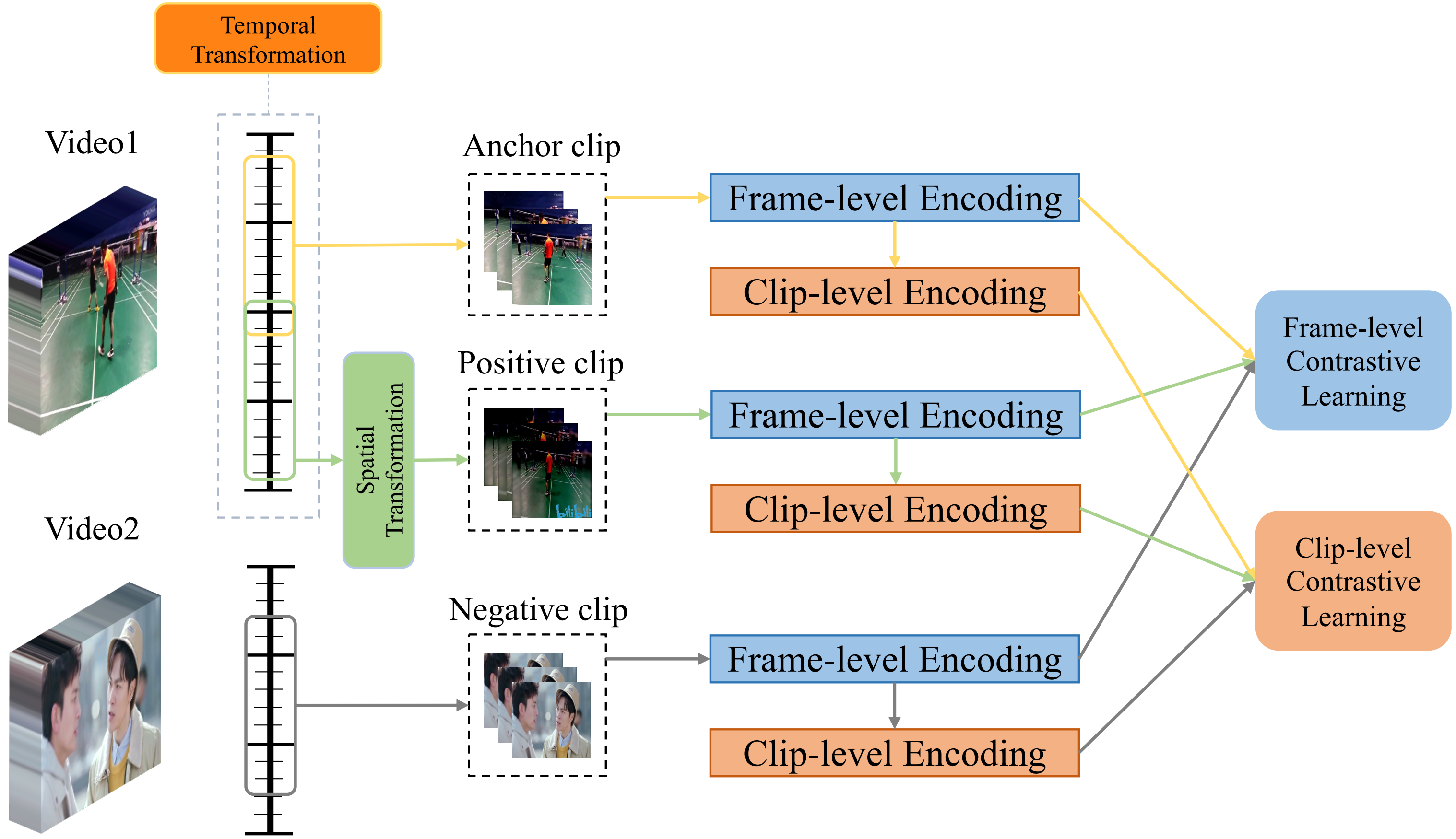


Guide Line

- Introduction
- Our Approach
- Experiment
- Conclusion

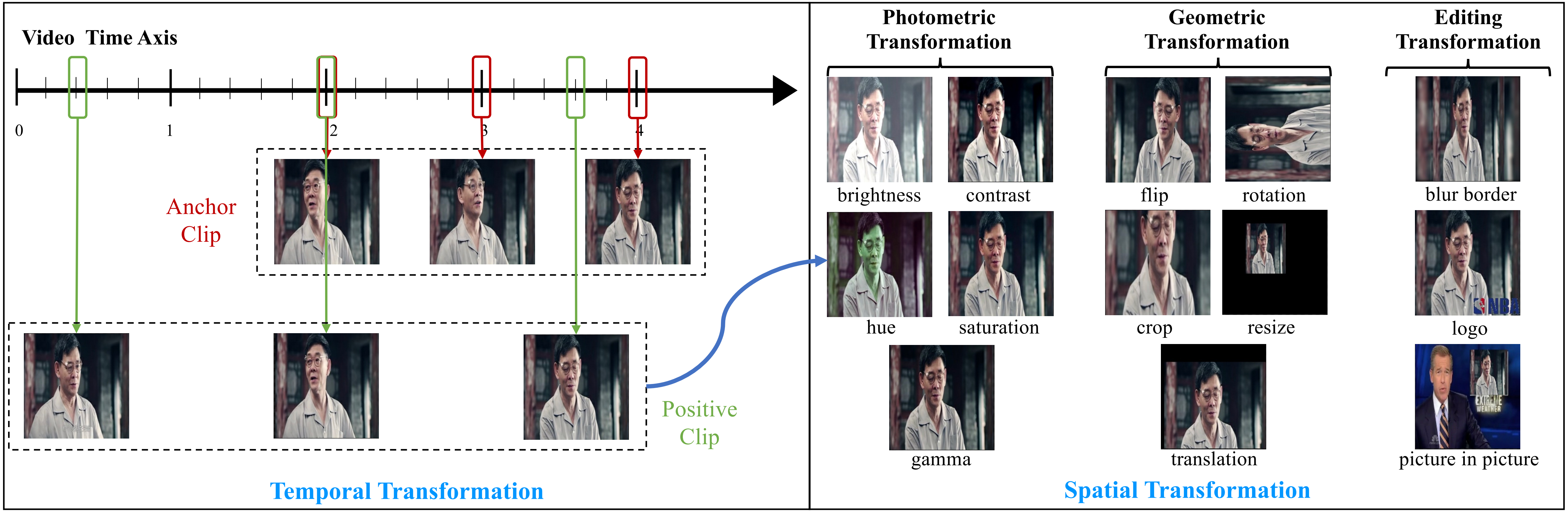
Our Approach

- **Frame-level Encoding**
  - Self-generation of Training Data
  - Spatial Structure Encoding
- **Clip-level Encoding**
  - Temporal Structure Encoding
  - Masked Frame Modeling
- **Video Similarity Calculation**



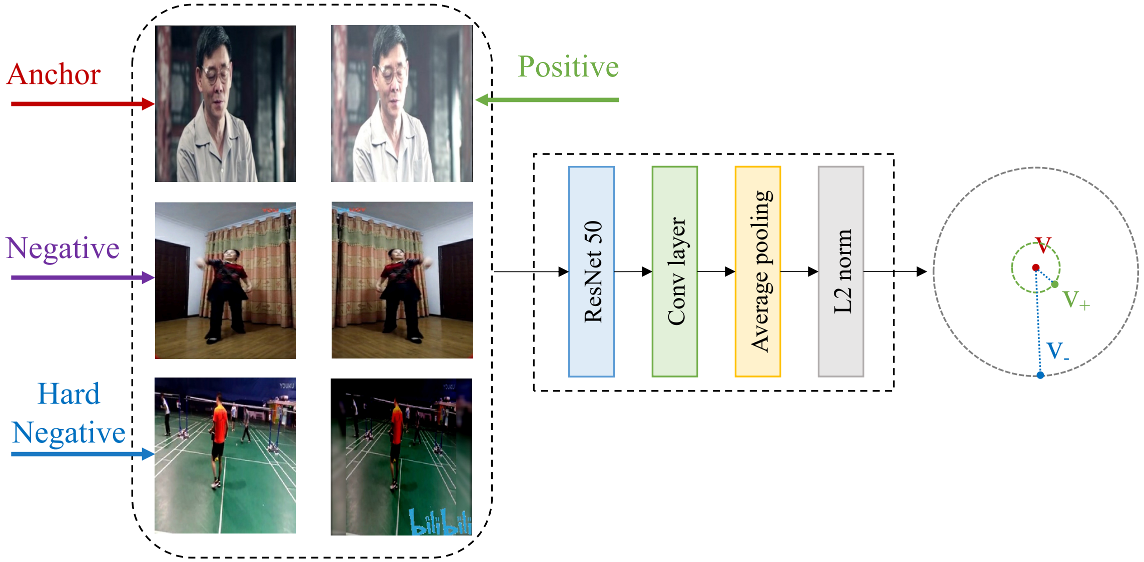
Frame-level Encoding

- **Self-generation of Training Data**
  - Temporal Transformation
  - Spatial Transformation



Frame-level Encoding

- **Spatial Structure Encoding**
  - Backbone: ResNet-50
  - Loss Function: adapted NCE loss



Given a set of frame-level features $S_F = \{(v^t, v^t_+)\}_{t=1}^{N}$

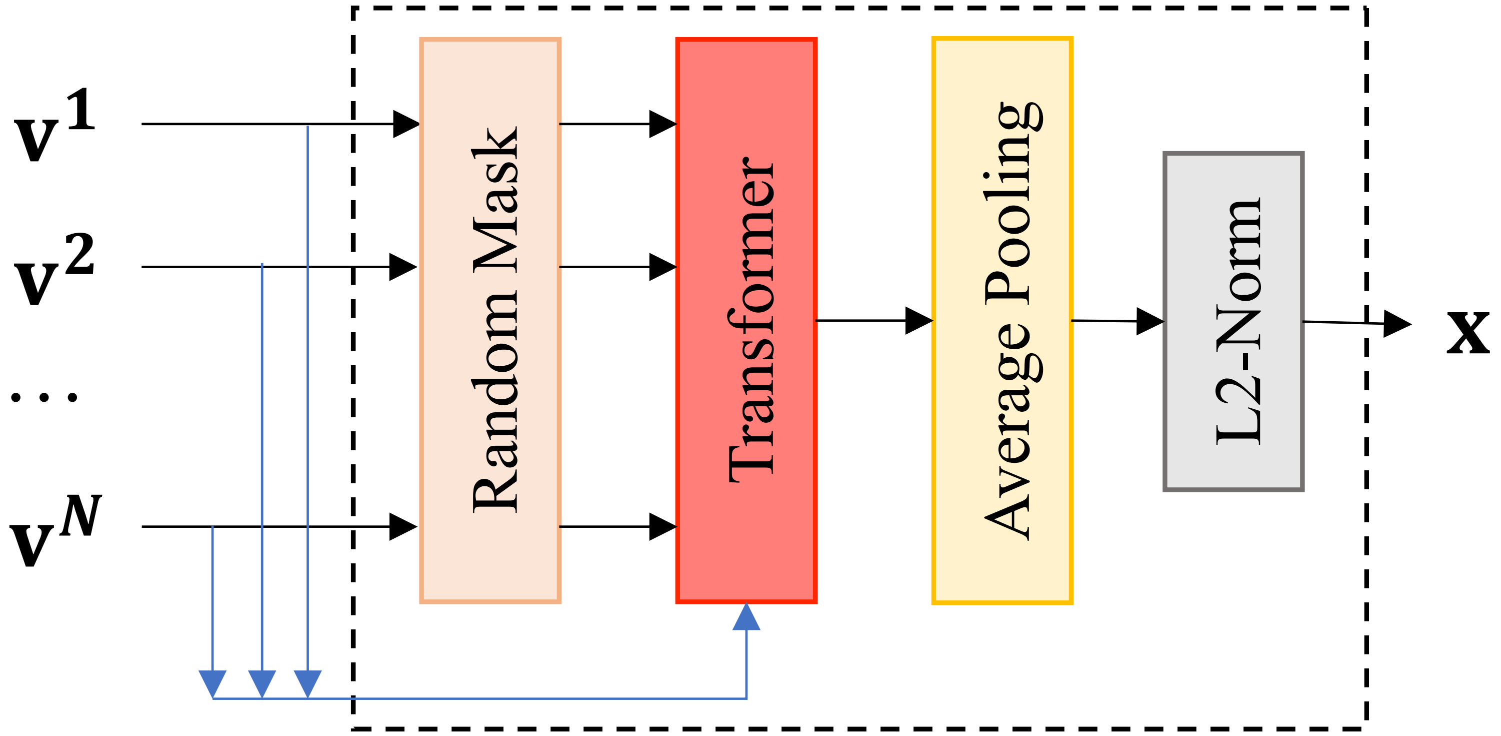
Calculate adapted noise contrastive estimation loss

$$L_F = \frac{1}{N}\sum_{t=1}^{N} -\mathbb{E}_{P_d}\log \boxed{P(D=1|v^t, v^t_+)} - (1-\mathbb{E}_{P_d})\log(1-\boxed{P(D=1|v^t, v^t_+)}) \quad (1)$$

$$P(D=1|v^t, v^t_+) = \frac{\exp({v^t}^{\mathrm{T}} v^t_+)}{\exp({v^t}^{\mathrm{T}} v^t_+) + \max\limits_{v_- \notin S_F} \exp({v^t}^{\mathrm{T}} v_-)} \quad (2)$$

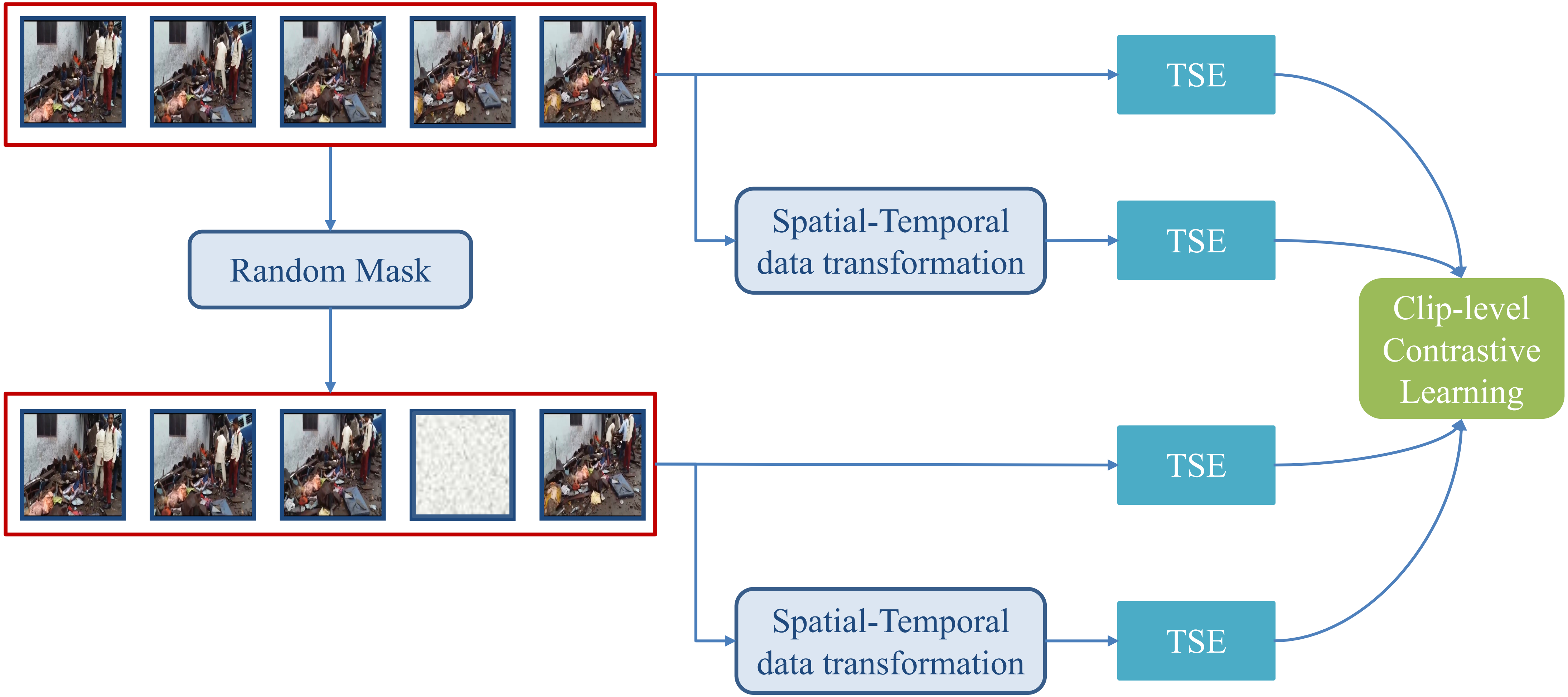
Clip-level Encoding

- **Clip-level Set Transformer Network**
  - Temporal Structure Encoding
  - Masked Frame Modeling



Clip-level Encoding

- **Masked Frame Modeling**

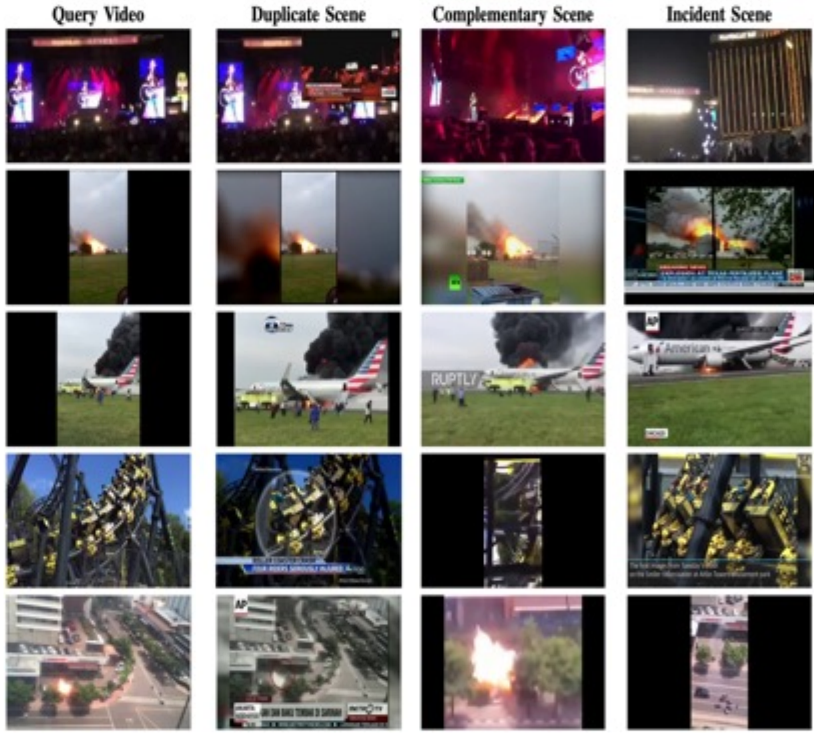


Guide Line

- Introduction
- Our Approach
- Experiment
- Conclusion

Dataset

- **Self-Transformation**
  - 3,000 hours' videos
  - Unlabeled data
- **FIVR-200K**
  - 225,960 videos
  - 100 queries
- **SVD**
  - 562,013 short videos
  - 1,206 queries



Comparisons with State-of-the-art Methods

- **On FIVR-200K dataset**
  - Compare with frame-level retrieval approach, our VRL approach outperforms all state-of-the-art methods except VisiL

Feature	Methods	Feature Dim/#bits	DSVR	CSVR	ISVR
Video-level	HC[36]	-	0.265	0.247	0.193
	DML[7]	500D	0.398	0.378	0.309
	TCA_c[9]	2048D	0.570	0.553	0.473
Frame-level	CNN-L[10]	4096D	0.710	0.675	0.572
	PPT[11]	4096D	0.775	0.740	0.632
	TN[12]	-	0.724	0.699	0.589
	TCA_f[9]	2048D	0.877	0.830	0.703
	VisiL[8]	9x3840D	0.892	0.841	0.702
	VRL_f	**512 bits**	**0.900**	**0.858**	**0.709**
Clip-level	**VRL**	**512 bits**	**0.876**	**0.835**	**0.686**

Comparisons with State-of-the-art Methods

- **On FIVR-200K dataset**
  - In frame-level features, our VRLf approach can achieve better retrieval performance than VisiL without any complex calculation

Feature	Methods	Feature Dim/#bits	DSVR	CSVR	ISVR
Video-level	HC[36]	-	0.265	0.247	0.193
	DML[7]	500D	0.398	0.378	0.309
	TCA_c[9]	2048D	0.570	0.553	0.473
Frame-level	CNN-L[10]	4096D	0.710	0.675	0.572
	PPT[11]	4096D	0.775	0.740	0.632
	TN[12]	-	0.724	0.699	0.589
	TCA_f[9]	2048D	0.877	0.830	0.703
	VisiL[8]	9x3840D	0.892	0.841	0.702
	VRL_f	**512 bits**	**0.900**	**0.858**	**0.709**
Clip-level	**VRL**	**512 bits**	**0.876**	**0.835**	**0.686**

Comparisons with State-of-the-art Methods

- **On FIVR-200K dataset**
  - Our VRL approach achieves significant improvements by 30.6%, 28.2%, 21.3% mAPs on the DSVR, CSVR and ISVR tasks

Feature	Methods	Feature Dim/#bits	DSVR	CSVR	ISVR
Video-level	HC[36]	-	0.265	0.247	0.193
	DML[7]	500D	0.398	0.378	0.309
	TCA_c[9]	2048D	0.570	0.553	0.473
Frame-level	CNN-L[10]	4096D	0.710	0.675	0.572
	PPT[11]	4096D	0.775	0.740	0.632
	TN[12]	-	0.724	0.699	0.589
	TCA_f[9]	2048D	0.877	0.830	0.703
	VisiL[8]	9x3840D	0.892	0.841	0.702
	VRL_f	**512 bits**	**0.900**	**0.858**	**0.709**
Clip-level	**VRL**	**512 bits**	**0.876**	**0.835**	**0.686**

Comparisons with State-of-the-art Methods

- **On SVD dataset**
  - Our VRL approach achieves best performance compared with both frame-level and video-level based methods

Feature	Methods	Feature Dim/#bits	Top-100 mAP
Video-level	DML[7]	500D	0.813
Frame-level	CNN-L[10]	4096D	0.610
	CNN-V[10]	4096D	0.251
	VRL$_f$	**512 bits**	**0.871**
Clip-level	**VRL**	**512 bits**	**0.860**

Effectiveness of Reducing Storage and Search Cost

- **On SVD dataset**
  - The storage of the frame-level features cost 1720.32 MB, while clip-level features only cost 366.98 MB, reducing the storage cost by 78.7%
  - Our VRL approach increases the retrieval speed by ~ 25times

Feature	Storage Space	Search Complexity
Frame-level	1720.32 MB	$O(M \times N)$
Clip-level	**366.98 MB**	$\sim \frac{1}{25} O(M \times N)$

Exploration of Flexible Retrieval Manners

- **On SVD dataset**
  - Provide more flexible retrieval manners, i.e. clip-to-clip retrieval and frame-to-clip retrieval
  - Use more fine-grained features (i.e. frame-level) can achieve better retrieval performance, which further verifies the effectiveness of clip-level encoding with masked frame modeling

Query	Database	Top-100 mAP
Clip-level	Clip-level	0.860
Frame-level	Clip-level	0.871

Ablation Study

- **Self-generation of Training Data**
  - VRL_f with all the three types of transformations achieves the best performance

Methods	Transformations			DSVR	CSVR	ISVR
	PT	GT	ET			
VRL_f	✓	✓	✓	**0.900**	**0.858**	**0.709**
A	✓	✓		0.868	0.818	0.673
B	✓		✓	0.881	0.825	0.662
C		✓	✓	0.868	0.815	0.649

- **Masked Frame Modeling**
  - Clip-level encoding with masked frame modeling improves the discrimination and robustness of the learned clip-level feature, and achieves better performance

Methods	SVD	FIVR-200K		
		DSVR	CSVR	ISVR
CE	0.854	0.870	0.834	**0.687**
CE w/ MFM	**0.860**	**0.876**	**0.835**	0.686

Guide Line

- Introduction
- Our Approach
- Experiment
- Conclusion

Conclusion

- We propose the VRL approach to encode the video in clip-level representation with contrastive learning to reduce the expensive cost of manual annotation, storage space and similarity search

- Frame-level encoding is to learn the discrimination and robustness of the learned feature with self-generation of training data

- Clip-level encoding is to reduce the redundancy of the frames in a clip, as well as make the model frame permutation and missing invariant, and support more flexible retrieval manners

Contact

Lab Homepage

Github Homepage

***M**ultimedia **I**nformation **P**rocessing **L**ab* (*MIPL*)

http://www.wict.pku.edu.cn/mipl/