CHAIN: Exploring Global-Local Spatio-Temporal Information for Improved Self-Supervised Video Hashing

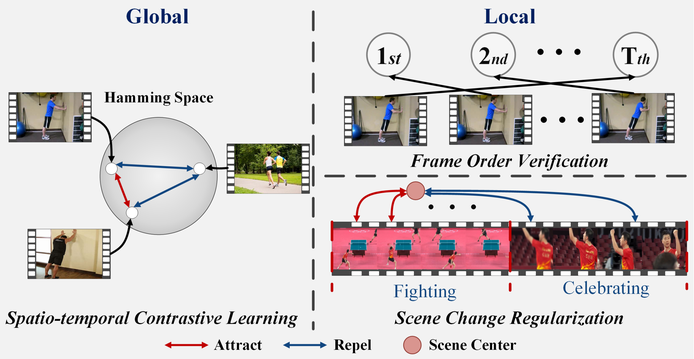
Rukai Wei[1], Yu Liu[1], Jingkuan Song[2], Heng Cui[1], Yanzhao Xie[1], Ke Zhou[1]

[1]Huazhong University of Science and Technology [2] University of Electronic Science and Technology of China

ACM Multimedia 2023

Introduction & Contribution

- Existing self-supervised video hashing methods have been effective in designing expressive temporal encoders, but have not fully utilized the temporal dynamics and spatial appearance of videos due to less challenging and unreliable learning tasks.



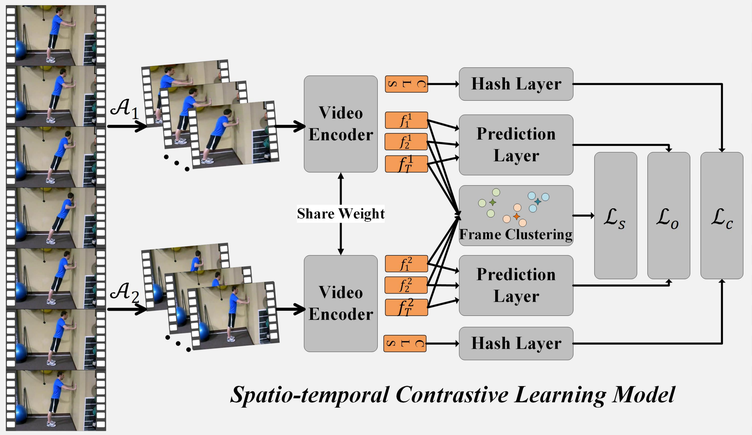
- Our proposed CHAIN incorporates three collaborative learning tasks to effectively explore both global and local spatio-temporal information for self-supervised hashing.

Our main contributions can be ourlined:

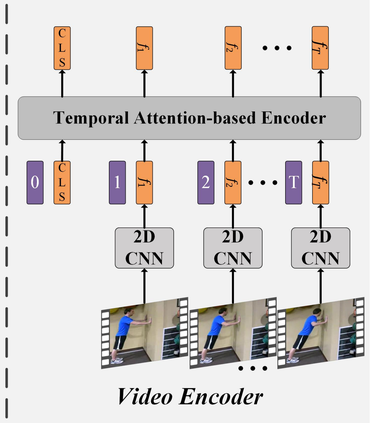
- For global spatio-temporal relationships between video instances, we propose a novel spatio-temporal contrastive learning framework, where we define both spatial and temporal augmentations to create positive pairs, enhancing the model to learn motion, scale, and viewpoint invariant hash codes for videos.
- For local temporal connections between video frames, we introduce a frame order verification task to predict the absolute temporal positions of video frames, achieving full exploitation of the inherent sequential structure.
- For local spatial relationships between video frames, we incorporate a scene change regularization task to distinguish various scenes within a video, allowing the model to capture the spatio-temporal variations for robust spatio-temporal modeling.

The Proposed CHAIN

- Our proposed CHAIN adopts a 2D CNN to extract frame-level representation, and leverages an attention-based temporal encoder to capture the long-range dependencies.
- We propose segment sampling based temporal augmentation and temporally-consistentspatial augmentation for contrastive learning to effectively sample high-quality positive pairs.
- We introduce frame order verification task, which predicts the absolute temporal order of frames within a video.
- In the scene change regularization task, we cluster all frame representations into several scenes using the AP algorithm and contrast between frame-level representations and scene prototypes with prototypical contrastive learning.



Spatio-temporal Contrastive Learning Model



Video Encoder

Comparisons with SOTAs

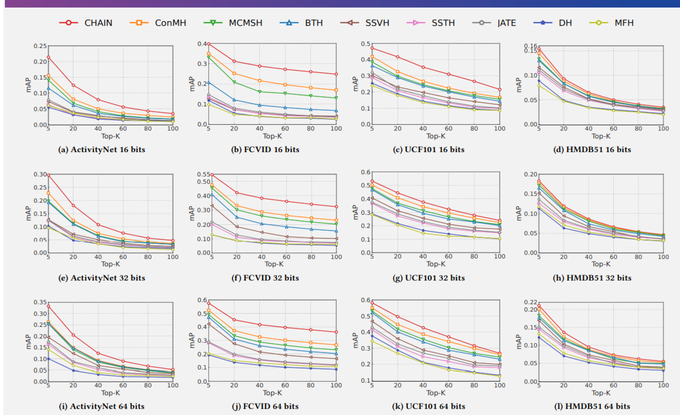


Figure 3: Retrieval performance compared with state-of-the-art methods in terms of mAP@K over four datasets.

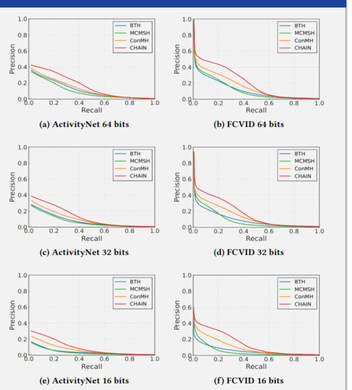


Figure 4: PR curves of the proposed CHAIN and the state-of-the-art baselines, including ConMH, MCMSH, and BTH on FCVID and ActivityNet.

- ✓ On the FCVID, UCF101, ActivityNet, and HMDB51 datasets, CHAIN consistently achieves higher mAP results than all state-of-the-arts methods.
- ✓ CHAIN achieves higher precision at the same recall rate than other methods.

Visualization

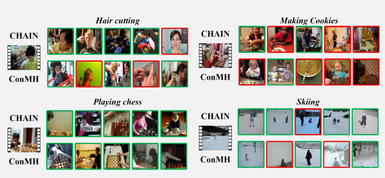


Figure 5: Top-5 retrieved results of CHAIN and ConMH on FCVID dataset. The video inside the green square is correctly retrieved, while the video inside the red square is incorrect.

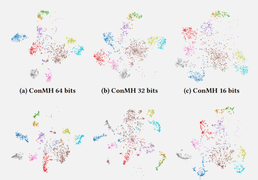


Figure 6: The t-SNE visualization of the learned 64-bit hash codes from the test set of FCVID. The scattered points of the same color indicate the same category. Note that we only visualize the first 10 classes.

- ✓ CHAIN exhibits a more stable performance in retrieving more relevant videos.
- ✓ The hash codes generated by CHAIN exhibit more distinct compactness for the same category and dispersion for dissimilar ones, in comparison to ConMH.