[Journal of Physics: Conference Series] STRNN: End-to-end deep learning framework for video partial copy detection

Yanzhu Hu, Zhongkai Mu and Xinbo Ai
Beijing University of Posts and Telecommunications

2022-02-21

Video Alignment?

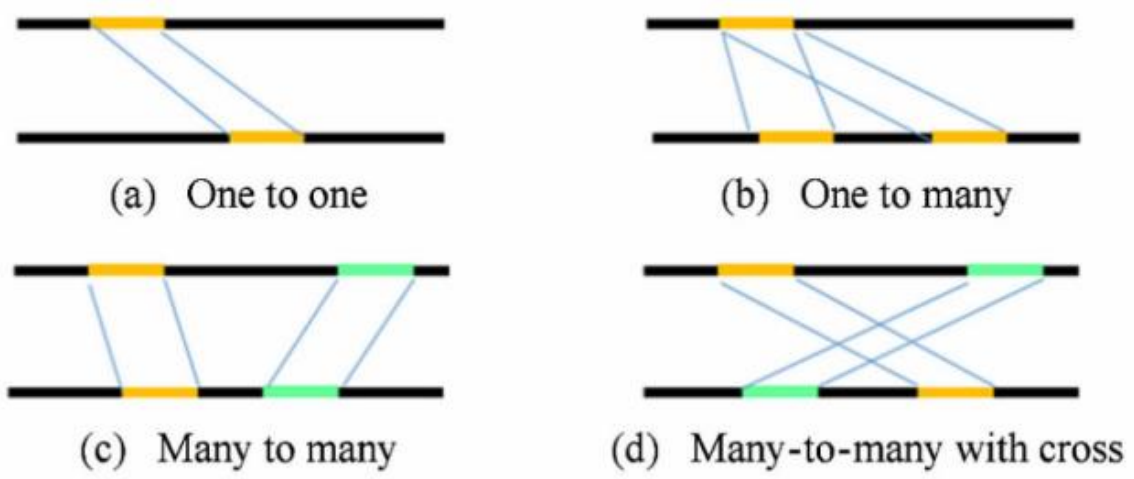
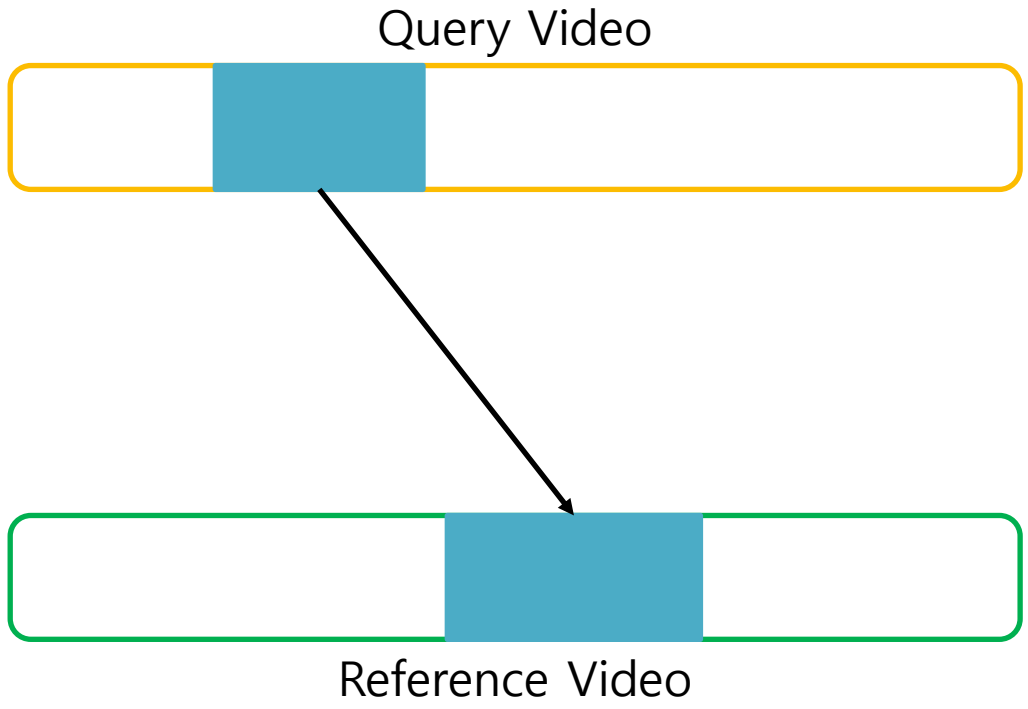


Fig.2 Four alignments mode of video partial copy clips



Video Alignment?

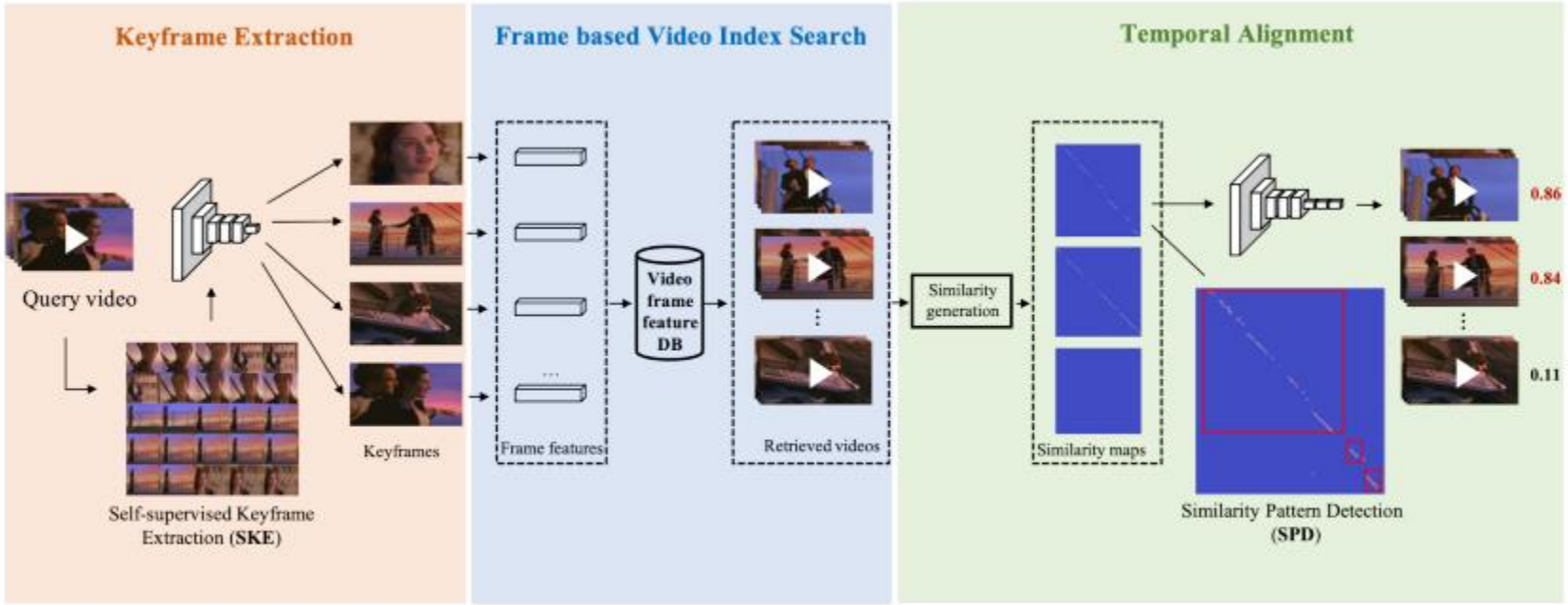


Figure 1: Query process of our proposed approach on Segment-level Content Based Video Retrieval (S-CBVR)

"Learning Segment Similarity and Alignment in Large-Scale Content Based Video Retrieval" published by ACMM,

STRNN's alignment methodology was taken and used in its SPD module.

Problem definition in Video Alignment

- Machine learning method
  - use local features such as SIFT to match and find similar video frame pairs, and then time alignment
- Deep learning method
  - Retrieval performance is good, however performance of time alignment is still disappointing

"Therefore, we propose a novel space-time feature fusion framework."

STRNN

STRNN

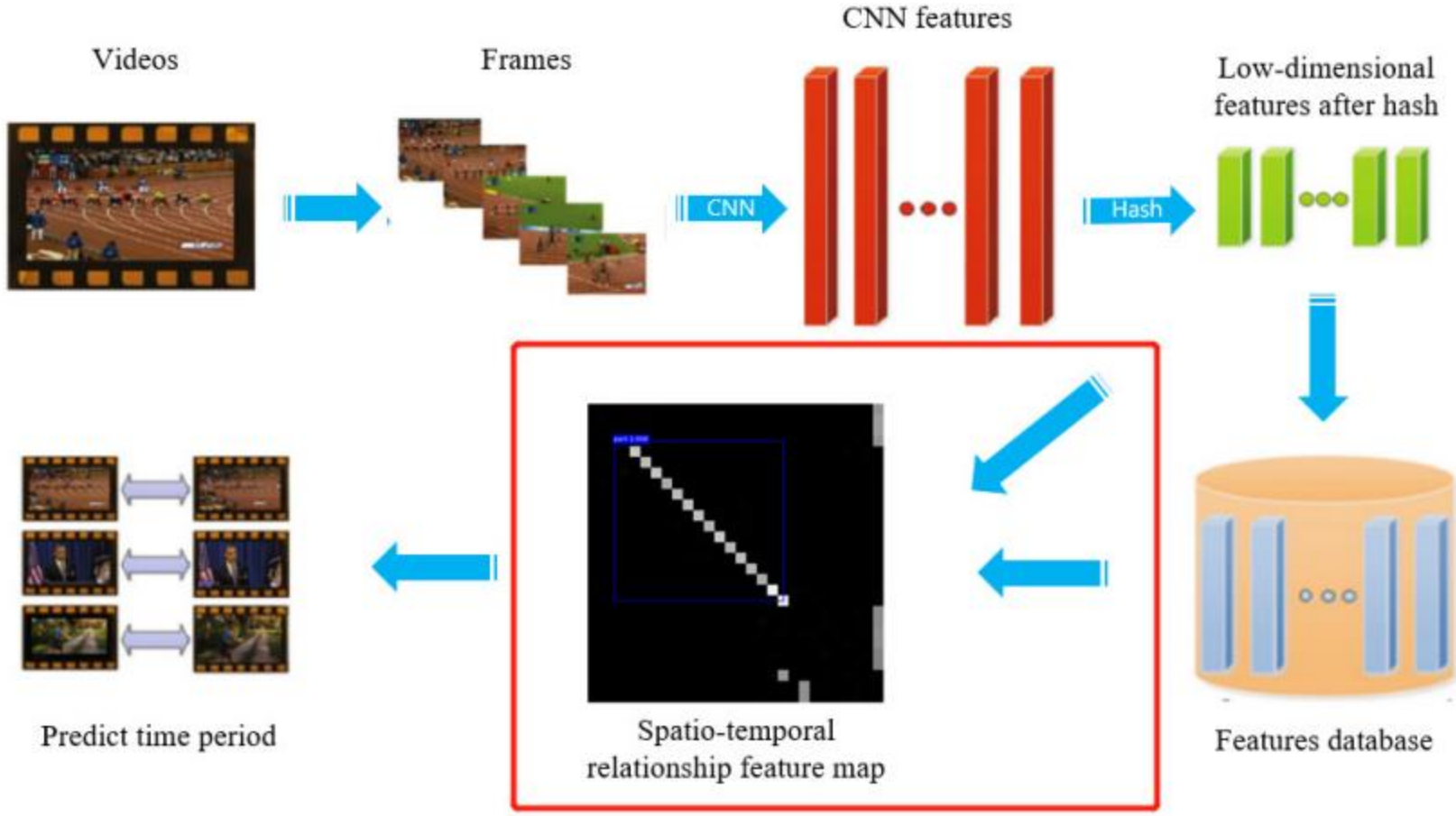
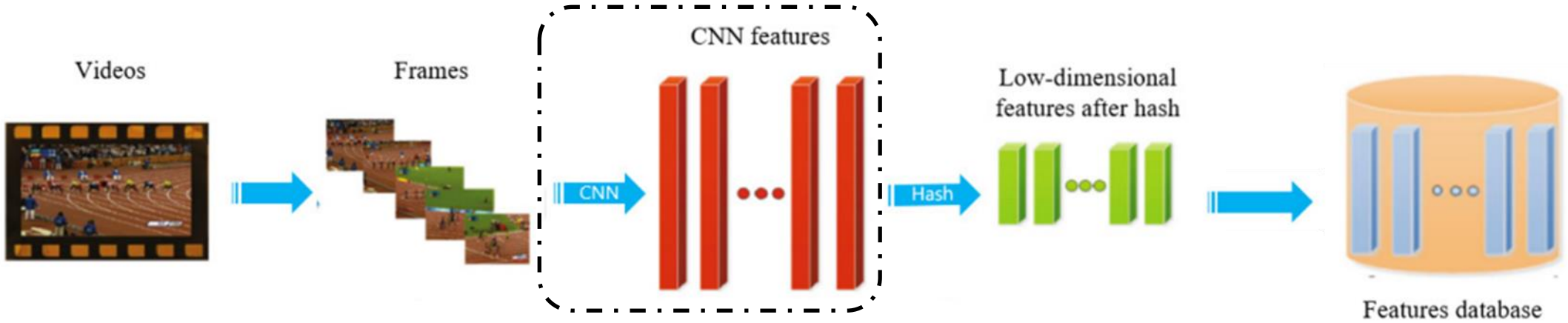
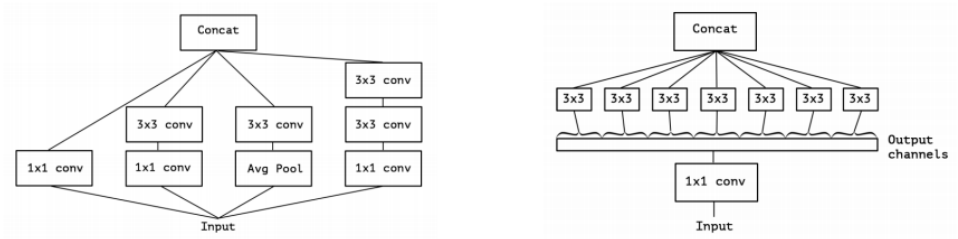


Fig.3 The framework of partial copy detection in videos based on STRNN

STRNN



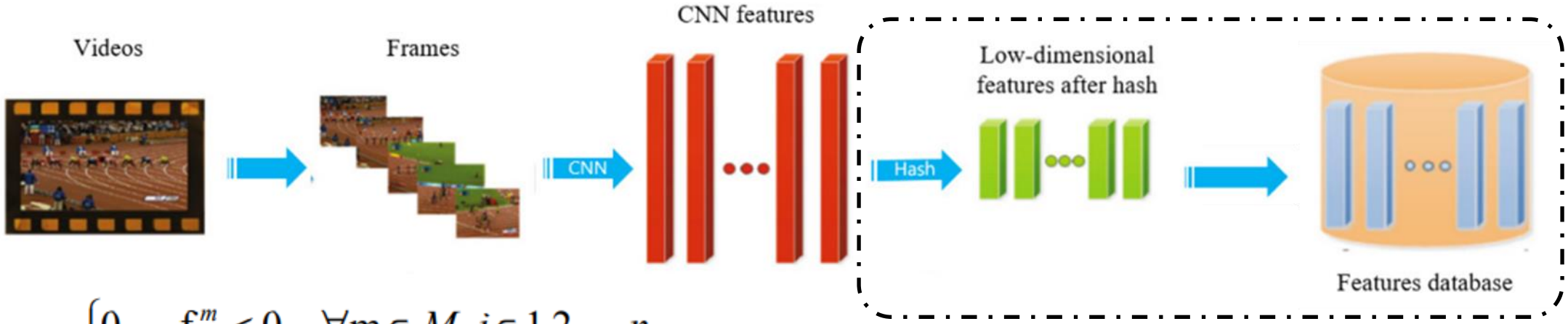


(a) Inception basic unit (b) Xception basic unit

Fig.4 Basic unit of Inception and Xception modules

- Used Xception model for CNN feature.
  - To achieve a complete separation of channel correlation and spatial correlation, refining the learning objectives of each convolution kernel.
- Extract frame-level features using the Global Average Pooling

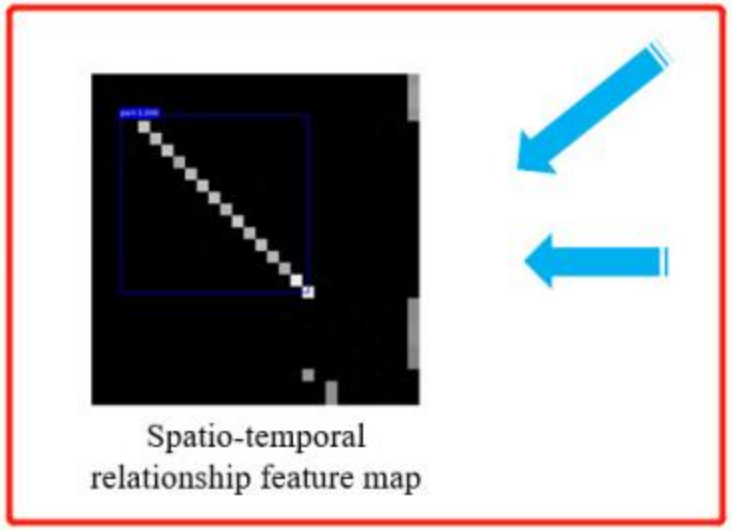
STRNN



$$s_i^m = \begin{cases} 0, & f_i^m < 0, \quad \forall m \in M, i \in 1,2,...,n \\ 1, & f_i^m > 0, \quad \forall m \in M, i \in 1,2,...,n \end{cases}$$

- Add a fully connected layer containing M neurons between this layer and the output layer and call it a hash layer.
- Add tanh activation function for binary mapping process.
- Use inverted file structure to index the quantized hash feature vectors filtering suspected copy frame.

STRNN



$$F = 255 * (R \times Q^T)$$

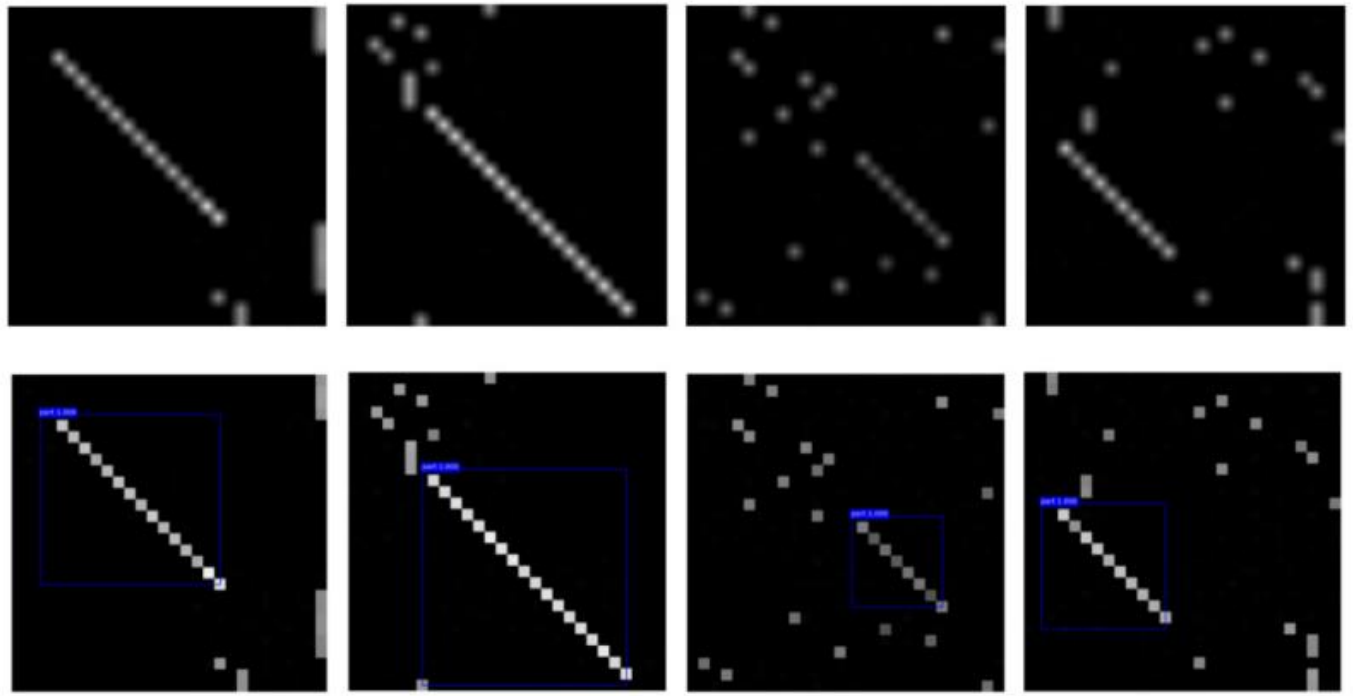
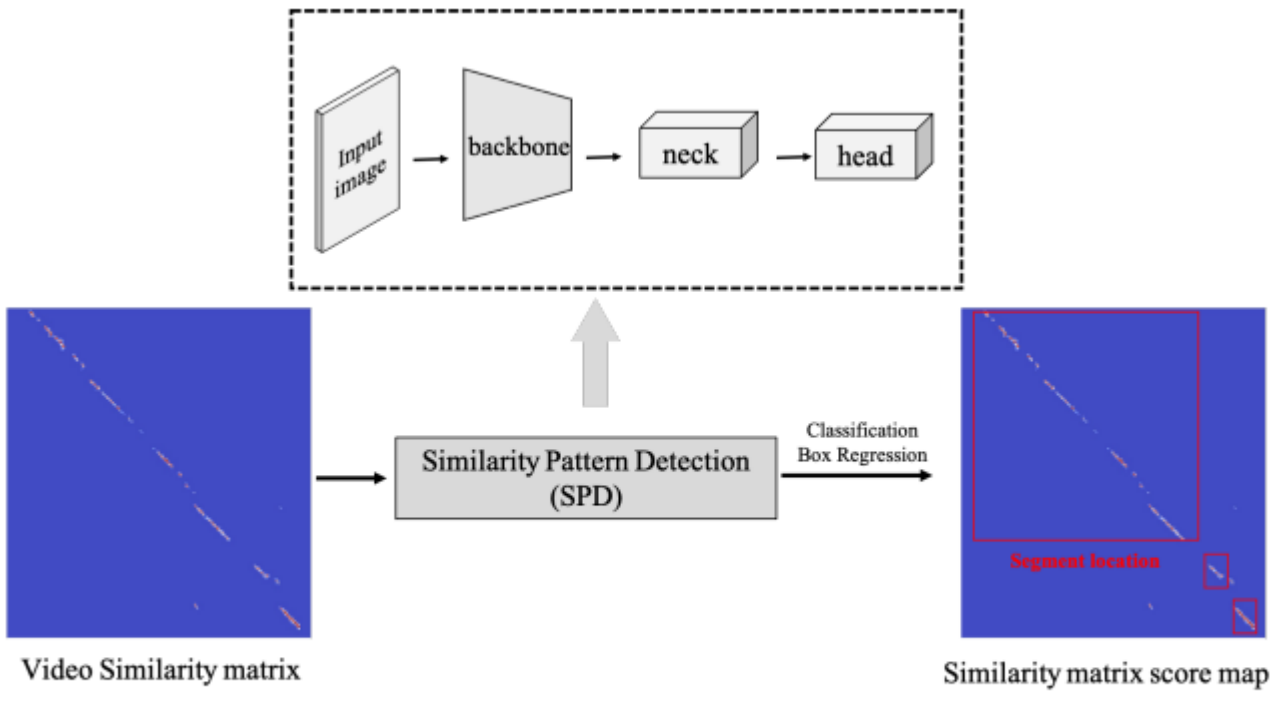


Fig.5 Spatio-temporal correlation feature map and partial copy regions detected by the model

- Spatio-temporal correlation feature map is obtained by multiplying query and reference feature array.
- Cosine values of the unit vectors computed by Q and R, range 0 to 1.
- RefineDet is used for object detection network to detect temporal alignment.

STRNN used in SSAN



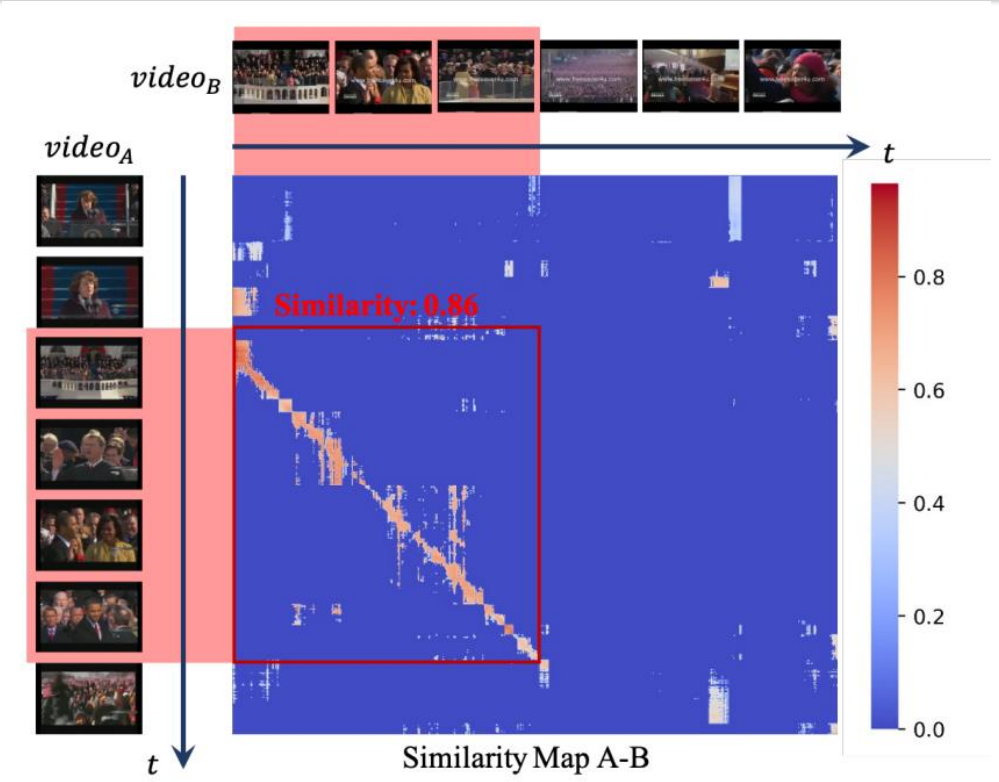
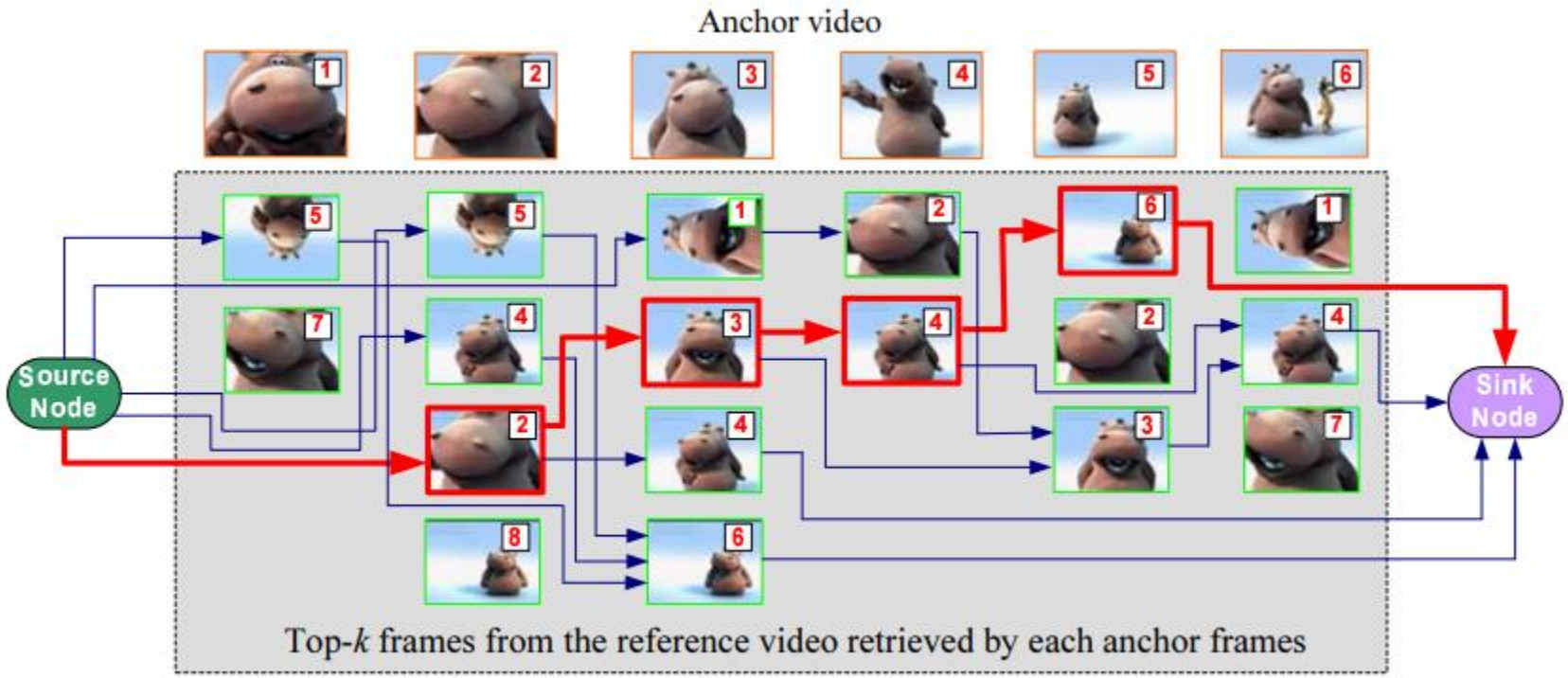


Figure 3: Similarity Pattern Detection (SPD) module

- Define alignment as a detection problem that detects similarity pattern on frame-to-frame similarity matrix S.
- Use BCE loss and GIoU loss for similarity pattern classification loss and bounding box location or similarity pattern location regression loss

Experiments

Experiments



- The graph-based Temporal Network (TN) takes matched frames as nodes and similarities between frames as weights of links to construct a network. And the matched clip is the weighted longest path in the network.

Experiments

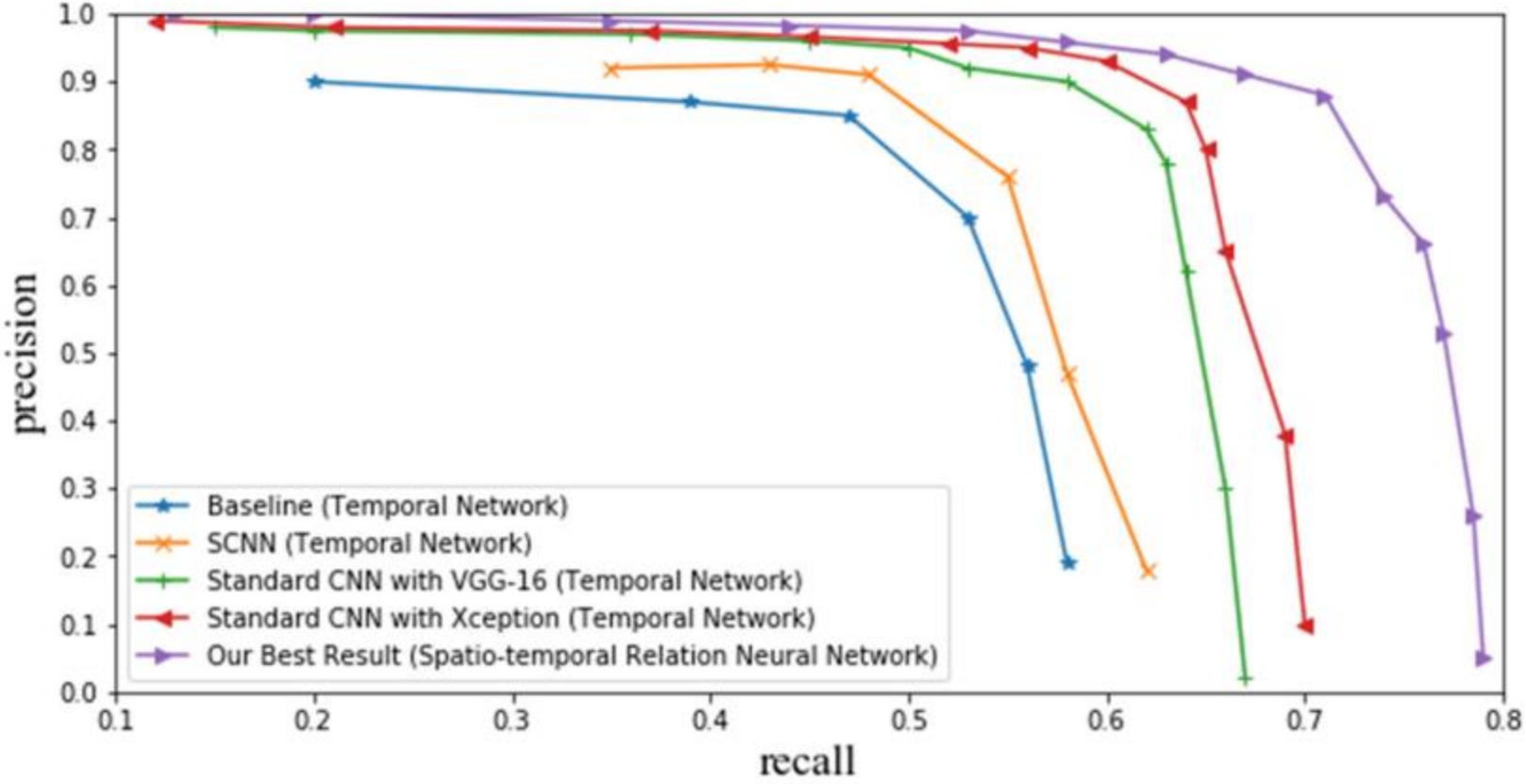


Fig.7 Precision-Recall curves for different methods on VCDB data-set

Q&A

Q&A