

# Self-supervised Video Retrieval Transformer Network

Xiangteng He, Yulin Pan, Mingqian Tang and Yiliang Lv  
Alibaba DAMO Academy

## Abstract

*Content-based video retrieval aims to find videos from a large video database that are similar to or even near-duplicate of a given query video. It plays an important role in many video related applications, including copyright protection, recommendation, filtering and etc.. Video representation and similarity search algorithms are crucial to any video retrieval system. To derive effective video representation, most video retrieval systems require a large amount of manually annotated data for training, making it costly inefficient. In addition, most retrieval systems are based on frame-level features for video similarity searching, making it expensive both storage wise and search wise. We propose a novel video retrieval system, termed SVRTN, that effectively addresses the above shortcomings. It first applies self-supervised training to effectively learn video representation from unlabeled data to avoid the expensive cost of manual annotation. Then, it exploits transformer structure to aggregate frame-level features into clip-level to reduce both storage space and search complexity. It can learn the complementary and discriminative information from the interactions among clip frames, as well as acquire the frame permutation and missing invariant ability to support more flexible retrieval manners. Comprehensive experiments on two challenging video retrieval datasets, namely FIVR-200K and SVD, verify the effectiveness of our proposed SVRTN method, which achieves the best performance of video retrieval on accuracy and efficiency.*

## 1. Introduction

These days, we have witnessed dramatic increase in the volume of videos generated over internet. At the same time, we have also observed a large number of videos that essentially steal contents from others, making video copyright protection and filtering an important demand. Content-based video retrieval addresses the problem by identifying a subset of videos from a large video database, which share similar contents with a given query video. It has drawn much attention, such as near-duplicate video retrieval [1] and fine-grained incident video retrieval [2].

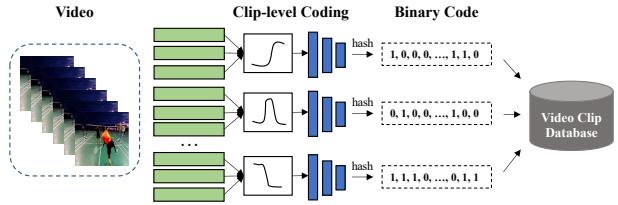


Figure 1. The illustration of clip-level video representation. Videos are split into shots via shot boundary detection, and further divided into clips at a fixed time interval. Then our SVRTN approach is applied to extract clip-level video representation. Finally, we binarize the clip-level features via hashing method for efficient retrieval.

To design an efficient and effective video retrieval system, two components are important, i.e. video representation and video search. For video representation, most existing methods [3, 4] apply supervised deep learning technologies to learn appropriate feature representation for accurate video content matching. Since a large amount of labeled videos are needed for training, it is costly to learn a robust and powerful video representation in this way.

For video search, most methods [4, 5, 6, 7] represent each video by a set of frame-level features, and the similarity between two videos are decided by the similarities between frames from the two videos followed by temporal alignment analysis, such as dynamic programming[6, 8], temporal network[7, 9, 10], and Temporal Hough Voting[11, 12]. The main shortcoming of these methods are of two folders. First, it needs to store all the frame-level features from all videos, making it storage expensive. Second, since the similarity measurement between two videos requires the similarity measurement between frames, making it computationally expensive. One common approach [3, 13] to address these limitations is to represent each video by a single vector (i.e. video-level features). Although these approaches help alleviate the problems of storage and computational cost, as pointed out in [14], they are insufficient to capture crucial details of individual videos, particularly for long videos.

To address the above problems, we propose a Self-supervised Video Retrieval Transformer Network

(SVRTN). It leverages self-supervised learning to learn video representation from unlabeled data, and exploits transformer structure to aggregate frame-level features into clip-level, as shown in Figure 1. More specifically, the key contributions of this work can be summarized as follow;

- **Self-supervised video representation learning** is proposed to learn the representation with the pairs of the video and its transformations, which are automatically generated by temporal and spatial transformations, thus avoiding the high cost in manual annotation. Due to the self-generation of training data, our SVRTN approach can learn better video representation from a large amount of unlabelled videos, leading to better generalization for its learned representation.
- **Clip-level set transformer network** is proposed to aggregate frame-level features into clip-level, leading to significant reduction in both storage space and search complexity. It can learn the complementary and variant information from the interactions among clip frames via self-attention mechanism, as well as acquires the frame permutation and missing invariant ability to handle the issue of missing frames, both of which increase the discrimination and robustness of the clip-level feature. Besides, it supports more flexible retrieval manners, such as clip-to-clip retrieval and frame-to-clip retrieval.

Comprehensive experiments on two challenging video retrieval datasets, namely FIVR-200K and SVD, verify the effectiveness of our SVRTN approach, which achieves the best performance of video retrieval on accuracy and efficiency. Compared with video-level methods, our SVRTN achieves the improvements of 30.6%, 28.2%, 21.3% mAPs on the DSVR, CSVR and ISVR tasks of FIVR-200K dataset, and 4.7% mAP on SVD dataset. Compared with frame-level methods, our SVRTN approach achieves comparable performance, and reduces about 78.7% of the feature storage cost and increases the retrieval speed by  $\sim 25$  times.

## 2. Related Work

Existing video retrieval methods can be divided into two categories: frame-level retrieval methods and video-level retrieval methods.

### 2.1. Frame-level Retrieval Methods

These methods generally extract frame-level features using CNN, and retrieve related frames by approximate nearest neighbor search. Various post-processing methods[11, 7, 10, 15, 16, 17] have been proposed to aggregate the frame-to-frame similarity matrix to video similarity score. Jiang et al. propose Temporal Hough Voting [11] to find

temporal alignments, which makes full use of the relative timestamp between matched frames. Tan et al. propose Graph-based Temporal Network [7] to detect the longest shared path between two compared videos. Hu and Lu [10] combine temporal network with a CNN+RNN feature encoder, to address the problem of partial copied detection. Another popular solution is based on Dynamic Programming(DP), which is applied to extract the biggest matched diagonal block from frame-to-frame similarity matrix, and tolerate limited horizontal and vertical movements for flexibility. Chou et al. [6] apply Bag-of-Words to represent frames, and propose m-pattern-based dynamic programming (mPDP) algorithm to localize near-duplicate segments and re-rank the retrieved videos. However, the above methods ignore exploiting spatial feature invariance, which is essential to video retrieval. Recently, Kordopatis-Zilos et al. [4] employ a region-level similarity calculation and aggregate region similarity matrix to frame similarities, which considers fine-grained spatial alignments and achieves high retrieval performance. These frame-level retrieval methods disregard the redundancy between successive frames, so that more computation cost will be needed, resulting in a low retrieval efficiency.

### 2.2. Video-level Retrieval Methods

Video-level retrieval methods encode the videos in video-level, and search for the  $k$ -nearest neighbors for the video-level feature of the query video in the embedding space. Various frame feature aggregation methods[3, 5, 18, 19, 20, 21] have been used to obtain a single video-level representation. Liang et al. [22] propose temporal pooling layer to aggregate the successive frames by the means of average pooling. Kordopatis-Zilos et al. [5] extract individual frame features from intermediate CNN layers, and adopt Bag-of-Words to compress them into a video-level representation, so that video similarity can be measured by calculating the cosine distance between the two video-level representations. Furthermore, Kordopatis-Zilos et al. [3] aggregate frame features by the means of average pooling, and introduce Deep Metric Learning(DML) to learn an embedding by minimizing the distance between related videos and maximizing the distance between irrelevant ones. Hash codes based methods [19, 23, 24, 25] are also widely used to encode unified spatial-temporal representation from videos. Song et al. [25] capture the temporal relationship between frames using an encoder-decoder architecture. Li et al. [23] apply the binary codes to capture spatial-temporal structure in a video by integrating the neighborhood attention mechanism into an RNN-based reconstruction scheme. However, these video-level retrieval methods generally perform worse than frame-level retrieval methods, which is mainly due to that single vector is hard to capture the entire spatio-temporal structure in a video sufficiently.

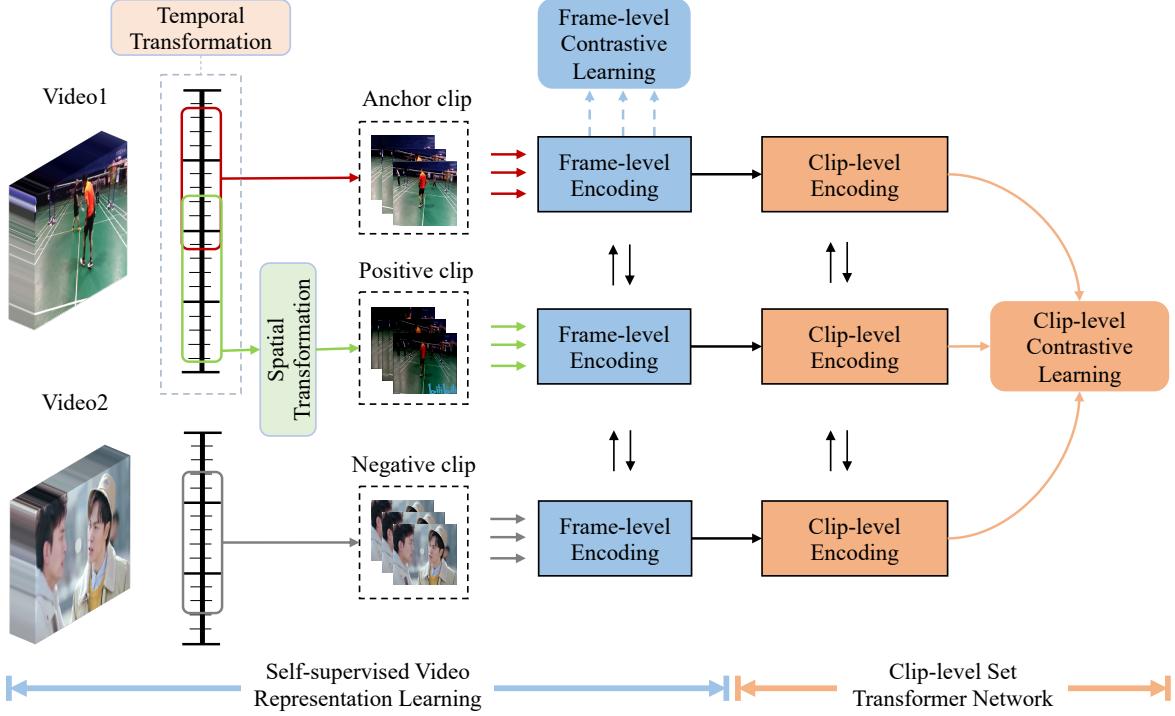


Figure 2. Overview of our SVRTN approach.

### 3. Self-supervised Video Retrieval Transformer Network

In this section, we present the proposed self-supervised video retrieval transformer network (SVRTN) for efficient retrieval by reducing the expensive cost of manual annotation, storage space and similarity search. It mainly consists of two components: self-supervised video representation learning and clip-level set transformer network, as shown in Figure 2. First, we automatically generate the video pairs via temporal and spatial transformations. Then, we utilize these video pairs as supervision to learn frame-level feature with contrastive learning. Finally, we aggregate the frame-level features into clip-level feature via self-attention mechanism, and increase the robustness via masked frame modeling.

#### 3.1. Self-supervised Video Representation Learning

Existing methods generally train their model with manual annotated video pairs. The more data, the better performance will be achieved [26]. However, the cost of annotation is too expensive to generate a large amount of training data. So it is normal that representation learning is restricted to the limited volume of training data. Inspired by the advance of recent self-supervised learning methods [27, 28, 29], we propose the self-supervised video representation learning to break the restriction, and exploit the

spatial-temporal invariant of representation to defense various video transformations.

##### 3.1.1 Self-generation of Training Data

First, we automatically collect a large amounts of videos from the video website. Then, temporal and spatial transformations are sequentially performed on these clips to construct the training data.

**(1) Temporal Transformation:** As shown in the left part of Figure 3, given a video, we first uniformly sample  $N$  frames with a fixed time interval  $r$  to generate the anchor clip, denoted as  $C = \{I^1, I^2, \dots, I^N\}$ . Then a frame  $I^m$  is randomly selected from the anchor clip as the identical content shared by anchor clip  $C$  and positive clip  $C_+$ . We regard the selected frame as the median frame of  $C_+$ , and uniformly sample  $\frac{N-1}{2}$  frames forward and backward respectively, with a different sample time interval  $r_+$ .

**(2) Spatial Transformation:** For each frame, we further perform spatial transformation. As shown in the right part of Figure 3, three types of spatial transformations are considered: (a) *Photometric transformation*. It includes the transformations of brightness, contrast, hue, saturation and gamma adjustment. (b) *Geometric transformation*. It includes the transformations of horizontal flip, rotation, crop, resize and translation. (c) *Editing transformation*. It includes the transformations of adding blurred background,

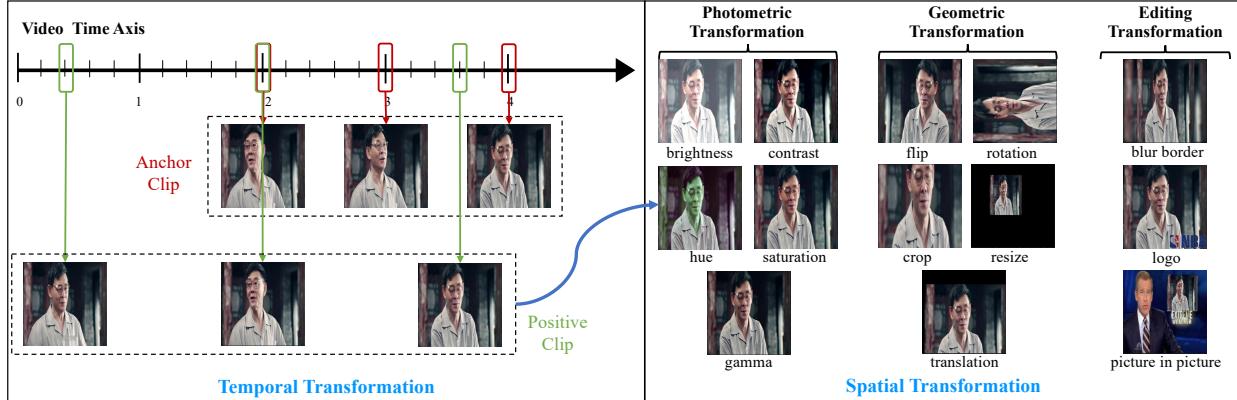


Figure 3. Illustration of self-generation of training data.

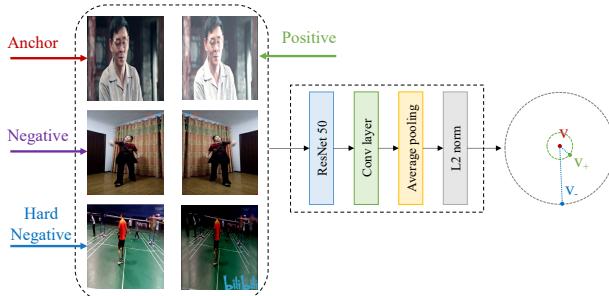


Figure 4. Model architecture.

adding logo, picture in picture and etc.. In training stage, we randomly select one transformation from each type of spatial transformation, and then apply them on frames from positive clips in sequence to generate the new positive clips.

### 3.1.2 Video Representation Learning

Since the supervised video pairs are generated, we use them to learn the video representation with frame-level contrastive loss. As shown in Figure 4, we adopts ResNet50 [30] as feature encoder, and then followed by a convolutional layer to reduce the channel number of the feature map, finally average pooling and  $L_2$  normalization are applied to obtain the frame-level feature.

The goal of video representation learning is to capture spatial structure from individual frames and ignore the impacts of various transformations, through minimizing the distance between features of the anchor clip frames and positive clip frames, as well as maximizing the distance between features of the anchor/positive clip frames and negative clip frames.

Specifically, given an anchor clip containing  $N$  frames

$C = \{I^1, I^2, \dots, I^N\}$ , then a positive clip is generated via temporal and spatial transformations, denoted as  $C_+ = \{I_+^1, I_+^2, \dots, I_+^N\}$ . We organize these frames in semantic-related pairs  $\{(I^t, I_+^t)\}_{t=1}^N$ . Then video representation learning is employed to encode spatial structure from individual frames, which is formulated as

$$v = f_S(I) \quad (1)$$

Since a set of frame-level features  $S_F = \{(v^t, v_+^t)\}_{t=1}^N$  is obtained, a contrastive learning is adopted to drive the features more discriminative and robust. The loss function is an adapted noise contrastive estimation loss [31], and its definition is as follows:

$$\begin{aligned} L_F = & \frac{1}{N} \sum_{t=1}^N -\mathbb{E}_{P_d} \log P(D = 1|v^t, v_+^t) \\ & - (1 - \mathbb{E}_{P_d}) \log(1 - P(D = 1|v^t, v_+^t)) \end{aligned} \quad (2)$$

where  $P_d$  denotes the actual data distribution and  $\mathbb{E}_{P_d} = 1$  indicates  $I^t$  and  $I_+^t$  share absolutely identical visual semantic. The probability of the encoded vectors  $v^t$  with  $v_+^t$  is from the data distribution  $P(D = 1|v^t, v_+^t)$  can be defined as :

$$P(D = 1|v^t, v_+^t) = \frac{\exp(v^t v_+^t)}{\exp(v^t v_+^t) + \max_{v_- \notin S_F} \exp(v^t v_-)} \quad (3)$$

where  $v_-$  indicates the feature of frame from the negative chip, which is semantic-irrelevant with anchor clip. It is noted that only the batch-hardest negative frame will contribute to the  $P(D = 1|v^t, v_+^t)$ , because the simple negative frames will decrease the discriminability of the learned feature.

### 3.2. Clip-level Set Transformer Network

Since the adjacent frames from one clip have the similar content, the frame-level features have high redundancy between each other, and the complementary information is not fully explored. Therefore, we aggregate the frame-level features into clip-level feature in this paper.

Specifically, given a clip, a set of frame-level features  $\{v^1, v^2, \dots, v^N\}$  are extracted through self-supervised video representation learning, then aggregated into a single clip-level feature  $x$ , which is defined as follows:

$$x = f_C(v^1, v^2, \dots, v^N) \quad (4)$$

To encode the clip-level feature, we propose an adapted Transformer [32], called clip-level set transformer network, whose architecture is shown in Figure 5. Instead of directly using Transformer to encode the clip-level feature, we apply the idea of set retrieval [33] in the clip-level encoding. It is noted that we only use one encoder layer with 8 attention heads, without position embedding. It enables our SVRTN approach has the abilities: (1) More robust. Increase the robustness of the learned clip-level features with the ability of frame permutation and missing invariant. (2) More flexible. Support more retrieval manners, including clip-to-clip retrieval and frame-to-clip retrieval.

#### 3.2.1 Clip-level Encoding

Similar with frame-level encoding in Section 3.1.2, given a set of clip-level features  $S_C = \{(x^b, x_+^b)\}_{b=1}^B$ , where  $B$  is the number of clips in a batch, a clip-level contrastive learning is adopted. The loss function is defined as follows:

$$\begin{aligned} L_C(x, x_+) = & \frac{1}{B} \sum_{b=1}^B -\mathbb{E}_{P_d} \log P(D = 1|x^b, x_+^b) \\ & - (1 - \mathbb{E}_{P_d}) \log(1 - P(D = 1|x^b, x_+^b)) \end{aligned} \quad (5)$$

where  $P_d$  denotes the actual data distribution and  $\mathbb{E}_{P_d}$  is set to 1 indicates the anchor clip and positive clip share absolutely identical visual semantic.  $P(D = 1|x^b, x_+^b)$  denotes the posterior probability that  $x$  with  $x_+$  is from the actual data distribution, its definition is similar with Equation (3). Clip-level encoding can learn the complementary information from the frames of the video clip via self-attention mechanism of Transformer, and hence the discrimination of features via attentively seeing the frames.

#### 3.2.2 Clip-level Encoding with Masked Frame Modeling

To increase the robustness of the learned clip-level features, we treat the frames of one clip as a set, and randomly mask some frames in clip-level encoding. For a given clip  $C$ , we

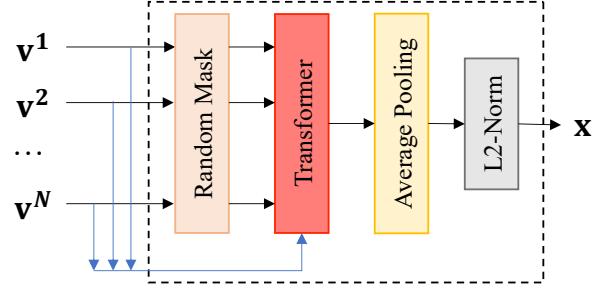


Figure 5. Architecture of our clip-level set transformer network.

randomly drop some frames to generate a new clip  $C'$ . Its goal is to eliminate the influence of frame blur or clip cut, and drive the model to have the ability that use any combination of any frames in the clip can retrieval its corresponding clips.

Specially, given a clip-level feature  $x$ , its new feature after conducting masked frame modeling is denoted as  $x'$ . Similarly, the corresponding positive clip-level feature and its new feature are denoted as  $x_+$  and  $x'_+$ . Therefore, we need to learn from the following loss functions:  $L_C(x, x'_+)$  and  $L_C(x', x_+)$ . So the final loss function of clip-level set transformer network is defined as follows:

$$L_C = L_C(x, x_+) + L_C(x, x'_+) + L_C(x', x_+) \quad (6)$$

### 3.3. Video Similarity Calculation

For each video, we first conduct shot boundary detection to segment the videos into shots, and then divide the shots into clips at a fixed time interval, i.e  $N$  seconds. Second, the sequence of successive frames is passed through the clip-level set transformer network to generate the clip-level feature. Finally, the clip-level feature is binarized by IsoHash[34] to further reduce the storage cost and search cost. When retrieving, we measure the clip-to-clip similarities with hamming distance. Given an  $M \times N$  clip-to-clip similarity matrix, the video similarity score can be calculated as follows:

$$Sim = \frac{1}{M} \sum_{i=1}^M \max_{j \in [1, N]} CS(i, j) \quad (7)$$

where  $CS(i, j)$  denotes the similarity score between clip  $i$  and  $j$ , and it is calculated as follows:

$$CS(i, j) = \max_{k \in K} \mathcal{H}(i, k) - \mathcal{H}(i, j) \quad (8)$$

in which  $K$  indicates the entire clip set and  $\mathcal{H}(\cdot, \cdot)$  indicates the hamming distance calculation.

## 4. Experiments

### 4.1. Datasets

Our SVRTN approach is trained on our constructed Self-Transformation dataset, and performs evaluations on two challenging video retrieval datasets, namely FIVR-200K and SVD, which focus on fine-grained incident video retrieval and near-duplicate video retrieval respectively. The detailed information is introduced as follows:

- **Self-Transformation** is constructed by collecting videos from video website<sup>1</sup>. It consists of 3,000 hours' videos, and temporal and spatial transformations are performed at training stage.
- **FIVR-200K** [2] consists of 225,960 videos and 100 queries. It is constructed for fine-grained incident video retrieval, including three retrieval tasks: (1) Duplicate scene video retrieval (DSVR) is to retrieval the videos sharing at least one scene that captured by the same camera, regardless of any transformation. (2) Complementary scene video retrieval (CSVR) is to retrieval the videos containing part of the same spatio-temporal segment with different views. (3) Incident scene video retrieval (ISVR) is to retrieval the videos capturing the same event without the same overlapped spatio-temporal segment. We evaluate our SVRTN approach on all the three tasks to verify its effectiveness.
- **SVD** [1] is constructed for short video retrieval task. It consists of 562,013 short videos with the duration less than 60 seconds. It contains 1,206 query videos, and over 30,000 labelled videos in which the negatives have extremely similar but different appearance. Besides, there are more than 500,000 hard negative unlabelled distraction videos to increase the retrieval difficulty.

### 4.2. Evaluation Metric

Following [1, 2], we apply the mean average precision (mAP) score to evaluate the video retrieval performance. We first calculate average precision (AP) score for each query, and then calculate their mean value as mAP score.

### 4.3. Effectiveness of Reducing Storage and Search Cost

To verify the effectiveness of our proposed SVRTN approach on reducing the storage and search cost, we compare the storage spaces and search complexities between frame-level retrieval and clip-level retrieval on SVD dataset. As shown in Table 1, the storage of the frame-level features cost 1720.32 MB, while clip-level features only cost 366.98

<sup>1</sup><https://www.youku.com/>

MB, reducing the storage cost by 78.7%. It is mainly because that our SVRTN approach first segments the videos into shots, and then divides the shots into clips, finally encodes the clips to represent the videos.

Suppose that the SVD dataset has  $m$  queries and  $n$  videos, and they are encoded by  $M$  frame-level features and  $N$  frame-level features respectively. So it needs  $\mathcal{O}(M \times N)$  similarity computation. However, depend on the above analyses, they can be encoded by  $\sim \frac{5}{M}$  and  $\sim \frac{5}{N}$  clip-level features respectively, so only  $\mathcal{O}(\sim \frac{M}{5} \times \sim \frac{N}{5}) = \sim \frac{1}{25} \mathcal{O}(M \times N)$  similarity computation is needed. In other words, our SVRTN approach increases the retrieval speed by  $\sim 25$  times, which verifies that clip-level video retrieval is an efficient retrieval paradigm to reduce the storage cost and search cost.

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

Table 1. Reduction of storage and search cost on SVD dataset.

### 4.4. Comparisons with State-of-the-art Methods

In this subsection, experimental results and analyses of comparing our proposed SVRTN approach with the state-of-the-art methods on FIVR-200K and SVD datasets are presented, which are shown in Table 2 and Table 3. It is noted that we evaluate our SVRTN approach on all the three tasks of FIVR-200K dataset, including DSVR, CSVR, ISVR.

#### 4.4.1 Comparisons with Frame-level Retrieval Methods

We then compare our SVRTN approach with 5 frame-level retrieval methods, which are briefly introduced as follows:

- CNN-L and CNN-V [5] are proposed to convert multiple intermediate CNN features into one vector via layer and vector aggregation schemes respectively.
- PPT [13] is a spatio-temporal pattern-based method under the hierarchical filter-and-refine framework.
- Temporal Network (TN) [7] is proposed to detect the longest shared path between two videos.
- VisiL [4] is proposed to calculate video-to-video similarity from refined frame-to-frame similarity matrices.

Compare with frame-level retrieval approach, our SVRTN approach outperforms all state-of-the-art methods except VisiL. It is noted that VisiL adopts a region-aligned matching scheme, which is impractical for large-scale retrieval task due to its low efficiency. While our SVRTN approach still achieves comparable retrieval performance with

Feature	Methods	Feature Dim/#bits	DSVR	CSVR	ISVR
Video-level	HC[35]	-	0.265	0.247	0.193
	DML[3]	500D	0.398	0.378	0.309
	TCA[13]	1024D	0.570	0.553	0.473
Frame-level	CNN-L[5]	4096D	0.710	0.675	0.572
	PPT[6]	4096D	0.775	0.740	0.632
	TN[7]	-	0.724	0.699	0.589
	VisiL[4]	9x3840D	0.892	0.841	0.702
	SVRTN <sub>f</sub>	512 bits	0.900	0.858	0.709
Clip-level	SVRTN	512 bits	0.876	0.835	0.686

Table 2. Comparisons with state-of-the-art methods on all three tasks of FIVR-200K dataset.

Feature	Methods	Feature Dim/#bits	Top-100 mAP
Video-level	DML[3]	500D	0.813
Frame-level	CNN-L[5]	4096D	0.610
	CNN-V[5]	4096D	0.251
	SVRTN <sub>f</sub>	512 bits	0.871
Clip-level	SVRTN	512 bits	0.860

Table 3. Comparisons with state-of-the-art methods on SVD dataset.

VisiL under the situation that only using binary codes and no any re-ranking process. Furthermore, when encoding the videos in frame-level features, our SVRTN<sub>f</sub> approach can achieve better retrieval performance than VisiL without any complex calculation. It is mainly because self-supervised video representation learning can boost the discrimination of the features due to the self-generation of training data, which has strong power in representation learning. Importantly, no annotated video pairs are needed, which efficiently reduces the expensive cost of manual annotation.

#### 4.4.2 Comparisons with Video-level Retrieval Methods

We first compare our SVRTN approach with 3 video-level retrieval methods, which are briefly introduced as follows:

- Hashing Codes (HC) [35] is proposed to learn a group of hash functions based on frame-level features, and then combine the hash codes into a single video vector.
- Deep Metric Learning (DML) [3] is proposed to early or late fuse the frame-level features into a single video vector, which is then fine-tuned by deep metric learning.
- Temporal Context Aggregation (TCA) [13] is proposed to learn a single video vector by aggregating frame-level features with self-attention.

Compare with video-level retrieval methods, clip-level retrieval methods needs more storage and search cost. To reduce these costs, we utilize hash codes and measure

hamming distances while other methods use floats and measure Euclidean or Cosine distances. Even so, our SVRTN approach achieves significant improvements by 30.6%, 28.2%, 21.3% mAPs on the DSVR, CSVR and ISVR tasks of FIVR-200K dataset, as well as 4.7% mAP on SVD dataset, which are shown in Table 2 and Table 3. It are mainly because: (1) Clip-level feature encoding can extract more abundant and complementary information from the interactions among clip frames. (2) Clip-level set transformer network can aggregate the frame features in one clip considering their different roles, which takes full advantage of each frame’s discrimination, and eliminates the redundancy between the adjacent frames. Besides, it acquires the frame permutation and missing invariant ability with masked frame modeling.

Methods	Transformations			DSVR	CSVR	ISVR
	PT	GT	ET			
SVRTN <sub>f</sub>	✓	✓	✓	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

Table 4. Impacts of different transformations on FIVR-200K dataset.

#### 4.5. Ablation Study

##### 4.5.1 Effectiveness of Self-supervised Video Representation Learning

We directly utilize the frame-level feature captured from self-supervised video representation learning to perform video retrieval, results are shown in Table 2 and Table 3 as “SVRTN<sub>f</sub>”. It outperforms than state-of-the-art methods on both two datasets, which verifies the effectiveness of the self-supervised video representation learning. Due to self-generation of training data, we can generate the training data as much as we want, which breaks the restriction of the expensive manual annotation cost.

Besides, we further evaluate the impact of each transformation on the retrieval performance of self-supervised

Methods	SVD	FIVR-200K		
		DSVR	CSV	ISVR
CE	0.854	0.870	0.834	<b>0.687</b>
<b>CE w/ MFM</b>	<b>0.860</b>	<b>0.876</b>	<b>0.835</b>	0.686

Table 5. Effectiveness of clip-level encoding with masked frame modeling.

video representation learning. The results of three tasks on FIVR-200K dataset are shown in Table 4, where “PT”, “GT” and “ET” denote photometric transformation, geometric transformation and editing transformation respectively, as well as the experiments of “A”, “B” and “C” denote training without editing transformation, geometric transformation and photometric transformation respectively. We can observe that “SVRTN<sub>f</sub>” with all the three types of transformations achieves the best performance, which verifies that each transformation plays an irreplaceable role on video representation learning. They provide rich supervision information to drive the model to approximate the real data distribution, which make the learned representation spatial-temporal invariant.

Clip Length	SVD	FIVR-200K		
		DSVR	CSV	ISVR
4s	0.861	0.883	0.841	0.693
6s	0.867	0.881	0.837	0.688
8s	0.860	0.876	0.835	0.686

Table 6. Impact of clip length on clip-level encoding.

#### 4.5.2 Effectiveness of Clip-level Set Transformer Network

We first evaluate the effectiveness of clip-level encoding with masked frame modeling on FIVR-200K and SVD datasets. Results are shown in Table 5, where “CE” and “MFM” denote clip-level encoding and masked frame modeling respectively. Clip-level encoding with masked frame modeling coerces the Transformer to learn the correlations between the anchor clip with missing information and positive clip, as well as the anchor clip and positive clip with missing information, which makes the transformer more robust and not sensitive to the frame missing. So it improves the discrimination and robustness of the learned clip-level feature, and achieves better performance.

Besides, we evaluate the impact of clip length to the retrieval performance of clip-level set transformer network. Table 6 shows the results of different clip length settings on FIVR-200K and SVD datasets. We can observe that our clip-level set transformer network is not very sensitive to the clip lengths. It is mainly because that we apply masked frame modeling in clip-level encoding, which drives the model to have the ability that any combination of

any frames in the clip can retrieval its corresponding clips. So to balance the retrieval accuracy and efficiency, we set the clip length as 8s in our experiments.

#### 4.6. Exploration of Flexible Retrieval Manners

As mentioned above, clip-level set transformer network provides more flexible retrieval manners, i.e. clip-to-clip retrieval and frame-to-clip retrieval. So we explore their retrieval performance on SVD dataset, as shown in Table 7. The videos in database are all encoded in clip-level features, only different in query encoding. We can see that 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, which can driven our SVRTN approach learn both clip-level and frame-level features. With more flexible retrieval manners, our SVRTN approach has more application prospects.

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

Table 7. Results of different retrieval manners.

### 5. Conclusion

This paper proposes the SVRTN approach to encode the video in clip-level representation with self-supervised learning to reduce the expensive cost of manual annotation, storage space and similarity search . It consists of two components: (1) Self-supervised video representation learning is proposed to automatically generate the pairs of the videos and their transformations as supervision information, which reduces the heavy labor consumption in annotating. Besides, with more self-generated supervised data, the discrimination and robustness of the learned feature are increased. (2) Clip-level set transformer network is proposed to reduce the redundancy of the frames in a clip, as well as learn the complementary and discriminative information from the interactions among clip frames. Besides, clip-level encoding with masked frame modeling make the model frame permutation and missing invariant, and support more flexible retrieval manners. Comprehensive experiments on two challenging video retrieval datasets, namely SVD and FIVR-200K, verify the effectiveness of our SVRTN approach, which achieves the best performance of video retrieval on accuracy and efficiency.

The future works will lie in two aspects: (1) How to design more efficient self-supervised learning? (2) How to transfer the motion information to RGB frames in training, but only use RGB frames in retrieval. Both of them will be explored to further improve the video retrieval performance.

## References

- [1] Qing-Yuan Jiang, Yi He, Gen Li, Jian Lin, Lei Li, and Wu-Jun Li. Svd: A large-scale short video dataset for near-duplicate video retrieval. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5281–5289, 2019. [1](#), [6](#)
- [2] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Ioannis Kompatsiaris. Fivr: Fine-grained incident video retrieval. *IEEE Transactions on Multimedia (TMM)*, 21(10):2638–2652, 2019. [1](#), [6](#)
- [3] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Yiannis Kompatsiaris. Near-duplicate video retrieval with deep metric learning. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 347–356, 2017. [1](#), [2](#), [7](#)
- [4] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Ioannis Kompatsiaris. Visil: Fine-grained spatio-temporal video similarity learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6351–6360, 2019. [1](#), [2](#), [6](#), [7](#)
- [5] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Yiannis Kompatsiaris. Near-duplicate video retrieval by aggregating intermediate cnn layers. In *International Conference on Multimedia Modeling (MMM)*, pages 251–263. Springer, 2017. [1](#), [2](#), [6](#), [7](#)
- [6] Chien-Li Chou, Hua-Tsung Chen, and Suh-Yin Lee. Pattern-based near-duplicate video retrieval and localization on web-scale videos. *IEEE Transactions on Multimedia (TMM)*, 17(3):382–395, 2015. [1](#), [2](#)
- [7] Hung-Khoon Tan, Chong-Wah Ngo, Richard Hong, and Tat-Seng Chua. Scalable detection of partial near-duplicate videos by visual-temporal consistency. In *Proceedings of the 17th ACM International Conference on Multimedia (ACM MM)*, pages 145–154, 2009. [1](#), [2](#), [6](#), [7](#)
- [8] Hao Liu, Qingjie Zhao, Hao Wang, Peng Lv, and Yanming Chen. An image-based near-duplicate video retrieval and localization using improved edit distance. *Multimedia Tools and Applications (MTA)*, 76(22):24435–24456, 2017. [1](#)
- [9] Yu-Gang Jiang and Jiajun Wang. Partial copy detection in videos: A benchmark and an evaluation of popular methods. *IEEE Transactions on Big Data (TBD)*, 2(1):32–42, 2016. [1](#)
- [10] Yaocong Hu and Xiaobo Lu. Learning spatial-temporal features for video copy detection by the combination of cnn and rnn. *Journal of Visual Communication and Image Representation (JVCIR)*, 55:21–29, 2018. [1](#), [2](#)
- [11] Yu-Gang Jiang, Yudong Jiang, and Jiajun Wang. Vcdb: a large-scale database for partial copy detection in videos. In *European Conference on Computer Vision (ECCV)*, pages 357–371. Springer, 2014. [1](#), [2](#)
- [12] Matthijs Douze, Hervé Jégou, and Cordelia Schmid. An image-based approach to video copy detection with spatio-temporal post-filtering. *IEEE Transactions on Multimedia (TMM)*, 12(4):257–266, 2010. [1](#)
- [13] Jie Shao, Xin Wen, Bingchen Zhao, and Xiangyang Xue. Temporal context aggregation for video retrieval with contrastive learning. *arXiv preprint arXiv:2008.01334*, 2020. [1](#), [6](#), [7](#)
- [14] Jingkuan Song, Yi Yang, Zi Huang, Heng Tao Shen, and Richang Hong. Multiple feature hashing for real-time large scale near-duplicate video retrieval. In *Proceedings of the 19th ACM International Conference on Multimedia (ACM MM)*, pages 423–432, 2011. [1](#)
- [15] Yang Feng, Lin Ma, Wei Liu, Tong Zhang, and Jiebo Luo. Video re-localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 51–66, 2018. [2](#)
- [16] Lorenzo Baraldi, Matthijs Douze, Rita Cucchiara, and Hervé Jégou. Lamv: Learning to align and match videos with kernelized temporal layers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7804–7813, 2018. [2](#)
- [17] Jérôme Revaud, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Event retrieval in large video collections with circulant temporal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2459–2466, 2013. [2](#)
- [18] Kaiyang Liao, Hao Lei, Yuanlin Zheng, Guangfeng Lin, Congjun Cao, Mingzhu Zhang, and Jie Ding. Ir feature embedded bof indexing method for near-duplicate video retrieval. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 29(12):3743–3753, 2018. [2](#)
- [19] Yang Cai, Linjun Yang, Wei Ping, Fei Wang, Tao Mei, Xian-Sheng Hua, and Shipeng Li. Million-scale near-duplicate video retrieval system. In *Proceedings of the 19th ACM International Conference on Multimedia (ACM MM)*, pages 837–838, 2011. [2](#)
- [20] Zhanning Gao, Gang Hua, Dongqing Zhang, Nebojsa Jojic, Le Wang, Jianru Xue, and Nanning Zheng. Er3: A unified framework for event retrieval, recognition and recounting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2253–2262, 2017. [2](#)
- [21] Xiao Wu, Alexander G Hauptmann, and Chong-Wah Ngo. Practical elimination of near-duplicates from web video search. In *Proceedings of the 15th ACM International Conference on Multimedia (ACM MM)*, pages 218–227, 2007. [2](#)
- [22] Venice Erin Liong, Jiwen Lu, Yap-Peng Tan, and Jie Zhou. Deep video hashing. *IEEE Transactions on Multimedia (TMM)*, 19(6):1209–1219, 2016. [2](#)
- [23] Shuyan Li, Zhixiang Chen, Jiwen Lu, Xiu Li, and Jie Zhou. Neighborhood preserving hashing for scalable video retrieval. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 8212–8221, 2019. [2](#)
- [24] Yanbin Hao, Tingting Mu, Richang Hong, Meng Wang, Ning An, and John Y Goulermas. Stochastic multiview hashing for large-scale near-duplicate video retrieval. *IEEE Transactions on Multimedia (TMM)*, 19(1):1–14, 2016. [2](#)

- [25] Jingkuan Song, Hanwang Zhang, Xiangpeng Li, Lianli Gao, Meng Wang, and Richang Hong. Self-supervised video hashing with hierarchical binary auto-encoder. *IEEE Transactions on Image Processing (TIP)*, 27(7):3210–3221, 2018. 2
- [26] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 843–852, 2017. 3
- [27] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 3
- [28] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1920–1929, 2019. 3
- [29] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 766–774, 2014. 3
- [30] 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 (CVPR)*, pages 770–778, 2016. 4
- [31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017. 5
- [33] Yujie Zhong, Relja Arandjelovic, and Andrew Zisserman. Compact deep aggregation for set retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 5
- [34] Weihao Kong and Wu-Jun Li. Isotropic hashing. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1646–1654, 2012. 5
- [35] Jingkuan Song, Yi Yang, Zi Huang, Heng Tao Shen, and Jiebo Luo. Effective multiple feature hashing for large-scale near-duplicate video retrieval. *IEEE Transactions on Multimedia (TMM)*, 15(8):1997–2008, 2013. 7