TransVCL: Attention-Enhanced Video Copy Localization Network with Flexible Supervision

Sifeng He[1*], Yue He[1], Minlong Lu[1], Chen Jiang[1], Xudong Yang[1], Feng Qian[1*], Xiaobo Zhang[1], Lei Yang[1], Jiandong Zhang[2]

[1]Ant Group
[2] Copyright Protection Center of China
{sifeng.hsf, youzhi.qf}@antgroup.com

Abstract

Video copy localization aims to precisely localize all the copied segments within a pair of untrimmed videos in video retrieval applications. Previous methods typically start from frame-to-frame similarity matrix generated by cosine similarity between frame-level features of the input video pair, and then detect and refine the boundaries of copied segments on similarity matrix under temporal constraints. In this paper, we propose TransVCL: an attention-enhanced video copy localization network, which is optimized directly from initial frame-level features and trained end-to-end with three main components: a customized Transformer for feature enhancement, a correlation and softmax layer for similarity matrix generation, and a temporal alignment module for copied segments localization. In contrast to previous methods demanding the handcrafted similarity matrix, TransVCL incorporates long-range temporal information between feature sequence pair using self- and cross- attention layers. With the joint design and optimization of three components, the similarity matrix can be learned to present more discriminative copied patterns, leading to significant improvements over previous methods on segment-level labeled datasets (VCSL and VCDB). Besides the state-of-the-art performance in fully supervised setting, the attention architecture facilitates TransVCL to further exploit unlabeled or simply video-level labeled data. Additional experiments of supplementing video-level labeled datasets including SVD and FIVR reveal the high flexibility of TransVCL from full supervision to semi-supervision (with or without video-level annotation). Code is publicly available at https://github.com/transvcl/TransVCL.

Introduction

Due to the explosive growth of pirated multimedia and increasing demands for preventing copyright infringements in recent years, content-based video retrieval (CBVR) becomes increasingly essential in applications such as copyright protection, video filtering, recommendation, etc. Besides overall video-level copy detection for given pairs of potential copied videos, finer-grained segment-level copy details with temporal boundaries are also desired particularly in large datasets or real-world scenarios so that applications such as copyright protection become easier and more intuitive. Figure 1 shows this segment-level video copy detection task.

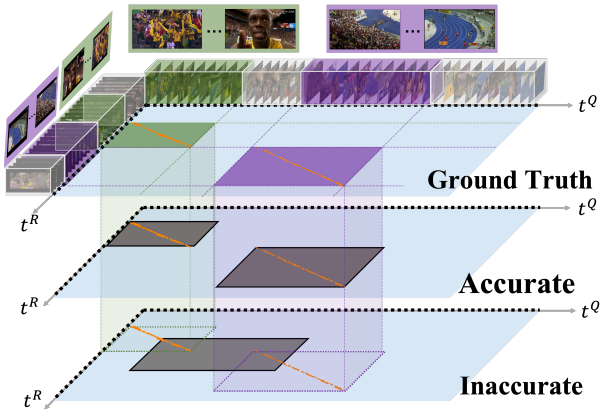

Figure 1: Video copy localization is to accurately detect the start and end timestamps of all copied segment pairs in the query and reference videos. The yellow highlight points on similarity matrix indicate high similarity scores of temporal corresponding frames between query and reference videos.

The green or purple bounding box corresponding to trimmed temporal section of two original videos (query and reference video) is a pair of copied segments. The goal of video copy localization task is to search and obtain all these copied segments pairs from query and reference videos as consistent as possible to ground-truth labels.

Many recent works (Baraldi et al. 2018; Jiang et al. 2021; Han et al. 2021; He et al. 2022) have been proposed to precisely localize copied segments within video pairs. (He et al. 2022) concludes a detailed benchmark pipeline for video copy localization. The pipeline starts with a pair of copied videos as input and then outputs the predicted copied segments. All the benchmarked copy localization algorithms follow the scheme that generates a handcrafted frame-to-frame similarity matrix (usually cosine similarity between frame-level feature sequence pairs), and then feed this correlation matrix to temporal alignment methods including TN (Tan et al. 2009), DP (Chou, Chen, and Lee 2015), SPD (Jiang et al. 2021) etc. This pipeline has good generalizability and extensibility, and it is suitable for most basic copy transformations.

However, this pipeline also has several limitations. First,

the frames of a video are commonly processed as individual images, making the modeling of long-range semantic dependencies difficult. In addition to the disregard for spatial-temporal relationships within each video or between the query and reference videos, the steps of this pipeline are separated and independent. Hence, the input of temporal alignment modules is the similarity correlation which might be severely compressed from initial feature pair. This leads to only local optimization in each module, thus limiting video copy localization performance.

Another important observation is that the data-driven neural networks have significantly outperformed most of conventional methods on video copy detection task (He et al. 2022; Kordopatis-Zilos et al. 2019b; Jiang et al. 2021). However, the assumption of having a sufficient amount of accurate labeled data for training may not hold, especially for video copy localization task which needs annotations with detailed copy starting and ending timestamps. Thus, a natural idea is to leverage abundant unlabeled data or only video-level labeled data to facilitate learning in the original localization task. This requires a large-capacity model that is adaptable to flexible manner of annotations.

To address the above-mentioned issues, we propose TransVCL, a novel approach to localize copied segments in videos. Motivated by the effectiveness of attention mechanism in capturing temporal dependencies, we propose to incorporate temporal information between frame-level features using Transformer (Vaswani et al. 2017) with self- and cross- attention layers to reason about the internal and mutual relationships between video descriptors, and the similarity matrices are then generated from a differentiable softmax matching layer. After that, the similarity matrices are fed to a temporal alignment network that detects the similarity patterns with detailed starting and ending timestamps under the supervision of labeled copied segments. Here, the components in our framework including feature enhancement, similarity generation and temporal alignment modules are jointly trained to obtain an end-to-end optimization.

Meanwhile, the above all revolve around segment-level copy detection on fine-grained frame correlations, which might lose awareness of global video-level representation. Attributed to the architecture of Transformer, we simply add an extra learnable token to the sequence, and this token is to aggregate information from the entire video sequence as a global representation. Hence, we supplement an auxiliary video-level binary classification task to further exploit video-level prediction from the relationship between query and reference video representation. This supplementary global loss is proved to facilitate video copy localization well, while also making better use of video-level labeled data. This enables our algorithm to be more flexibly extended to semi-supervised or weakly supervised learning.

We evaluate our proposed method on two current segment-level annotated video copy dataset VCSL (He et al. 2022) and VCDB (Jiang, Jiang, and Wang 2014). The experiments show that TransVCL outperforms other algorithms by a large margin with much more accurate copied segment localization. Detailed visualization reveals that TransVCL automatically learns to optimize the correlation similarity distribution to a cleaner pattern, and then localize the copied segments on the optimized noiseless similarity matrix. Moreover, we test TransVCL in semi-supervised setting with using small fraction of labeled dataset and the remainder as unlabeled sets or weakly labeled sets, and TransVCL shows significant accuracy improvements while adding unlabeled or weakly labeled data. At last, we also utilize publicly available datasets (SVD (Jiang et al. 2019), FIVR (Kordopatis-Zilos et al. 2019a)) with only video-level labels to further improve video copy localization performance, and this simulates the real-world applications with continuous streamed unlabeled or weakly labeled data.

Related Work

Video Copy Localization

Video copy localization, i.e., segment-level video copy detection, is applied to a pair of potential copied videos, including a query video $V^Q = \{v_m^Q\}_{m=1}^q$ and a reference video $V^R = \{v_n^R\}_{n=1}^r$, where v_m^Q and v_n^R are m-th frame in the query video and n-th frame in the reference video respectively, and q and r are the numbers of frames in these two videos. There might exist one or more copied segments between V^Q and V^R. The objective is to seek the accurate corresponding temporal copied segments $\{v_m^Q\}_{m=t_{s_1}^Q}^{t_{e_1}^Q}$, $\{v_m^Q\}_{m=t_{s_2}^Q}^{t_{e_2}^Q}$, ... in V^Q and $\{v_n^R\}_{n=t_{s_1}^R}^{t_{e_1}^R}$, $\{v_n^R\}_{n=t_{s_2}^R}^{t_{e_2}^R}$, ... in V^R, where t_{s_i} and t_{e_i} are the start and end timestamps and i is the index of multiple copied segments. The frames between $t_{s_i}^Q$ and $t_{e_i}^Q$ in V^Q are the transformed copies of the frames between $t_{s_i}^R$ and $t_{e_i}^R$ in V^R. In the segment-level labeled video copy dataset (e.g. VCDB (Jiang, Jiang, and Wang 2014), VCSL (He et al. 2022)), the detailed copied segment information $[\{t_{s_1}^Q, t_{e_1}^Q, t_{s_1}^R, t_{e_1}^R\}, \{t_{s_2}^Q, t_{e_2}^Q, t_{s_2}^R, t_{e_2}^R\}$, ...] are provided for full supervision of copy localization. There also exist some video retrieval datasets (e.g. SVD (Jiang et al. 2019), FIVR (Kordopatis-Zilos et al. 2019a)) that only indicate video-level copy information without detailed copied segment localization annotation.

Due to the temporal accuracy requirements on copied segment boundaries, copy localization methods always start with frame-level features rather than clip-level features. Meanwhile, for a fair comparison between copy localization algorithms, the frame-level features tend to be fixed. Then a common approach is to generate a frame-to-frame similarity matrix by taking into account of continuous frame sequences, as shown in Figure 1. In order to obtain the detailed copied localization, a simple method is to vote temporally by temporal Hough Voting (HV) (Douze, Jegou, and Schmid 2010). The graph-based Temporal Network (TN) (Tan et al. 2009) 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. Another method is Dynamic Programming (DP) (Chou, Chen, and Lee 2015), whose goal is to find a diagonal block with the largest similarity. Inspired by temporal matching kernel (Poullot et al. 2015), LAMV(Baraldi et al. 2018) transforms the kernel into a differentiable layer to

conduct temporal alignment. SPD (Jiang et al. 2021) formulates temporal alignment as an object detection task based on the frame-to-frame similarity matrix, achieving a state-of-the-art segment-level copy localization performance.

Transformer for Videos

Transformer (Vaswani et al. 2017) has become the de facto standard for sequence modeling in natural language processing (NLP) due to its simplicity and computation efficiency. Recently, Transformers are also getting more attention in basic computer vision tasks, such as image classification (Dosovitskiy et al. 2020), object detection (Carion et al. 2020) and semantic segmentation (Strudel et al. 2021). As for video task, Transformers have also been applied to video retrieval (Shao et al. 2021), video super resolution (Liu et al. 2022), video action recognition (Yang et al. 2022), etc. Inspired by the success of Transformers in many vision tasks, we first explore the Transformer architecture for video copy localization task.

Semi-supervised and Weakly-supervised Learning

Semi-supervised learning (SSL) exploits the potential of unlabeled data to facilitate model learning with limited annotated data. Most of the existing SSL methods focus on image classification (Hu et al. 2021; Taherkhani et al. 2021) and object detection (Sohn et al. 2020; Zhou et al. 2021) task. On the other hand, the weakly supervised setting, where only global-level category labels are required during training, has drawn increasing attention in image segmentation (Lee et al. 2021a; Zhou et al. 2022) and video action localization (Lee et al. 2021b; Shen and Elhamifar 2022). In this paper, we introduce these two simple yet effective settings to video copy localization for the first time. Owing to the good adaptability of TransVCL, our approach can be conveniently extended to flexible supervision without requiring complex design. We believe that the specific network designed only for semi-supervised or weakly supervised video copy localization will also attract increasing attention in the future, but we will not discussed in this paper.

TransVCL Details

In this section, we introduce the framework details of TransVCL. An overview is presented in Figure 2, consisting of feature enhancement, similarity generation and segment localization modules. Another advantage of flexible extension to semi-supervision, which shows great generalizability of our proposed network, will also be introduced.

Feature Enhancement with Attention

The key to our formulation lies in generating high-quality similarity relationship between the input frame-level feature pairs for next-step alignment. Recall that the frame-level features are extracted independently from a fixed image embedding network. We denote the frame-level feature pair as $F^Q = \left\{f_m^Q\right\}_{m=1}^q$ from query video V^Q and $F^R = \left\{f_n^R\right\}_{n=1}^r$ from reference video V^R, and f_m^Q is m-th frame-level feature of query video and f_n^R is n-th frame-level feature of reference video. To further consider the inter- and intra- dependencies between these features, a natural choice is Transformer, which is particularly suitable for modeling the mutual relationship within or between F^Q and F^R with the attention mechanism. Since the frame-level features have no notion of the temporal information, we first add the fixed sine and cosine temporal encoding to the initial features, which is consistent with positional encoding in (Vaswani et al. 2017). The sum of initial frame-level feature and temporal encoding also makes the next-step matching process consider not only the feature similarity but also their temporal distance.

Inspired from the leverage of class token in some previous works, like BERT (Devlin et al. 2018),ViT (Dosovitskiy et al. 2020), we prepend a learnable embedding to the sequence of frame-level features on each video, whose value in convergent state serves as the global feature of the given video. Concretely, the input features to our attention modules are presented as:

$$\begin{aligned} F^Q_{\text{input}} &= [f^Q_{\text{class}}; f^Q_0; f^Q_1; ...] + f_{\text{tem}} \\ F^R_{\text{input}} &= [f^R_{\text{class}}; f^R_0; f^R_1; ...] + f_{\text{tem}} \end{aligned} \tag{1}$$

where f_{class} is the learnable class token embedding served as a global video feature, and f_{tem} is the temporal encoding.

Then we perform stacked self- and cross- attention modules to enhance the initial input features. Specifically, for self-attention module, which considers only the internal relationship inside the frame sequence of a single video, the query, key and value in the attention mechanism are from the same feature sequence, either F^Q or F^R. For cross-attention module, the key and value are same (e.g., F^Q) but different in the query (e.g., F^R) to introduce their mutual dependencies. This process is performed for both F^Q and F^R symmetrically, which is indicated in Eq. 2. The formulas of our attention modules are illustrated as below,

$$\begin{aligned} F^Q_{\text{output}} &= \mathfrak{T}(F^Q_{\text{input}}, F^R_{\text{input}}) \\ F^R_{\text{output}} &= \mathfrak{T}(F^R_{\text{input}}, F^Q_{\text{input}}) \end{aligned} \tag{2}$$

where

$$\begin{aligned} \mathfrak{T} &= [\mathfrak{T}_{\text{self}}, \mathfrak{T}_{\text{cross}}] \\ \mathfrak{T}_{\text{self}}(x, \cdot) &= \text{MHA}(x, x, x) \\ \mathfrak{T}_{\text{cross}}(x, y) &= \text{MHA}(x, y, y) \end{aligned} \tag{3}$$

and $\mathfrak{T}$ is Transformer, MHA is multi-head attention mechanism with the same query, key, value as input in Transformer but with different focus aspects.

After attention-based feature enhancement, we can additionally obtain the video-level feature representation for either query or reference videos which is the first token in F^Q_{output} or F^R_{output}. Similar with Next Sentence Prediction in Bert (Devlin et al. 2018), a binary classification head is attached for further utilization of weak labels (i.e., video-level copy annotations), as well as negative labeled information for uncopied video pairs. Our classification head is implemented by a MLP with one hidden layer with the concatenation of both class tokens:

$$y = \text{MLP}([F^Q_{\text{output}}[0]; F^R_{\text{output}}[0]]) \tag{4}$$

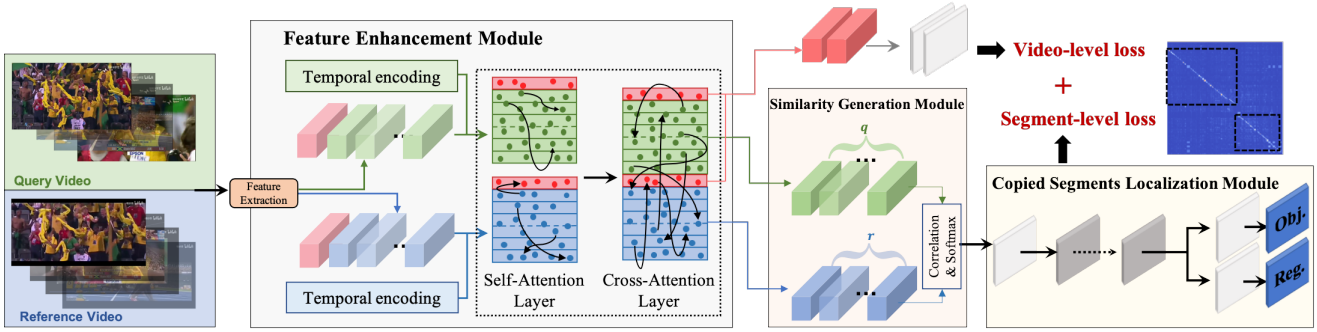

Figure 2: Overview of TransVCL, which is joint trained with three modules including feature enhancement, similarity generation and segment localization. We supplement a video-level loss to facilitate copy localization and further exploit weak labels.

where y is the predictive probability of F^Q and F^R being copied in video-level representation. This will be an auxiliary task for copied segments localization (the third modules in TransVCL).

To pursue better efficiency, we propose to use Linear Transformer (Katharopoulos et al. 2020) instead of standard Transformer architecture, an efficient variant of the vanilla attention layer in Transformer, whose complexity is reduced from $O(n^2)$ to $O(n)$, where n is sequence length. Details of linear attention are attached in Supplementary Material.

Similarity Matrix Generation

The frame-to-frame similarity matrix based on frame-level features is capable of representing similarity relationship between the copied videos and it has proved to be feasible to search and localize the copied segments (He et al. 2022). However, aforementioned matching step with generating a similarity map is handcrafted without optimization in learning process. Inspired from local feature matching network (Rocco et al. 2018; Sun et al. 2021), we apply the differentiable matching layers with a dual-softmax operator to similarity generation. In detail, the similarity matrix is first calculated by correlation expression $S_{mn} = \frac{1}{\tau} \cdot \left\langle \tilde{f}_m^Q, \tilde{f}_n^R \right\rangle$ with a tunable hyper-parameter τ temperature, and m and n are the temporal frame index of query and reference video respectively, $\tilde{f}$ indicates the attention-enhanced frame-level feature. Then we apply softmax on both dimensions of S to obtain probability of soft mutual nearest neighbor matching:

$$\tilde{S} = \text{softmax}(S(m,\cdot))_n \cdot \text{softmax}(S(\cdot,n))_m \quad (5)$$

Copied Segments Localization

The objective of copy localization task is to find the detailed copied segment information, i.e. $[\{t_{s_1}^Q, t_{e_1}^Q, t_{s_1}^R, t_{e_1}^R\}, \{t_{s_2}^Q, t_{e_2}^Q, t_{s_2}^R, t_{e_2}^R\}, ...]$, within the given feature pairs. Interestingly, these 4-dimensional copied segment pairs with start and end timestamps can be formulated as bounding boxes with top-left and bottom-right coordinates in similarity map $\tilde{S}$ as shown in Figure 1 and Figure 2. The temporal boundaries of ground truth segment-level labels can also be expressed as bounding box coordinates. Therefore, we treat the task of locating all the copied segment pairs as an object detection task on aforementioned similarity map $\tilde{S}$. Here, high frame similarities between copied segment pairs compose specific infringement patterns, and this can be learned by the object detection network in a data-driven manner. The training loss of segment-level supervision is defined as:

$$L_{\text{seg}}(F^Q, F^R, p^*, t^*) = \sum_i [L_{obj}(p_i, p_i^*) + \lambda L_{reg}(t_i, t_i^*)] \quad (6)$$

, where i is the index of a trimmed copied segment pair detected from input video pair. p_i is the predictive probability of the segment pair being positive (copied). t_i is the 4-dimensional coordinates of the copied segment pair. p_i^* is the binary label (copied or not) from ground-truth. t_i^* is temporal location of the ground-truth copied segment pair. We use BCE Loss for training obj branch to predict the trimmed segment pair is similar or not, and IoU Loss for training reg branch to regress their temporal boundaries (coordinates).

In addition, we utilize the obtained video-level copied prediction from Eq.4 for auxiliary binary classification loss:

$$L_{\text{video}}(F^Q, F^R, y^*) = L_{\text{cls}}(y, y^*) \quad (7)$$

, where y^* is the ground truth video-level binary label. This weak label can be easily obtained from both segment-level or video-level annotated video copy dataset. The final loss consists of the losses from both segment and video-level:

$$L = L_{\text{seg}} + L_{\text{video}} \quad (8)$$

Extension to Flexible Supervision

Due to the network design with both video-level and segment-level task taken into consideration, we can easily adapt TransVCL to semi-supervision and weak supervision. In detail, only part of the provided data of this section are fully labeled with segment-level, and the remainder of data are unlabeled or weakly labeled with only global info (only video-level annotations of copied or not).

Inspired from semi-supervised object detection task (Sohn et al. 2020), a simple framework is proposed to verify the transferability of TransVCL to semi-supervised learning shown in Figure 3. First, we train a teacher model on initial segment-level labeled video feature pairs, and this step

is exactly the same as training TransVCL in fully supervised setting above. Second, pseudo labels of unlabeled data and weakly labeled data are generated while inference on the trained teacher model. Third, we use a high threshold value for the confidence-based thresholding to control the quality of pseudo labels comprised of copied segment localization. In this step, the weak supervision with video-level info can also help us filter out the low-quality pseudo labels, and this detailed process will also improve the performance with weakly labeled data. The final step is to train TransVCL network with both labeled and pseudo-labeled data by jointly minimizing the supervised and unsupervised (or weakly supervised) loss as follows:

$$\tilde{L} = L_s(F^Q, F^R, p^*, t^*) + \lambda_u L_u(F^Q, F^R, y^*) \quad (9)$$

, where y^* is the video-level copied labels only provided in weakly supervised situation. Both L_s and L_u are consistent with Eq. 8, and p^* and t^* in L_u come from pseudo labels.

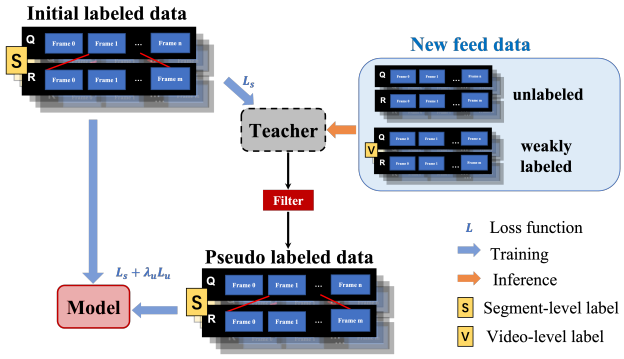

Figure 3: The semi-supervised and weakly semi-supervised setting of video copy localization task.

Experiments

In this section, we first train and evaluate TransVCL on publicly available segment-level video copy datasets with fully supervised setting. The visualization of intermediate results demonstrates the feasibility of our method. The performance of TransVCL to flexible supervision is also provided.

Datasets and Evaluation

Following the video copy segment localization benchmark (He et al. 2022), we use VCSL dataset for method comparisons. In addition, we also evaluate results on VCDB dataset with smaller scale of data but also containing ground truth copied segments. Besides these two existing datasets with detailed segment-level copied annotations, we further utilize video-level copy dataset (FIVR and SVD) as weakly labeled data for semi-supervised evaluation.

As for evaluation protocol, we take two untrimmed videos as input to detect all the potential trimmed copied segments between the two videos. We adopt the F-score, the harmonic mean of segment-level recall and precision (He et al. 2022), as the standard metric. This metric not only intuitively indicates segment-level alignment accuracy, but also reflects video-level performance. In addition, coarse-grained video-level false rejection rate (FRR) and false acceptance rate (FAR) are also discussed in ablation study.

Implementation Details

The entire TransVCL network is trained end-to-end with pairs of frame-level video feature sequences extracted from the given query and reference videos as input. Due to the unavailability of some raw video files with invalid video links in VCSL dataset, we directly use the official provided ISC features, which are also claimed as SOTA frame-level features in Facebook AI Image Similarity Challenge (Papakipos et al. 2022). For VCDB and other datasets with raw video files, we first extract one frame per second and then extract ISC features following (Yokoo 2021) like VCSL, obtaining a 256-dimension embedding for each frame. In the training stage of our experiments, all the video feature sequences are truncated with the maximum length of 1200 (i.e., 20min) and padded zeros for sequence length lower than 1200. Therefore, the input to TransVCL network is pairs of video features with uniform size of 1200×256. In the feature enhancement component, the head of Transformer is 8. In the similarity generation component, temperature τ of dual-softmax is 0.1 and the similarity matrix is reshaped to (640, 640). In copied segment localization module, we adopt the anchor-free detection network YOLOX (Ge et al. 2021) with simple design, and the regression loss weight λ is 5 as default. The entire model is trained using SGD with momentum 0.9, batch size of 64, initial learning rate of 0.01 and weight decay of 0.0005. The detailed implementation can also be inferred in our codes.

Comparison with Previous Methods

We compare our method to the previously reported copy localization algorithms in VCSL benchmark (He et al. 2022). In addition, we also evaluation these methods on VCDB dataset as shown in Table 1. All these methods are trained (if necessary) on close to 100k pairs of copied video pairs and an approximately equal number of automatically generated negative uncopied pairs between different video categories in train set of VCSL dataset. As can be shown from Table 1, TransVCL outperforms all other methods by a large margin on VCSL dataset, with F-score of 66.51% (+4.55% compared with previous SOTA SPD). On VCDB dataset, TransVCL also achieves the best F-score of 75.37%. Based on the attention-enhanced features and end-to-end optimization, our proposed method obtains the highest accuracy on copy segments localization. In detail, TransVCL achieves the best results on over 60% of the topic categories (7/11) including the most common daily life videos and the most difficult kichiku videos. Meanwhile, TransVCL shows significantly higher accuracy ($+10\%$ higher F-score than second-best) on short copied segments within videos (copy duration percentages $< 40\%$). For other topic categories and copy percentages scenarios, TransVCL mostly achieves top-2 highest performances which are very close (only $0.5 \sim 1\%$ F-score) to the highest scores. If we look into some individual hard cases, TransVCL also presents considerably better localization results on some complex video transformations

including backwards-running, acceleration or deceleration, short copied segments, mushup, etc. The detailed statistic comparison and some visualization cases are given in Supplementary Material.

Dataset	Method	Recall	Precision	**F-score**↑
VCSL	HV	86.94	36.82	51.73
	TN	75.25	51.80	61.36
	DP	49.48	60.61	54.48
	DTW	45.10	56.67	50.23
	SPD	56.49	68.60	61.96
	TransVCL	65.59	67.46	**66.51**
VCDB	HV	89.23	58.70	70.81
	TN	80.06	69.20	74.23
	DP	63.84	73.54	68.35
	DTW	61.78	72.26	66.61
	SPD	71.00	78.82	74.71
	TransVCL	76.69	74.09	**75.37**

Table 1: Video copy localization performance comparison of our proposed TransVCL with previous benchmarked approaches on VCSL and VCDB datasets.

Ablation Study

We ablate different settings of feature enhancement module on VCSL dataset in Table 2. Here, the basic localization model has the same setting as the aforementioned copied segment localization module. Compared with the basic localization model with conventional frame-to-frame cosine similarity map as input, the F-score is significantly improved after incorporating the attention module and joint optimization. Meanwhile, the class token along with the auxiliary branch of video-level classification loss introduces a broader guidance, bringing additional performance gains.

In Table 2, the global class token design as an auxiliary task is proved to facilitate the copied segment localization accuracy. To further demonstrate the feasibility of our design, we also show coarse-grained copy detection results with video-level FRR/FAR in Table 3. The video-level performance of basic localization model slightly improves compared with previous SOTA SPD algorithm, and the error rate is significantly lower after joint optimization with our designed attention module. Meanwhile, the video-level result of our method is also better than any others reported in video copy segment localization benchmark (He et al. 2022).

Method	Recall	Precision	**F-score**↑
basic localization model	60.24	64.70	62.39
+self-att&cross-att	63.96	65.54	64.74
+att+video-level loss	60.19	71.43	65.33
+att+temporal encoding	65.88	66.79	66.34
full model	65.59	67.46	66.51

Table 2: Ablation of detailed settings in TransVCL network.

Method	FRR ↓	FAR ↓	**F-score** ↑
SPD	0.2974	0.0958	79.08
basic localization model	0.2527	0.1105	81.22
full model	0.1666	0.0173	90.19

Table 3: Video-level copy detection performance gain from the attention module.

Visualization of Intermediate Results

To better understand the effect of our proposed method, we visualize self- and cross- attention weights in Figure 4(a). The self-attention represented as arcs establishes long-range semantic connections within the video, thus expanding the receptive field from individual frames or clips to almost the entire video. Meanwhile, the cross-attention enhances the correlations between the input video pair. For example, the semantic correlation of frames such as the relevant screenshot of runway in Bolt's 100-meter race is established even with long temporal interval in Figure 4(a). As a result, informative frames describing key and relevant moments of the event get higher response, and the redundant frames are suppressed, leading to significantly cleaner frame-to-frame similarity maps shown in Figure 5. This learned intermediate result reduces the difficulty of follow-up copied segment localization and obtains obvious performance gains.

Moreover, we also observe how global feature integrates information across the video sequence and further improves the video copy detection performance. The red curve in Figure 4(b) presents the attention weights between the class token and different frame features in reference video. Here, query video is the same Bolt's 100-meter race video in Figure 4(a). It is interesting that the global feature's attention learns to peak at the temporal location of video highlights, i.e., clips of 100-meter race scene. This learned mechanism also enhances the effective information and reduces noisy disturbance, bringing a more robust video copy detection accuracy. In addition, the visualization in Figure 4(b) also shows the potential of our method for extending to other video tasks such as video summarization or event retrieval.

Performance in Semi-supervised Setting

Our proposed method can also be flexibly extended to semi-supervised learning, and we experiment two settings here. The first is to randomly sample 1%, 2%, 5% and 10% of labeled training data in VCSL and use rest of training data as an unlabeled or weakly labeled set. The 1%, 2%, 5% and 10% protocol also means that the scale ratio between unlabeled (or weakly labeled) data and labeled data is 99×, 49×, 19× and 9×. Besides sampling within a single dataset, we also use the entire VCSL as a labeled set and additional video-level labeled dataset (FIVR, SVD) as a weakly labeled set. In detail, FIVR and SVD provide $\sim$ 17.8k weakly labeled copied video pairs, which are corresponding to 1.8× and 0.18× scale of 10%VCSL and 100%VCSL ($\sim$ 97.7k) respectively. This second protocol evaluates the potential to improve our SOTA video copy localization method with weakly labeled data. In addition, semi-supervision also in-

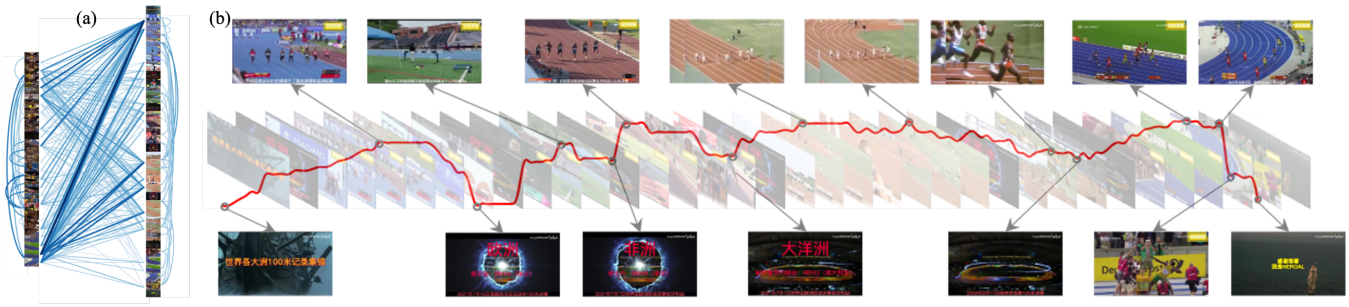

Figure 4: (a)Visualization of attention weights. The left frames are from query video and right frames are from reference video. The thickness of lines indicate strength of attention weights. Zoom in for details. (b)Attention weights about the global feature (class token) to frame feature sequence. After training, local maximum clips of attention (red line) are learned to be the highlight moments of reference videos (top row), and local minimum points are mostly unimportant transitions (bottom row).

Method	1%VCSL	2%VCSL	5%VCSL	10%VCSL	100%VCSL
Supervised on labeled data only	23.10	31.95	47.71	59.07	66.51
Semi-supervised (w/o weak label)	37.28 (+14.18)	44.80 (+12.85)	56.11 (+ 8.40)	61.98 (+2.91)	/
Semi-supervised (with weak label)	44.87 (+21.77)	50.50 (+18.55)	58.03 (+10.32)	62.60 (+3.53)	/

Table 4: Comparison in F-score for different semi-supervised settings on VCSL dataset. 1%, 2%, 5% and 10% of labeled training data are randomly sampled as a labeled set (with detailed copied segments locations) and we use the rest of labeled training data as an unlabeled or weakly labeled (only video-level copy annotation) set.

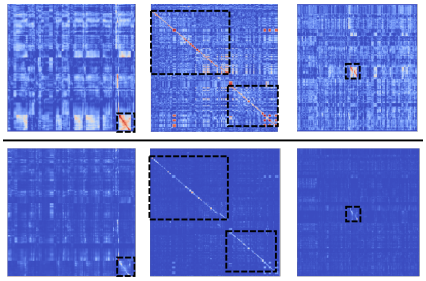

Figure 5: Learned intermediate results, i.e., similarity matrices, before (upper row) and after (bottom row) attention-based feature enhancement for example copied video pairs.

troduces another hyper-parameters λ_u as the unsupervised loss weight in Eq. 9. After studying the impact of λ_u on 10% protocol (details in Supplementary Material), the best performance is obtained when $\lambda_u = 0.5$. Two semi-supervised protocol results are summarized in Table 4 and Table 5.

For 1%, 2%, 5% and 10% protocols, we demonstrate that TransVCL can be effectively extended to semi-supervised learning with significant F-score improvements from unlabeled data, especially under high ratio between unlabeled and labeled data. Meanwhile, the semi-supervised results are further improved from unlabeled data to weakly labeled data, bringing a considerable performance gain with controllable annotation labor. This improvement is also obvious when the scale of unlabeled data is much larger than labeled data, and this is extremely suitable for online streamed data in practical use. From the cross dataset performance in Table 5, F-score is also improved and the SOTA performance is further refreshed from 66.51 to 67.13 while only adding 0.18× scale of more data with weak label. Notably, on the same 10% labeled data, FIVR & SVD brings slightly higher improvements (+3.96) than 90% VCSL (+3.53), even though the weakly labeled data of former is much less than latter (1.8× vs 9×). Our proposed network is not only suitable for weakly supervised scenarios, but also has a good capacity to the data diversity.

Method	F-score	Δ
Supervised on 10% VCSL	59.07	/
+FIVR & SVD (weak label)	63.04	+3.96
Supervised on 100% VCSL	66.51	/
+FIVR & SVD (weak label)	67.13	+0.62

Table 5: Cross dataset semi-supervised learning performance of supplementing video-level labeled datasets.

Conclusion

We have presented a novel network named TransVCL with joint optimization of multiple components for segment-level video copy detection and demonstrated its strong performance. Enhanced by attention mechanism, TransVCL incorporates temporal correlation within and across input videos and captures more accurate copied segments. Meanwhile, the flexible extension to semi-supervised and weakly semi-supervised scenarios further demonstrates the advantages of TransVCL. We hope our new perspective will pave a way towards a new paradigm for accurate and data-efficient video copy localization.

Acknowledgments

This work is partly supported by R&D Program of DCI Technology and Application Joint Laboratory.

References

Baraldi, L.; Douze, M.; Cucchiara, R.; and Jegou, H. 2018. LAMV: Learning to Align and Match Videos with Kernelized Temporal Layers. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7804–7813.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chou, C.-L.; Chen, H.-T.; and Lee, S.-Y. 2015. Pattern-based near-duplicate video retrieval and localization on web-scale videos. *IEEE Transactions on Multimedia*, 17(3): 382–395.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Douze, M.; Jegou, H.; and Schmid, C. 2010. An Image-Based Approach to Video Copy Detection With Spatio-Temporal Post-Filtering. *IEEE Transactions on Multimedia*, 12(4): 257–266.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

Han, Z.; He, X.; Tang, M.; and Lv, Y. 2021. Video Similarity and Alignment Learning on Partial Video Copy Detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4165–4173.

He, S.; Yang, X.; Jiang, C.; et al. 2022. A Large-scale Comprehensive Dataset and Copy-overlap Aware Evaluation Protocol for Segment-level Video Copy Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21086–21095.

Hu, Z.; Yang, Z.; Hu, X.; and Nevatia, R. 2021. Simple: similar pseudo label exploitation for semi-supervised classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15099–15108.

Jiang, C.; Huang, K.; He, S.; Yang, X.; Zhang, W.; Zhang, X.; Cheng, Y.; Yang, L.; Wang, Q.; Xu, F.; et al. 2021. Learning Segment Similarity and Alignment in Large-Scale Content Based Video Retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1618–1626.

Jiang, Q.-Y.; He, Y.; Li, G.; Lin, J.; Li, L.; and Li, W.-J. 2019. SVD: A Large-Scale Short Video Dataset for Near-Duplicate Video Retrieval. In *Proceedings of International Conference on Computer Vision*.

Jiang, Y.-G.; Jiang, Y.; and Wang, J. 2014. VCDB: a large-scale database for partial copy detection in videos. In *European conference on computer vision*, 357–371. Springer.

Katharopoulos, A.; Vyas, A.; Pappas, N.; and Fleuret, F. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, 5156–5165. PMLR.

Kordopatis-Zilos, G.; Papadopoulos, S.; Patras, I.; and Kompatsiaris, I. 2019a. FIVR: Fine-grained incident video retrieval. *IEEE Transactions on Multimedia*, 21(10): 2638–2652.

Kordopatis-Zilos, G.; Papadopoulos, S.; Patras, I.; and Kompatsiaris, I. 2019b. ViSiL: Fine-grained spatio-temporal video similarity learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6351–6360.

Lee, J.; Yi, J.; Shin, C.; and Yoon, S. 2021a. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2643–2652.

Lee, P.; Wang, J.; Lu, Y.; and Byun, H. 2021b. Weakly-supervised temporal action localization by uncertainty modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1854–1862.

Liu, C.; Yang, H.; Fu, J.; and Qian, X. 2022. Learning Trajectory-Aware Transformer for Video Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5687–5696.

Papakipos, Z.; Tolias, G.; Jenicek, T.; Pizzi, E.; Yokoo, S.; Wang, W.; Sun, Y.; Zhang, W.; Yang, Y.; Addicam, S.; et al. 2022. Results and findings of the 2021 Image Similarity Challenge. In *NeurIPS 2021 Competitions and Demonstrations Track*, 1–12. PMLR.

Poullot, S.; Tsukatani, S.; Nguyen, A.; Jégou, H.; and Satoh, S. 2015. Temporal Matching Kernel with Explicit Feature Maps. In *Proceedings of the 23rd ACM international conference on Multimedia*, 381–390.

Rocco, I.; Cimpoi, M.; Arandjelović, R.; Torii, A.; Pajdla, T.; and Sivic, J. 2018. Neighbourhood consensus networks. *Advances in neural information processing systems*, 31.

Shao, J.; Wen, X.; Zhao, B.; and Xue, X. 2021. Temporal Context Aggregation for Video Retrieval With Contrastive Learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Shen, Y.; and Elhamifar, E. 2022. Semi-Weakly-Supervised Learning of Complex Actions from Instructional Task Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3344–3354.

Sohn, K.; Zhang, Z.; Li, C.-L.; Zhang, H.; Lee, C.-Y.; and Pfister, T. 2020. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7262–7272.

Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; and Zhou, X. 2021. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8922–8931.

Taherkhani, F.; Dabouei, A.; Soleymani, S.; Dawson, J.; and Nasrabadi, N. M. 2021. Self-supervised wasserstein pseudo-labeling for semi-supervised image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12267–12277.

Tan, H.-K.; Ngo, C.-W.; Hong, R.; and Chua, T.-S. 2009. Scalable Detection of Partial Near-Duplicate Videos by Visual-Temporal Consistency. MM '09, 145–154. New York, NY, USA: Association for Computing Machinery. ISBN 9781605586083.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yang, J.; Dong, X.; Liu, L.; Zhang, C.; Shen, J.; and Yu, D. 2022. Recurring the Transformer for Video Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14063–14073.

Yokoo, S. 2021. Contrastive learning with large memory bank and negative embedding subtraction for accurate copy detection. *arXiv preprint arXiv:2112.04323*.

Zhou, Q.; Yu, C.; Wang, Z.; Qian, Q.; and Li, H. 2021. Instant-teaching: An end-to-end semi-supervised object detection framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4081–4090.

Zhou, T.; Zhang, M.; Zhao, F.; and Li, J. 2022. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4299–4309.