



City Research Online

City St George's, University of London

Citation: Kordopatis-Zilos, G., Tzelepis, C., Papadopoulos, S., Kompatsiaris, I. & Patras, I. (2022). DnS: Distill-and-Select for Efficient and Accurate Video Indexing and Retrieval. *International Journal of Computer Vision*, 130(10), pp. 2385-2407. doi: 10.1007/s11263-022-01651-3

This is the published version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: <https://openaccess.city.ac.uk/id/eprint/31351/>

Link to published version: <https://doi.org/10.1007/s11263-022-01651-3>

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).



DnS: Distill-and-Select for Efficient and Accurate Video Indexing and Retrieval

Giorgos Kordopatis-Zilos^{1,2} · Christos Tzelepis² · Symeon Papadopoulos¹ · Ioannis Kompatsiaris¹ · Ioannis Patras²

Received: 24 June 2021 / Accepted: 11 July 2022 / Published online: 5 August 2022
© The Author(s) 2022

Abstract

In this paper, we address the problem of high performance and computationally efficient content-based video retrieval in large-scale datasets'. Current methods typically propose either: (i) fine-grained approaches employing spatio-temporal representations and similarity calculations, achieving high performance at a high computational cost or (ii) coarse-grained approaches representing/indexing videos as global vectors, where the spatio-temporal structure is lost, providing low performance but also having low computational cost. In this work, we propose a Knowledge Distillation framework, called Distill-and-Select (DnS), that starting from a well-performing fine-grained Teacher Network learns: (a) Student Networks at different retrieval performance and computational efficiency trade-offs and (b) a Selector Network that at test time rapidly directs samples to the appropriate student to maintain both high retrieval performance and high computational efficiency. We train several students with different architectures and arrive at different trade-offs of performance and efficiency, i.e., speed and storage requirements, including fine-grained students that store/index videos using binary representations. Importantly, the proposed scheme allows Knowledge Distillation in large, unlabelled datasets—this leads to good students. We evaluate DnS on five public datasets on three different video retrieval tasks and demonstrate (a) that our students achieve state-of-the-art performance in several cases and (b) that the DnS framework provides an excellent trade-off between retrieval performance, computational speed, and storage space. In specific configurations, the proposed method achieves similar mAP with the teacher but is 20 times faster and requires 240 times less storage space. The collected dataset and implementation are publicly available: <https://github.com/mever-team/distill-and-select>.

1 Introduction

Due to the popularity of Internet-based video sharing services, the volume of video content on the Web has reached unprecedented scales. For instance, YouTube reports that more than 500 h of content are uploaded every minute.¹ This puts considerable challenges for all video analysis problems, such as video classification, action recognition, and video retrieval, that need to achieve high performance at low computational and storage requirements in order to deal with the large scale of the data. The problem is particularly hard in the case of content-based video retrieval, where, given a query video, one needs to calculate its similarity with all videos in a database to retrieve and rank the videos based on relevance. In such scenario, this requires efficient indexing, i.e., storage of the representations extracted from the videos in the

Communicated by Shin'ichi Satoh.

✉ Giorgos Kordopatis-Zilos
georgekordopatis@iti.gr

Christos Tzelepis
c.tzelepis@qmul.ac.uk

Symeon Papadopoulos
papadop@iti.gr

Ioannis Kompatsiaris
ikom@iti.gr

Ioannis Patras
i.pstras@qmul.ac.uk

¹ Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece

² Queen Mary University of London, Mile End Road, London E1 4NS, UK

¹ <https://www.youtube.com/yt/about/press/> accessed June 2021

dataset, and fast calculations of the similarity between pairs of them.

Depending on whether the spatio-temporal structure of videos is stored/indexed and subsequently taken into consideration during similarity calculation, research efforts fall into two broad categories, namely *coarse-* and *fine-grained* approaches. *Coarse-grained* approaches address this problem by aggregating frame-level features into single video-level vector representations (that are estimated and stored at indexing time) and then calculating the similarity between them by using a simple function such as the dot-product or the Euclidean distance (at retrieval time). The video-level representations can be global vectors (Gao et al., 2017; Kordopatis-Zilos et al., 2017b; Lee et al., 2020), hash codes (Song et al., 2011, 2018; Yuan et al., 2020), Bag-of-Words (BoW) (Cai et al., 2011; Kordopatis-Zilos et al., 2017a; Liao et al., 2018), or concept annotations (Markatopoulou et al., 2017, 2018; Liang & Wang, 2020). These methods have very low storage requirements, allow rapid similarity estimation at query-time, but they exhibit low retrieval performance, since they disregard the spatial and temporal structure of the videos and are therefore vulnerable to clutter and irrelevant content. On the other hand, *fine-grained* approaches extract (and store at indexing time) and use in the similarity calculation (at retrieval time) representations that respect the spatio-temporal structure of the original video, i.e., they have a temporal or a spatio-temporal dimension/index. Typically, such methods consider the sequence of frames in the similarity calculation and align them, e.g., by using Dynamic Programming (Chou et al., 2015; Liu et al., 2017), Temporal Networks (Tan et al., 2009; Jiang & Wang, 2016), or Hough Voting (Douze et al., 2010; Jiang et al., 2014); or consider spatio-temporal video representation and matching based on Recurrent Neural Networks (RNN) (Feng et al., 2018; Bishay et al., 2019), Transformer-based architectures (Shao et al., 2021), or in the Fourier domain (Poullot et al., 2015; Baraldi et al., 2018). These approaches achieve high retrieval performance but at considerable computation and storage cost.

In an attempt to exploit the merits of both fine- and coarse-grained methods, some works tried to utilize them in a single framework (Wu et al., 2007; Chou et al., 2015; Liang & Wang, 2020), leading to methods that offer a trade-off between computational efficiency and retrieval performance. Typically, these approaches first rank videos based on a coarse-grained method, in order to filter the videos with similarity lower than a predefined threshold, and then re-rank the remaining ones based on the similarity calculated from a computationally expensive fine-grained method. However, setting the threshold is by no means a trivial task. In addition, in those approaches, both coarse- and fine-grained components are typically built based on hand-crafted features with traditional aggregations (e.g., BoW) and heuristic/non-learnable approaches for similarity calculation—this results in sub-

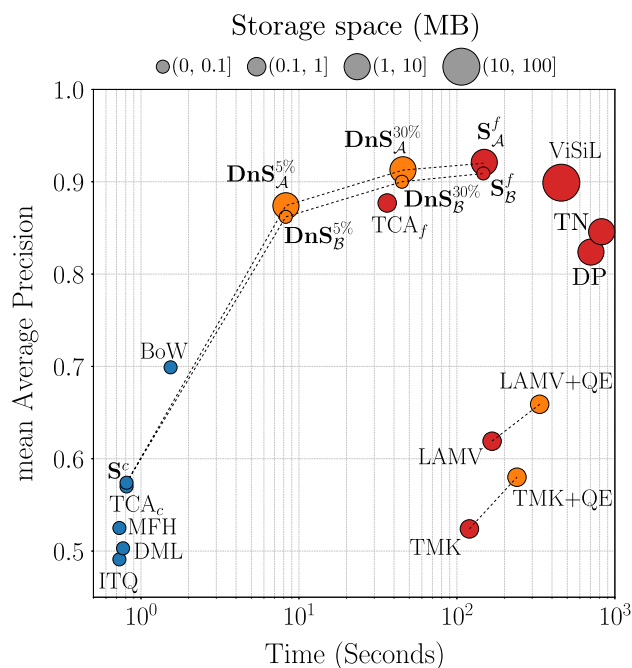


Fig. 1 Performance of our proposed **DnS** framework and its variants for several dataset percentages sent for re-ranking (denoted in bold) evaluated on the DSVR task of FIVR-200K in terms of mAP, computational time per query in seconds, and storage space per video in megabytes (MB), in comparison to state-of-the-art methods. Coarse-grained methods are in blue, fine-grained in red, and re-ranking in orange (Color figure online)

optimal performance. We will be referring to such approaches as *re-ranking* methods.

Figure 1 illustrates the retrieval performance, time per query, and storage space per video of several methods from the previous categories. Fine-grained approaches achieve the best results but with a significant allocation of resources. On the other hand, coarse-grained approaches are very lightweight but with considerably lower retrieval performance. Finally, the proposed re-ranking method provides a good trade-off between accuracy and efficiency, achieving very competitive performance with low time and storage requirements.

Knowledge Distillation is a methodology in which a student network is being trained so as to approximate the output of a teacher network, either in the labelled dataset in which the teacher was trained, or in other, potentially larger unlabelled ones. Depending on the student's architecture and the size of the dataset, different efficiency-performance trade-offs can be reached. These methods have been extensively used in the domain of image recognition (Yalniz et al., 2019; Touvron et al., 2020; Xie et al., 2020); however, in the domain of video analysis, they are limited to video classification methods (Bhardwaj et al., 2019; Garcia et al., 2018; Crasto et al., 2019; Stroud et al., 2020), typically performing distillation at feature level across different modalities. Those methods

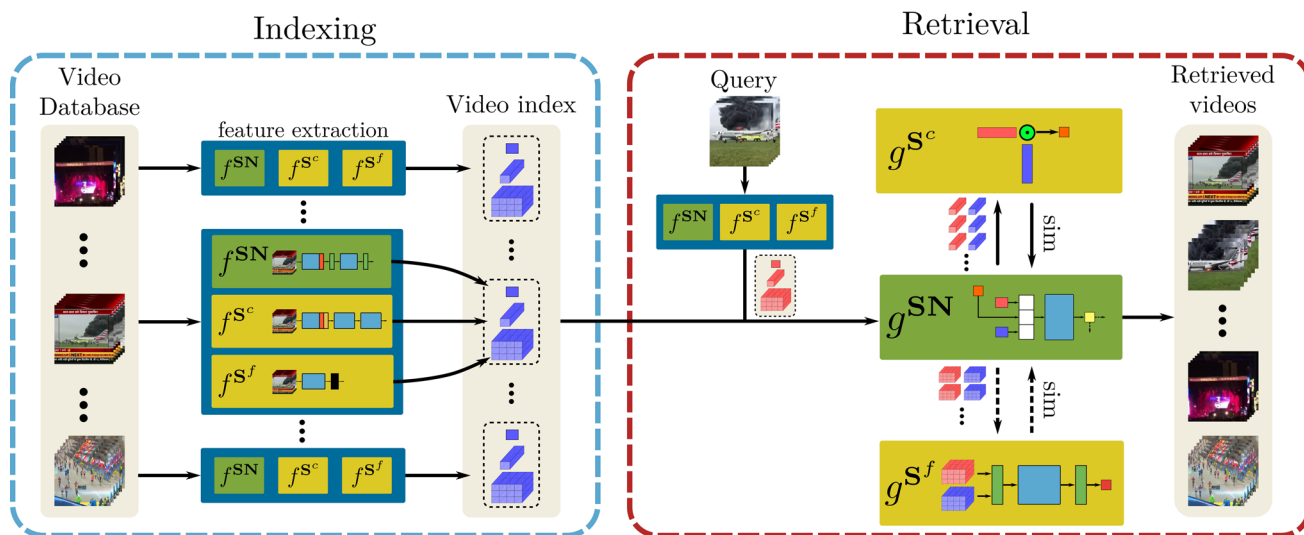


Fig. 2 Overview of the proposed framework. It consists of three networks: a coarse-grained student S^c , a fine-grained student S^f , and a selector network SN . Processing is split into two phases, *Indexing* and *Retrieval*. During indexing (blue box), given a video database, three representations needed by our networks are extracted and stored in a video index, i.e., for each video, we extract a 3D tensor, a 1D vector, and a scalar that captures video self-similarity. During retrieval (red box), given a query video, we extract its features, which, along with the indexed ones, are processed by the SN . It first sends all the 1D vectors of query-target pairs to S^c for an initial similarity calculation. Then, based

on the calculated similarity and the self-similarity of the videos, the selector network judges which query-target pairs have to be re-ranked with the S^f , using the 3D video tensors. Straight lines indicate continuous flow, i.e., all videos/video pairs are processed, whereas dashed lines indicate conditional flow, i.e., only a number of selected videos/video pairs are processed. Our students are trained with Knowledge Distillation based on a fine-grained teacher network, and the selector network is trained based on the similarity difference between the two students (Color figure online)

typically distill the features of a stream of the network operating in a (computationally) expensive modality (e.g., optical flow field, or depth) into the features of a cheaper modality (e.g., RGB images) so that only the latter need to be stored/extracted and processed at test time. This approach does not scale well on large datasets, as it requires storage or re-estimation of the intermediate features. Furthermore, current works arrive at fixed trade-offs of performance and computational/storage efficiency.

In this work, we propose to address the problem of high retrieval performance and computationally efficient content-based video retrieval in large-scale datasets. The proposed method builds on the framework of Knowledge Distillation, and starting from a well-performing, high-accuracy-high-complexity teacher, namely a fine-grained video similarity learning method (ViSiL) (Kordopatis-Zilos et al., 2019b), trains (a) both fine-grained and coarse-grained student networks on a large-scale unlabelled dataset and (b) a selection mechanism, i.e., a learnable re-ranking module, that decides whether the similarity estimated by the coarse-grained student is accurate enough, or whether the fine-grained student needs to be invoked. By contrast to other re-ranking methods that use a threshold on the similarity estimated by the fast network (the coarse-grained student in our case), our selection mechanism is a trainable, lightweight neural net-

work. All networks are trained so as to extract representations that are stored/indexed, so that each video in the database is indexed by the fine-grained spatio-temporal representation (3D tensor), its global, vector-based representation (1D vector), and a scalar self-similarity measure that is extracted by the feature extractor of the selector network, and can be seen as a measure of the complexity of the videos in question. The latter is expected to be informative of how accurate the coarse-grained, video-level similarity is, and together with the similarity rapidly estimated by the coarse-grained representations, is used as input to the selector. We note that, by contrast to other Knowledge Distillation methods in videos that address classification problems and typically perform distillation at intermediate features, the students are trained on a similarity measure provided by the teacher—this allows training on large scale datasets as intermediate features of the networks do not need to be stored, or estimated multiple times. Due to the ability to train on large unlabeled datasets, more complex models, i.e., with more trainable parameters, can be employed leading to even better performance than the original teacher network. An overview of the proposed framework is illustrated in Fig. 2.

The main contributions of this paper can be summarized as follows:

- We build a re-ranking framework based on a Knowledge Distillation scheme and a Selection Mechanism that allows for training our student and selector networks using large unlabelled datasets. We employ a teacher network that is very accurate but needs a lot of computational resources to train several student networks and the selector networks, and use them to achieve different performance-efficiency trade-offs.
- We propose a selection mechanism that, given a pair of a fine- and a coarse-grained student, learns whether the similarity estimated by the fast, coarse-grained student is accurate enough, or whether the slow, fine-grained student needs to be invoked. To the best of our knowledge, we are the first to propose such a trainable selection scheme based on video similarity.
- We propose two fine-grained and one coarse-grained student architectures. We develop: (i) a fine-grained attention student, using a more complex attention scheme than the teacher's, (ii) a fine-grained binarization student that extracts binarized features for the similarity calculation, and (iii) a coarse-grained attention student that exploits region-level information, and the intra- and inter-video relation of frames for the aggregation.
- We evaluate the proposed method on five publicly available datasets and compare it with several state-of-the-art methods. Our fine-grained student achieves state-of-the-art performance on two out of four datasets, and our DnS approach retains competitive performance with more than 20 times faster retrieval per query and 99% lower storage requirements compared to the teacher.

The remainder of the paper is organised as follows. In Sect. 2, the related literature is discussed. In Sect. 3, the proposed method is presented in detail. In Sect. 4, the datasets and implementation are presented. In Sect. 5, the results and ablation studies are reported. In Sect. 6, we draw our conclusions.

2 Related Work

This section gives an overview of some of the fundamental works that have contributed to content-based video retrieval and knowledge distillation.

2.1 Video Retrieval

The video retrieval methods can be roughly classified, based on the video representations and similarity calculation processes employed, in three categories: coarse-grained, fine-grained, and re-ranking approaches.

2.1.1 Coarse-Grained Approaches

Coarse-grained approaches represent videos with a global video-level signature, such as an aggregated feature vector or a binary hash code, and use a single operation for similarity calculation, such as a dot product. A straightforward approach is the extraction of global vectors as video representations combined with the dot product for similarity calculation. Early works (Wu et al., 2007; Huang et al., 2010) extracted hand-crafted features from video frames, i.e., color histograms, and aggregated them to a global vector. More recent works (Gao et al., 2017; Kordopatis-Zilos et al., 2017b; Lee et al., 2018, 2020) rely on CNN features combined with aggregation methods. Also, other works (Cai et al., 2011; Kordopatis-Zilos et al., 2017a) aggregate video content to Bag-of-Words (BoW) representation (Sivic and Zisserman, 2003) by mapping frames to visual words and extracting global representations with tf-idf weighting. Another popular direction is the generation of hash codes for the entire videos combined with Hamming distance (Song et al., 2011, 2018; Liong et al., 2017; Yuan et al., 2020). Typically, the hashing is performed via a network trained to preserve relations between videos. Coarse-grained methods provide very efficient retrieval covering the scalability needs of web-scale applications; however, their retrieval performance is limited, typically outperformed by the fine-grained approaches.

2.1.2 Fine-Grained Approaches

Fine-grained approaches extract video representations, ranging from video-level to region-level, and calculate similarity by considering spatio-temporal relations between videos based on several operations, e.g., a dot product followed by a max operation. Tan et al. (2009) proposed a graph-based Temporal Network (TN) structure, used for the detection of the longest shared path between two compared videos, which has also been combined with frame-level deep learning networks (Jiang & Wang, 2016; Wang et al., 2017). Additionally, other approaches employ Temporal Hough Voting (Douze et al., 2010) to align matched frames by means of a temporal Hough transform. Another solution is based on Dynamic Programming (DP) (Chou et al., 2015), where the similarity matrix between all frame pairs is calculated, and then the diagonal blocks with the largest similarity are extracted. Another direction is to generate spatio-temporal representations with the Fourier transform in a way that accounts for the temporal structure of video similarity (Poullot et al., 2015; Baraldi et al., 2018). Finally, some recent works rely on attention-based schemes to learn video comparison and aggregation by training either attentional RNN architectures (Feng et al., 2018; Bishay et al., 2019), transformer-based networks for temporal aggregation (Shao et al., 2021), or multi-attentional networks that

extract multiple video representations (Wang et al., 2021). Fine-grained methods achieve high retrieval performance; however, they do not scale well to massive datasets due to their high computational and storage requirements.

2.1.3 Video Re-Ranking

Re-ranking is a common practice in retrieval systems. In the video domain, researchers have employed it to combine methods from the two aforementioned categories (i.e., coarse- and fine-grained) to overcome their bottleneck and achieve efficient and accurate retrieval (Wu et al., 2007; Douze et al., 2010; Chou et al., 2015; Yang et al., 2019; Liang & Wang, 2020). Typical methods deploy a coarse-grained method as an indexing scheme to quickly rank and filter videos, e.g., using global vectors (Wu et al., 2007) or BoW representations (Chou et al., 2015; Liang & Wang, 2020). Then, a fine-grained algorithm, such as DP (Chou et al., 2015), Hough Voting (Douze et al., 2010) or frame-level matching (Wu et al., 2007), is applied on the videos that exceed a similarity threshold in order to refine the similarity calculation. Another re-ranking approach employed for video retrieval is Query Expansion (QE) (Chum et al., 2007). It is a two-stage retrieval process where, after the first stage, the query features are re-calculated based on the most similar videos retrieved, and the query process is executed again with the new query representation. This has been successfully employed with both coarse-grained (Douze et al., 2013; Gao et al., 2017; Zhao et al., 2019) and fine-grained (Pouillot et al., 2015; Baraldi et al., 2018) approaches. Also, an attention-based trainable QE scheme has been proposed in Gordo et al. (2020) for image retrieval. However, even though the retrieval performance is improved with QE, the total computational time needed for retrieval is doubled as the query process is applied twice.

2.2 Knowledge Distillation

Knowledge Distillation (Hinton et al., 2015) is a training scheme that involves two networks, a teacher network that is usually trained with supervision and a student network that leverages teacher's predictions for improved performance. A thorough review of the field is provided in Gou et al. (2021). Knowledge Distillation has been employed on various computer vision problems, i.e., image classification (Yalniz et al., 2019; Touvron et al., 2020; Xie et al., 2020), object detection (Li et al., 2017; Shmelkov et al., 2017; Deng et al., 2019), metric learning (Park et al., 2019; Peng et al., 2019), action recognition (Garcia et al., 2018; Thoker & Gall, 2019; Stroud et al., 2020), video classification (Zhang & Peng, 2018; Bhardwaj et al., 2019), video captioning (Pan et al., 2020; Zhang et al., 2020), and representation learning (Tavakolian et al., 2019; Piergiovanni et al., 2020).

Relevant works in the field of Knowledge Distillation distill knowledge based on the relations between data samples (Park et al., 2019; Tung & Mori, 2019; Liu et al., 2019; Lassance et al., 2020; Peng et al., 2019). Student networks are trained based on the distances between samples calculated by a teacher network (Park et al., 2019), the pairwise similarity matrix between samples within-batch (Tung & Mori, 2019), or by distilling graphs constructed based on the relations of the samples, using the sample representations as vertices and their distance as the edges to build an adjacency matrix (Liu et al., 2019; Lassance et al., 2020).

In the video domain, several approaches have been proposed for the improvement of the computational efficiency of the networks (Bhardwaj et al., 2019; Zhang & Peng, 2018; Garcia et al., 2018). Some works (Bhardwaj et al., 2019) proposed a Knowledge Distillation setup for video classification where the student uses only a fraction of the frames processed by the teacher, or multiple teachers are employed to construct a graph based on their relations, and then a smaller student network is trained (Zhang & Peng, 2018). Also, a popular direction is to build methods for distillation from different modalities and learn with privileged information to increase the performance of a single network, i.e., using depth images (Garcia et al., 2018), optical flow (Crasto et al., 2019; Stroud et al., 2020; Piergiovanni et al., 2020), or multiple modalities (Luo et al., 2018; Piergiovanni et al., 2020). In video retrieval, Knowledge Distillation has been employed for feature representation learning on frame-level using the evaluation datasets (Liang et al., 2019).

2.3 Comparison to Previous Approaches

In this section, we draw comparisons of the proposed approach to the related works from the literature with respect to the claimed novelties.

Proposed Framework There is no similar prior work in the video domain that builds a re-ranking framework based on Knowledge Distillation and a trainable Selection Mechanism based on which the re-ranking process is performed. Other works (Chou et al., 2015; Yang et al., 2019; Liang & Wang, 2020) rely on outdated hand-crafted methods using simple re-ranking approaches based on similarity thresholding, the selection of which is a non-trivial task. By contrast, in this work, a framework is proposed that starts from an accurate but heavy-weight teacher to train (a) both a fine-grained and coarse-grained student network on a large unlabelled dataset and (b) a selection mechanism, i.e., a learnable module based on which the re-ranking process is performed.

Knowledge Distillation To the best of our knowledge, there is no prior work in the video domain that trains a pairwise function that measures video similarity with distillation. Works that use a similar loss function for distillation are Park et al. (2019) and Tung and Mori (2019); however, these

approaches have been proposed for the image domain. Video-based approaches (Bhardwaj et al., 2019; Zhang & Peng, 2018; Garcia et al., 2018; Liang et al., 2019) distill information between intermediate representations, e.g., video/frame activations or attention maps—this is costly due to the high computational requirements of the teacher. By contrast, in our training scheme the teacher’s similarities of the video pairs used during training can be pre-computed—this allows training in large datasets in an unsupervised manner (i.e., without labels). Finally, these distillation methods end up with a single network that either offers compression or better performance—by contrast, in the proposed framework, we are able to arrive at different accuracy/speed/storage trade-offs.

Network architectures We propose three student network architectures that are trained with Knowledge Distillation in an unsupervised manner on large unannotated datasets avoiding in this way overfitting (cf. Sect. 5.1.2). Two fine-grained students are built based on our prior work in Kordopatis-Zilos et al. (2019b), with some essential adjustments to mitigate its limitations. A fine-grained attention student is developed using a more complex attention mechanism, which outperforms the Teacher when trained on the large unlabeled dataset. Also, a fine-grained binarization student is introduced with a binarization layer that has significantly lower storage requirements. Prior works have used binarization layers with coarse-grained approaches (Liong et al., 2017; Song et al., 2018; Yuan et al., 2020), but none learns a fine-grained similarity function based on binarized regional-level descriptors. Furthermore, a coarse-grained student is built. Its novelties are the use of a trainable region-level aggregation scheme—unlike other works that extract frame-level descriptors—and the combination of two aggregation components on frame-level that considers intra- and inter-video relations between frames. Prior works have employed a transformer encoder to capture intra-video frame relations (Shao et al., 2021), or a NetVLAD to capture inter-video ones (Miech et al., 2017); however, none combines the two components together.

3 Distill-and-Select

This section presents the Distill-and-Select (DnS) method for video retrieval. First, we describe the developed retrieval pipeline, which involves a fine-grained and a coarse-grained student network trained with Knowledge Distillation, and a selector network, acting as a re-ranking mechanism (Sect. 3.1). Then, we discuss the network architectures/alternatives employed in our proposed approach that offer different performance-efficiency trade-offs (Sect. 3.2). Finally, the training processes followed for the training of the proposed networks are presented (Sect. 3.3).

3.1 Approach Overview

Figure 2 depicts the DnS framework. It consists of three networks: (i) a coarse-grained student (S^c) that provides very fast retrieval speed but with low retrieval performance, (ii) a fine-grained student (S^f) that has high retrieval performance but with high computational cost, and (iii) a selector network (SN) that routes the similarity calculation of the video pairs and provides a balance between performance and time efficiency.

Each video in the dataset is stored/indexed using three representations: (i) a spatio-temporal 3D tensor f^{S^f} that is extracted (and then used at retrieval time) by the fine-grained student S^f , (ii) a 1D global vector f^{S^c} that is extracted (and then used at retrieval time) by the coarse-grained student S^c , and (iii) a scalar f^{SN} that summarises the similarity between different frames of the video in question that is extracted (and then used at retrieval time) by the selector network SN. The indexing process that includes the feature extraction is illustrated within the blue box in Figs. 2, 3, 4, 5 and is denoted as $f^X(\cdot)$ for each network X . At retrieval-time, given an input query-target video pair, the selector network sends to the coarse-grained student S^c the global 1D vectors so that their similarity is rapidly estimated (i.e., as the dot product of the representations) g^{S^c} . This coarse similarity and the self-similarity scalars for the videos in question are then given as input to the selector SN, which takes a binary decision g^{SN} on whether the calculated coarse similarity needs to be refined by the fine-grained student. For the small percentage of videos that this is needed, the fine-grained network calculates the similarity g^{S^f} based on the spatio-temporal representations. The retrieval process that includes the similarity calculation is illustrated within the red box in Figs. 2, 3, 4, 5 and is denoted as $g^X(\cdot, \cdot)$ for each network X .

In practice, we apply the above process on every query-target video pair derived from a database, and a predefined percentage of videos with the largest confidence score calculated by the selector is sent to the fine-grained student for re-ranking. With this scheme, we achieve very fast retrieval with very competitive retrieval performance.

3.2 Network Architectures

In this section, the architectures of all networks included in the DnS framework are discussed. First, the teacher network that is based on the ViSiL architecture is presented 3.2.1. Then, we discuss our student architectures, which we propose under a Knowledge Distillation framework that addresses the limitations introduced by the teacher; i.e., high resource requirements, both in terms of memory space for indexing, due to the region-level video tensors, and computational time for retrieval, due to the fine-grained similarity

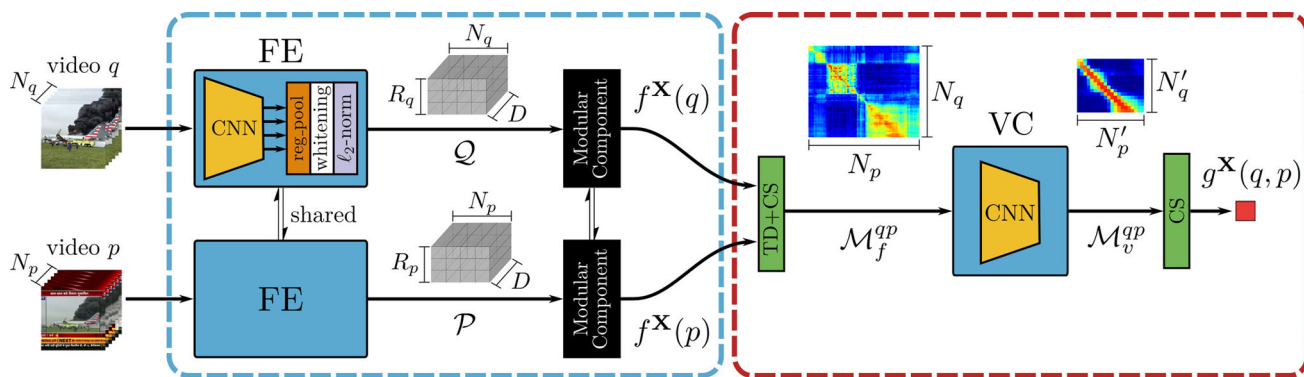


Fig. 3 Illustration of ViSiL (Kordopatis-Zilos et al., 2019b) architecture used for the teacher and fine-grained students, i.e., \mathbf{X} can take three values: \mathbf{T} , \mathbf{S}_A^f , and \mathbf{S}_B^f . During indexing, a 3D video tensor is extracted based on a Feature Extraction (FE) process, applying regional pooling, whitening, and ℓ^2 -normalization on the activations of a CNN. Then, a modular component is applied according to the employed network, i.e., an attention scheme for \mathbf{T} and \mathbf{S}_A^f , and a binarization layer for \mathbf{S}_B^f .

During retrieval, the Tensor Dot (TD) followed by Chamfer Similarity (CS) are applied on the representations of a video pair to generate their frame-to-frame similarity matrix, which is propagated to a Video Comparator (VC) CNN that captures the temporal patterns. Finally, CS is applied again to derive a single video-to-video similarity score (Color figure online)

calculation. More precisely, three students are proposed, two fine-grained and one coarse-grained variant, each providing different benefits. The fine-grained students are both using the ViSiL architecture. The first fine-grained student simply introduces more trainable parameters, leading to better performance with similar computational and storage requirements to the teacher 3.2.2. The second fine-grained student optimizes a binarization function that hashes features into a Hamming space and has very low storage space requirements for indexing with little performance sacrifice 3.2.3. The third coarse-grained student learns to aggregate the region-level feature vectors in order to generate a global video-level representation and needs considerably fewer resources for indexing and retrieval but at notable performance loss 3.2.3. Finally, we present the architecture of the selector network for indexing and retrieval 3.2.3. Our framework operates with a specific combination of a fine-grained and coarse-grained student and a selector network. Each combination achieves different trade-offs between retrieval performance, storage space, and computational time.

3.2.1 Baseline Teacher (T)

Here, we will briefly present the video similarity learning architecture that we employ as the teacher and which builds upon the ViSiL (Kordopatis-Zilos et al., 2019b) architecture (Fig. 3).

Feature Extraction/Indexing (f^T): Given an input video, we first extract region-level features from the intermediate convolution layers (Kordopatis-Zilos et al., 2017a) of a backbone CNN architecture by applying region pooling (Tolias et al., 2016) on the feature maps. These are further PCA-whitened (Jégou & Chum, 2012) and ℓ^2 -normalized. We denote the aforementioned process as Feature Extraction

(FE), and we employ it in all of our networks. FE is followed by a modular component, as shown in Fig. 3, that differs for each fine-grained student. In the case of the teacher, an attention mechanism is employed imposing that frame regions are weighted based on their saliency via a visual attention mechanism over region vectors based on an ℓ^2 -normalized context vector. The context vector is a trainable vector $\mathbf{u} \in \mathbb{R}^D$ that weights each region vector independently based on their dot-product. It is learned through the training process. Also, no fully-connected layer is employed to transform the region vectors for the attention calculation. We refer to this attention scheme as ℓ^2 -attention. The output representation of an input video x is a region-level video tensor $\mathcal{X} \in \mathbb{R}^{N_x \times R_x \times D}$, where N_x is the number of frames, R_x is the number of regions per frame, and D is the dimensionality of the region vectors—this is the output of the indexing process, and we denote it as $f^T(x)$.

Similarity Calculation/Retrieval (g^T): At retrieval time, given two videos, q and p , with N_q and N_p number of frames and R_q and R_p regions per frame, respectively, for every pair of frames, we first calculate the frame-to-frame similarity based on the similarity of their region vectors. More precisely, to calculate the *frame-to-frame* similarity on videos q and p , we calculate the Tensor Dot combined with Chamfer Similarity on the corresponding video tensors $f^T(q) = Q \in \mathbb{R}^{N_q \times R_q \times D}$ and $f^T(p) = P \in \mathbb{R}^{N_p \times R_p \times D}$ as follows

$$\mathcal{M}_f^{qp} = \frac{1}{R_q} \sum_{i=1}^{R_q} \max_{1 \leq j \leq R_p} Q_{\cdot(3,1)} P^\top(\cdot, i, j, \cdot), \tag{1}$$

where $\mathcal{M}_f^{qp} \in \mathbb{R}^{N_q \times N_p}$ is the output frame-to-frame similarity matrix, and the Tensor Dot axes indicate the channel

dimension of the corresponding video tensors. Also, the Chamfer Similarity is implemented as a max-pooling operation followed by an average-pooling on the corresponding dimensions. This process leverages the geometric information captured by region vectors and provides some degree of spatial invariance. Also, it is worth noting that this frame-to-frame similarity calculation process is independent of the number of frames and region vectors; thus, it can be applied on any video pair with arbitrary sizes and lengths.

To calculate the *video-to-video* similarity, the generated similarity matrix \mathcal{M}_f^{qp} is fed to a Video Comparator (VC) CNN module (Fig. 3), which is capable of learning robust patterns of within-video similarities. The output of the network is the refined similarity matrix $\mathcal{M}_v^{qp} \in \mathbb{R}^{N'_q \times N'_p}$. In order to calculate the final video-level similarity for two input videos q, p , i.e., $g^T(q, p)$, the hard tanh (Htanh) activation function is applied on the values of the aforementioned network output followed by Chamfer Similarity in order to obtain a single value, as follows

$$g^T(q, p) = \frac{1}{N'_q} \sum_{i=1}^{N'_q} \max_{1 \leq j \leq N'_p} \text{Htanh}(\mathcal{M}_v^{qp}(i, j)). \quad (2)$$

In that way, the VC takes temporal consistency into consideration by applying learnable convolutional operations on the frame-to-frame similarity matrix. Those enforce local temporal constraints while the Chamfer-based similarity provides invariance to global temporal transformations. Hence, similarly to the frame-to-frame similarity calculation, this process is a trade-off between respecting the video-level structure and being invariant to some temporal differences.

3.2.2 Fine-Grained Attention Student (S_A^f)

The first fine-grained student adopts the same architecture as the teacher (Sect. 3.2.1, Fig. 3), but uses a more complex attention scheme in the modular component, employed for feature weighting, as proposed in Yang et al. (2016).

Feature Extraction/Indexing ($f^{S_A^f}$): The Feature Extraction (FE) process is used to extract features, similar to the teacher. In the modular component shown in Fig. 3, we apply an attention weighting scheme as follows. Given a region vector $\mathbf{r}: \mathcal{X}(i, j, \cdot) \in \mathbb{R}^D$, where $i = 1, \dots, N_x$, $j = 1, \dots, N_y$, a non-linear transformation is applied, which is implemented as a fully-connected layer with tanh activation function, to form a hidden representation \mathbf{h} . Then, the attention weight is calculated as the dot product between \mathbf{h} and the context vector \mathbf{u} , followed by the sigmoid function,

as

$$\begin{aligned} \mathbf{h} &= \tanh(\mathbf{r} \cdot W_a + b_a), \\ \alpha &= \text{sig}(\mathbf{u} \cdot \mathbf{h}), \\ \mathbf{r}' &= \alpha \mathbf{r}, \end{aligned} \quad (3)$$

where $W_a \in \mathbb{R}^{D \times D}$ and $b_a \in \mathbb{R}^D$ are the weight and bias parameters of the hidden layer of the attention module, respectively, and $\text{sig}(\cdot)$ denotes the element-wise sigmoid function. We will be referring to this attention scheme as *h-attention*. The resulting 3D representation is the indexing output $f^{S_A^f}(x)$ for an input video x .

Similarity Calculation/Retrieval ($g^{S_A^f}$): To calculate similarity between two videos, we build the same process as for the teacher, i.e., we employ a Video Comparator (VC) and use the same frame-to-frame and video-to-video functions to derive $g^{S_A^f}(q, p)$ for two input videos q, p (Fig. 3).

In comparison to the teacher, this student (a) has very similar storage requirements, since in both cases, the videos are stored as non-binary spatio-temporal features, (b) has similar computational cost, since the additional attention layer introduces only negligible overhead, and (c) typically reaches better performance, since it has slightly higher capacity and can be trained in a much larger, unlabelled dataset.

3.2.3 Fine-Grained Binarization Student (S_B^f)

The second fine-grained student also adopts the same architecture as the teacher (Sect. 3.2.1, Fig. 3), except for the modular component where a binarization layer is introduced, as discussed below.

Feature Extraction/Indexing ($f^{S_B^f}$): $f^{S_B^f}$ is the part of the indexing of the student S_B^f that extracts a binary representation for an input video that will be stored and used at retrieval time. It uses the architecture of the teacher, where the modular component is implemented as a binarization layer (Fig. 3). This applies a binarization function $b: \mathbb{R}^D \rightarrow \{-1, 1\}^L$ that hashes the region vectors $\mathbf{r} \in \mathbb{R}^D$ to binary hash codes $\mathbf{r}_B \in \{-1, 1\}^L$ as

$$b(\mathbf{r}) = \text{sgn}(\mathbf{r} \cdot W_B), \quad (4)$$

where $W_B \in \mathbb{R}^{D \times L}$ denote the learnable weights and $\text{sgn}(\cdot)$ denotes the element-wise sign function.

However, since sgn is not a differentiable function, learning binarization parameters via backpropagation is not possible. To address this, we propose an approximation of the sign function under the assumption of small uncertainty in its input. More specifically, let $\text{sgn}: x \mapsto \{\pm 1\}$, where x is drawn from a uni-variate Gaussian distribution with given mean μ and fixed variance σ^2 , i.e., $x \sim \mathcal{N}(\mu, \sigma^2)$. Then,

the expected value² of the sign of x is given analytically as follows

$$\mathbb{E}[\text{sgn}(x)] = \text{erf}\left(\frac{\mu}{\sqrt{2\sigma^2}}\right), \tag{5}$$

where $\text{erf}(\cdot)$ denotes the error function. This is differentiable and therefore can serve as an activation function on the binarization parameters, that is,

$$b(\mathbf{r}) = \text{erf}\left(\frac{\mathbf{r} \cdot \mathbf{W}_B}{\sqrt{2\sigma^2}}\right), \tag{6}$$

where we use as variance an appropriate constant value (empirically set to $\sigma = 10^{-3}$). During training, we use (6), while during evaluation and hash code storage we use (4). After applying this operation to an arbitrary video x with N_x frames and R_x regions, we arrive at a binary tensor $\mathcal{X}_B \in \{\pm 1\}^{N_x \times R_x \times L}$, which is the indexing output $f^{\mathcal{S}_B^f}(x) = \mathcal{X}_B$ used by this student.

Similarity Calculation/Retrieval ($g^{\mathcal{S}_B^f}$): In order to adapt the similarity calculation processes with the binarization operation, the Hamming Similarity (HS) combined with Chamfer Similarity is employed as follows. Given two videos q, p with $f^{\mathcal{S}_B^f}(q) = \mathcal{Q}_B \in \{\pm 1\}^{N_q \times R_q \times L}$ and $f^{\mathcal{S}_B^f}(p) = \mathcal{P}_B \in \{\pm 1\}^{N_p \times R_p \times L}$ their binary tensors, respectively, we first calculate the HS between the two tensors with the use of Tensor Dot to calculate the similarity of all region pair combinations of the two videos and then apply Chamfer Similarity to derive the frame-to-frame similarity matrix $\mathcal{M}_B^{qp} \in \mathbb{R}^{N_q \times N_p}$. That is,

$$HS^{qp} = \mathcal{Q}_B \cdot (3,1) \mathcal{P}_B^\top / L, \tag{7}$$

$$\mathcal{M}_B^{qp} = \frac{1}{R_q} \sum_{i=1}^{R_q} \max_{1 \leq j \leq R_p} HS^{qp}(\cdot, i, j, \cdot),$$

Finally, a Video Comparator (VC) is applied on the frame-to-frame similarity matrices in order to calculate the final video-to-video similarity, similarly to (2) in the original teacher (Fig. 3)—this is denoted as $g^{\mathcal{S}_B^f}(q, p)$ for two input videos q, p .

In comparison to the teacher, this student (a) has remarkably lower storage requirements, since the binary spatio-temporal representations are 32 times smaller than the corresponding float ones (full precision), (b) has similar computational cost, as the architecture is very similar, and (c) reaches better performance since it is trained at a larger (despite being unlabelled) dataset. Note that this student only uses a binary input but is not a binarized network.

² $\mathbb{E}[\text{sgn}(x)] = 1 \cdot \mathbb{P}(x > 0) + (-1) \cdot \mathbb{P}(x < 0) = 1 - 2 \cdot \mathbb{P}(x < 0) = 1 - 2\mathbb{P}\left(z < -\frac{\mu}{\sqrt{2\sigma^2}}\right) \implies \mathbb{E}[\text{sgn}(x)] = 1 - 2\Phi\left(-\frac{\mu}{\sqrt{2\sigma^2}}\right) = \text{erf}\left(\frac{\mu}{\sqrt{2\sigma^2}}\right)$, where $z \sim \mathcal{N}(0, 1)$ denotes the standard Gaussian and Φ its CDF.

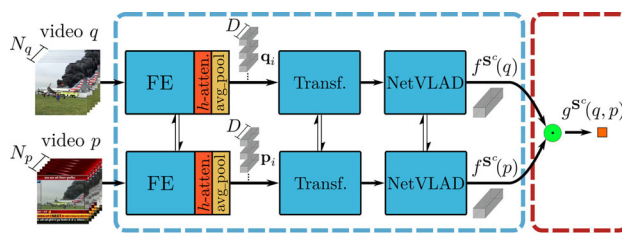


Fig. 4 Illustration of the architecture of the coarse-grained student \mathcal{S}^c , consisting of three main components. During indexing, the FE process with attention weighting and average pooling is applied to extract frame-level features. Then, they are processed by a Transformer network and aggregated to 1D vectors by a NetVLAD module. During retrieval, the video similarity derives from a simple dot product between the extracted representations (Color figure online)

3.2.4 Coarse-Grained Student (\mathcal{S}^c)

The coarse-grained student introduces an architecture that extracts video-level representations that are stored and can be subsequently used at retrieval time so as to rapidly estimate the similarity between two videos as the cosine similarity of their representations. An overview of the coarse student is shown in Fig. 4.

Feature Extraction/Indexing ($f^{\mathcal{S}^c}$): The proposed coarse-grained student comprises of three components. First, we extract weighted region-level features with Feature Extraction (FE), using the attention module given by (3), and then average pooling is applied across the spatial dimensions of the video tensors lead to frame-level representations for the videos; i.e., $\mathbf{x}_i = \frac{1}{R_x} \sum_{k=1}^{R_x} \alpha_k \mathbf{r}_k$, where $\mathbf{x}_i \in \mathbb{R}^D$ is the frame-level vector of the i -th video frame, R_x is the number of regions, and α_k is the attention weight calculated by (3). In that way, we apply a trainable scheme to aggregate the region-level features that focuses on the information-rich regions. Second, a transformer (Vaswani et al., 2017) network architecture is used to derive frame-level representations that capture long-term dependencies within the frame sequence, i.e., it captures the intra-video relations between frames. Following Shao et al. (2021), the encoder part of the Transformer architecture is used, which is composed of a multi-head self-attention mechanism and a feed-forward network. Finally, a NetVLAD (Arandjelovic et al., 2016) module aggregates the entire video to a single vector representation (Miech et al., 2017). This component learns a number of cluster centers and a soft assignment function through the training process, considering all videos in the training dataset. Therefore, it can be viewed as it encodes the inter-video relations between frames. Given input a video x , the output $f^{\mathcal{S}^c}(x)$ is a 1D video-level vector that is indexed and used by the coarse-grained student during retrieval.

Similarity Calculation/Retrieval ($g^{\mathcal{S}^c}$): Once feature representations have been extracted, the similarity calculation is a simple dot product between the 1D vectors of the com-

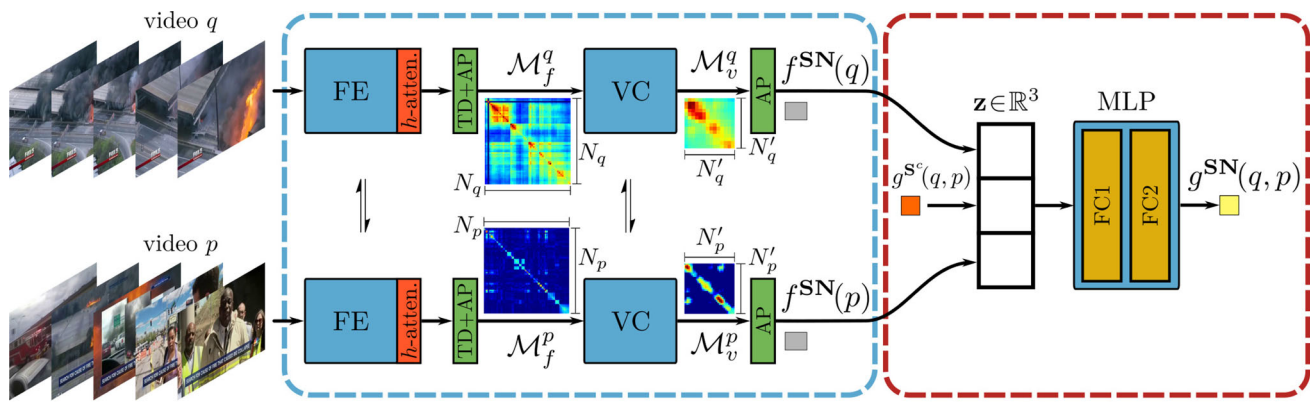


Fig. 5 Illustration of the Selector Network architecture. During indexing, the self-similarity of the videos is calculated according to the following scheme. First, region-level attention-weighted features are extracted. Then, the frame-to-frame self-similarity matrix derives with a Tensor Dot (DT) and Average Pooling (AP), which is propagated to

a VC module to capture temporal patterns. The final self-similarity is calculated based on an AP on the VC output. During retrieval, given a video pair, a 3-dimensional vector is composed by the self-similarity of each video and their similarity calculated by the S^c . The feature vector is fed to an MLP to derive a confidence score (Color figure online)

pared videos, i.e., $g^{S^c}(q, p) = f^{S^c}(q) \cdot f^{S^c}(p)$ for two input videos q, p .

In comparison to the original teacher, this student (a) has remarkably lower storage requirements for indexing, since it stores video-level representations instead of spatio-temporal ones, (b) has significantly lower computational cost at retrieval time, since the similarity is calculated with a single dot-product between video-level representations, and (c) has considerably lower performance, since it does not model spatio-temporal relations between videos during similarity calculation.

3.2.5 Selector Network (SN)

In the proposed framework, at retrieval time, given a pair of videos, the role of the selector is to decide whether the similarity that is calculated rapidly based on the stored coarse video-level representations is accurate enough (i.e., similar to what a fine-grained student would give), or whether a fine-grained similarity, based on the spatio-temporal, fine-grained representations needs to be used, and a new, refined similarity measure needs to be estimated. Clearly, this decision needs to be taken rapidly and with a very small additional storage requirement for each video.

The proposed selector network is shown in Fig. 5. At retrieval time, a simple Multi-Layer Perceptron (MLP) takes as input a three dimensional vector, $\mathbf{z} \in \mathbb{R}^3$, with the following features: (a) the similarity between a pair of videos q, p , as calculated by S^c (Sect. 3.2.4), and (b) the fine-grained self-similarities $f^{SN}(q)$ and $f^{SN}(p)$, calculated by a trainable NN (Fig. 5). Since $f^{SN}(x)$ depends only on video x , it can be stored together with the representations of the video x with negligible storage cost. Having $f^{SN}(q)$ and $f^{SN}(p)$ pre-computed, and g^{S^c} rapidly computed by the coarse-grained

student, the use of selector at retrieval time comes at a negligible storage and computational cost. Both the self-similarity function f^{SN} that extracts features at indexing time and the MLP that takes the decision at retrieval time, which are parts of the Selector Network SN, are jointly trained.

In what follows, we describe the architecture of the selector, starting from the network that calculates the fine-grained similarity f^{SN} . This is a modified version of the ViSiL architecture that aims to derive a measure that captures whether there is large spatio-temporal variability in its content. This is expected to be informative on whether the fine-grained student needs to be invoked. The intuition is that for videos with high f^{SN} , i.e., not high spatio-temporal variability, their video-level representations are sufficient to calculate their similarity, i.e., the similarity estimated by the coarse-grained student is accurate enough.

Feature Extraction/Indexing (f^{SN}): Given a video x as input, features are extracted based on the Feature Extraction (FE), using the attention module as in (3), to derive a video tensor $\mathcal{X} \in \mathbb{R}^{N_x \times R_x \times D}$. Then, the frame-to-frame self-similarity matrix is calculated, as

$$\mathcal{M}_f^x = \frac{1}{R_x^2} \sum_{i=1}^{R_x} \sum_{j=1}^{R_x} \mathcal{X}_{\cdot(3,1)} \mathcal{X}^\top(\cdot, i, j, \cdot), \quad (8)$$

where, $\mathcal{M}_f^x \in \mathbb{R}^{N_x \times N_x}$ is the symmetric frame-to-frame self-similarity matrix. Note that (8) is a modified version of (1), where the Chamfer Similarity is replaced by the average operator. In this case, we calculate the average similarity of a region with all other regions in the same frame—the use of Chamfer Similarity would have resulted in estimating the similarity of a region with the most similar region in the current frame, that is itself.

Similarly, a Video Comparator (VC) CNN network is employed (same as ViSiL, Fig. 3) that is fed with the self-similarity matrix in order to extract the temporal patterns and generate a refined self-similarity matrix $\mathcal{M}_v^x \in \mathbb{R}^{N'_x \times N'_x}$. To extract a final score (indexing output) that captures self-similarity, we modify (2) as

$$f^{\text{SN}}(x) = \frac{1}{N'_x{}^2} \sum_{i=1}^{N'_x} \sum_{j=1}^{N'_x} \mathcal{M}_v^x(i, j), \tag{9}$$

that is, the average of the pair-wise similarities of all video frames. Note that we also do not use the hard tanh activation function, as we empirically found that it is not needed.

Confidence Calculation/Retrieval (g^{SN}): Given a pair of videos and their similarity predicted by the S^c , we retrieve the indexed self-similarity scores, and then we concatenate them with the S^c similarity, forming a three-dimensional vector $\mathbf{z} \in \mathbb{R}^3$ for the video pair, as shown in Fig. 5. This vector is given as input to a two-layer MLP using Batch Normalization (Ioffe & Szegedy, 2015) and ReLU (Krizhevsky et al., 2012) activation functions. For an input video pair q, p , the retrieval output $g^{\text{SN}}(q, p)$ is the confidence score of the selector network that the fine-grained student needs to be invoked.

3.3 Training Process

In this section, we go through the details of the procedure followed for the training of the underlying networks of the proposed framework, i.e., the teacher, the students, and the selector.

3.3.1 Teacher Training

The teacher network is trained with supervision on a labelled video dataset \mathcal{V}_l , as shown in Fig. 6. The videos are organized in triplets (v, v^+, v^-) of an anchor, a positive (relevant), and a negative (irrelevant) video, respectively, where $v, v^+, v^- \in$

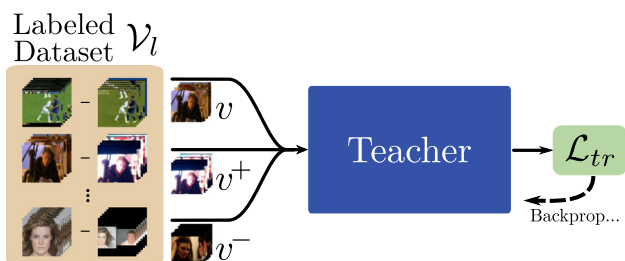


Fig. 6 Illustration of the training process of the teacher networks. It is trained with supervision with video triplets derived from a labelled dataset, minimizing triplet loss

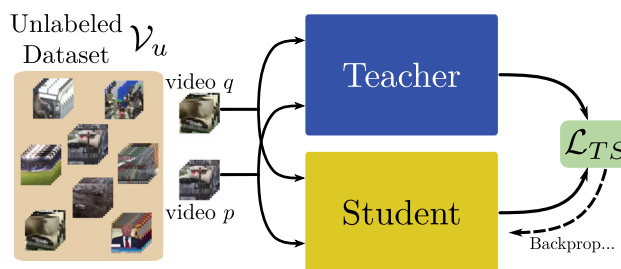


Fig. 7 Illustration of the training process of the student networks. They are trained on an unlabelled dataset by minimizing the difference between their video similarity estimations and the ones calculated by the teacher network

\mathcal{V}_l , and the network is trained with the triplet loss

$$\mathcal{L}_{tr} = \max \left(0, g^{\text{T}}(v, v^-) - g^{\text{T}}(v, v^+) + \gamma \right), \tag{10}$$

where γ is a margin hyperparameter. In addition, a similarity regularization function is used that penalizes high values in the input of hard tanh that would lead to saturated outputs. Following other works, we use data augmentation (i.e., color, geometric, and temporal augmentations) on the positive samples v^+ .

3.3.2 Student Training

An overview of the student training process is illustrated in Fig. 7. Let $\mathcal{V}_u = \{v_1, v_2, \dots, v_n\}$ be a collection of unlabelled videos and $g^{\text{T}}(q, p), g^{\text{S}}(q, p)$ be the similarities between videos $q, p \in \mathcal{V}_u$, estimated by a teacher network \mathbf{T} and a student network \mathbf{S} , respectively. \mathbf{S} is trained so that g^{S} approximates g^{T} , with the L_1 loss,³ that is,

$$\mathcal{L}_{TS} = \left\| g^{\text{T}}(q, p) - g^{\text{S}}(q, p) \right\|_1. \tag{11}$$

Note that the loss is defined on the output of the teacher. This allows for a training process in which the scores of the teacher are calculated for a number of pairs in the unlabelled dataset only once, and then being used as targets for the students. This is in contrast to methods where the loss is calculated on intermediate features of \mathbf{T} and \mathbf{S} , and cannot, thus, scale to large-scale datasets as they have considerable storage and/or computational/memory requirements. In this setting, the selection of the training pairs is crucial. Since it is very time consuming to apply the teacher network \mathbf{T} to every pair of videos in the dataset ($O(n^2)$ complexity) and randomly selecting videos would result in mostly pairs with low similarity scores, here, we follow Kordopatis-Zilos et al.

³ We have experimented with other losses, i.e., L_2 and Huber loss, with no considerable performance difference.

(2019a) and generate a graph to extract its connected components, which are considered as video clusters. Each video included in a video cluster is considered as an anchor, and we form pairs with the videos belonging to the same cluster, which are treated as positive pairs. Also, based on the anchor video, we form pairs with the 50 most similar videos that belong to the other clusters and the 50 most similar videos that belong to no cluster, which are treated as negative pairs. At each epoch, one positive and one negative pair are selected for each anchor video to balance their ratio.

3.3.3 Selector Training

Typically, the similarity between two videos q and p that is estimated by a fine-grained student S^f , leads to better retrieval scores than the one estimated by the coarse-grained student S^c . However, for some video pairs, the difference between them (i.e., $\|g^{S^c}(q, p) - g^{S^f}(q, p)\|_1$) is small and, therefore, having negligible effect to the ranking and on whether the video will be retrieved or not. The selector is a network that is trained to distinguish between those video pairs, and pairs of videos that exhibit large similarity differences. For the former, only the coarse-grained student S^c will be used; for the latter, the fine-grained student S^f will be invoked.

The selector network is trained as a binary classifier, with binary labels obtained by setting a threshold t on $\|g^{S^c}(q, p) - g^{S^f}(q, p)\|_1$, that is,

$$l(q, p) = \begin{cases} 1 & \text{if } \|g^{S^c}(q, p) - g^{S^f}(q, p)\|_1 > t, \\ 0 & \text{otherwise.} \end{cases} \quad (12)$$

Video pairs are derived from \mathcal{V}_u , and Binary Cross-Entropy is used as a loss function, as shown in Fig. 8. We use the same mining process used for the student training, and at each epoch, a fixed number of video pairs is sampled for the two classes. We reiterate here that the selector is trained in an end-to-end manner, i.e., both the self-similarity feature extraction network f^{SN} , given by (9), and the decision-making MLP (Fig. 5) are optimized jointly during training.

4 Evaluation Setup

In this section, we present the datasets (Sect. 4.1), evaluation metrics (Sect. 4.2), and implementation details (Sect. 4.3) adopted during the experimental evaluation of the proposed framework.

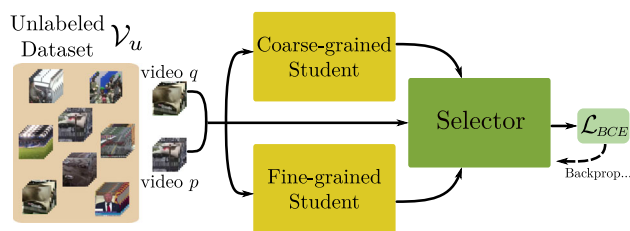


Fig. 8 Illustration of the training process of the selector network. It is trained on an unlabelled dataset, exploiting the similarities calculated by a coarse- and fine-grained student. Note that the fine-grained student is applied on all video pairs only during training time. During retrieval, only a portion of the dataset is sent to it

4.1 Datasets

4.1.1 Training Datasets

VCDB (Jiang et al., 2014) was used as the training dataset to generate triplets for the training of the teacher model. The dataset consists of videos derived from popular video platforms (i.e., YouTube and Metacafe) and has been developed and annotated as a benchmark for partial copy detection. It contains two subsets, namely, the core and the distractor subsets. The former one contains 28 discrete sets composed of 528 videos with over 9000 pairs of copied segments. The latter subset is a corpus of approximately 100,000 randomly collected videos that serve as distractors.

DnS-100K is the dataset collected for the training of the students. We followed the collection process from our prior work (Kordopatis-Zilos et al., 2019a) for the formation of the FIVR-200K dataset in order to collect a large corpus of videos with various relations between them. First, we built a collection of the major news events that occurred in recent years by crawling Wikipedia’s “Current Event” page.⁴ To avoid overlap with FIVR-200K, where the crawling period was from 2013–2017, we only considered the news events from the years 2018–2019. Then, we retained only the news events associated with armed conflicts and natural disasters by filtering them based on their topic. Afterwards, the public YouTube API⁵ was used to collect videos by providing the event headlines as queries. The results were filtered to contain only videos published at the corresponding event start date and up to one week after the event. At the end of this process, we had collected a corpus of 115,792 videos. Following the mining scheme described in Sect. 3.3.2, we arrived at 21,997 anchor videos with approximately 2.5M pairs.

⁴ https://en.wikipedia.org/wiki/Portal:Current_events.

⁵ <https://developers.google.com/youtube/>.

4.1.2 Evaluation Datasets

FIVR-200K (Kordopatis-Zilos et al., 2019a) was used as a benchmark for Fine-grained Incident Video Retrieval (FIVR). It consists of 225,960 videos collected based on the 4687 events and 100 video queries. It contains video-level annotation labels for near-duplicate (ND), duplicate scene (DS), complementary scene (CS), and incident scene (IS) videos. FIVR-200K includes three different tasks: a) the Duplicate Scene Video Retrieval (DSVR) task, where only videos annotated with ND and DS are considered relevant, b) the Complementary Scene Video Retrieval (CSVR) task, which accepts only the videos annotated with ND, DS, or CS as relevant, and c) Incident Scene Video Retrieval (ISVR) task, where all labels are considered relevant. For quick comparisons of different configurations, we also used **FIVR-5K**, a subset of FIVR-200K, as provided in Kordopatis-Zilos et al. (2019b).

CC_WEB_VIDEO (Wu et al., 2007) simulates the Near-Duplicate Video Retrieval (NDVR) problem. It consists of 24 query sets and 13,112 videos. The collection consists of a sample of videos retrieved by submitting 24 popular text queries to popular video sharing websites, i.e., YouTube, Google Video, and Yahoo! Video. For every query, a set of video clips was collected, and the most popular video was considered to be the query video. Subsequently, all videos in the video set were manually annotated based on their near-duplicate relation to the query video. We also use the ‘cleaned’ version, as provided in Kordopatis-Zilos et al. (2019b).

SVD (Jiang et al., 2019) was used for the NDVR problem, tailored for short videos in particular. It consists of 562,013 short videos crawled from a large video-sharing website, namely, Douyin.⁶ The average length of the collected videos is 17.33 s. The videos with more than 30,000 likes were selected to serve as queries. Candidate videos were selected and annotated based on a three-step retrieval process. A large number of probably negative unlabelled videos were also included to serve as distractors. Hence, the final dataset consists of 1206 queries with 34,020 labelled video pairs and 526,787 unlabelled videos. The queries were split into two sets, i.e., training and test set, with 1000 and 206 queries, respectively. In this paper, we only use the test set for the evaluation of the retrieval systems.

EVVE (Revaud et al., 2013) was designed for the Event Video Retrieval (EVR) problem. It consists of 2,375 videos and 620 queries. The main task on this dataset is the retrieval of all videos that capture the event depicted by a query video. The dataset contains 13 major events that were provided as queries to YouTube. Each event was annotated by one annotator, who first produced a precise definition of the event.

⁶ <http://www.douyin.com>.

However, we managed to download and process only 1906 videos and 504 queries (that is, $\approx 80\%$ of the initial dataset) due to the unavailability of the remaining ones.

4.2 Evaluation Metric

To evaluate retrieval performance, we use the *mean Average Precision* (mAP) metric, as defined in Wu et al. (2007), which captures the quality of video rankings. For each query, the *Average Precision* (AP) is calculated as

$$AP = \frac{1}{n} \sum_{i=0}^n \frac{i}{r_i}, \quad (13)$$

where n is the number of relevant videos to the query video and r_i is the rank of the i -th retrieved relevant video. The mAP is calculated by averaging the AP scores across all queries. Also, for the evaluation of the selector, we use the plot of mAP with respect to the total dataset percentage sent to the fine-grained student. The objective is to achieve high retrieval performance (in terms of mAP) with low dataset percentage.

4.3 Implementation Details

All of our models have been implemented with the PyTorch (Paszke et al., 2019) library. For the teacher, we have re-implemented ViSiL (Kordopatis-Zilos et al., 2019b) following the same implementation details, i.e., for each video, we extracted 1 frame per second and used ResNet-50 (He et al., 2016) for feature extraction using the output maps of the four residual blocks, resulting in $D = 3840$. The PCA-whitening layer was learned from 1M region vectors sampled from VCDB. In all of our experiments, the weights of the feature extraction CNN and whitening layer remained fixed. We sampled 2000 triplets in each epoch. The teacher was trained for 200 epochs with 4 videos per batch using the raw video frames. We employed Adam optimization (Kingma & Ba, 2015) with learning rate 10^{-5} . Other parameters were set to $\gamma = 0.5$, $r = 0.1$ and $W = 64$, similarly to Kordopatis-Zilos et al. (2019b).

For the students, we used the same feature extraction process as in the teacher, and the same PCA-whitening layer was used for whitening and dimensionality reduction. We empirically set $D = 512$ as the dimensions of the reduced region vectors. The students were trained with a batch size of 64 video pairs for 300 epochs, using only the extracted video features. Also, during training, we applied temporal augmentations, i.e., random frame drop, fast forward, and slow motion, with 0.1 probability each. We employed Adam optimization (Kingma & Ba, 2015) with learning rate 10^{-5} and 10^{-4} for the course- and fine-grained students, respectively. For the fine-grained binarization student, the binarization

layer was initialized with the ITQ algorithm (Gong et al., 2012), learned on 1M region vectors sampled from our dataset, as we observed better convergence than random initialization, and with $L = 512$ bits. For the coarse-grained student's training, the teacher's similarities were rescaled to $[0, 1]$ leading to better performance. Also, we used one layer in the transformer, with 8 heads for multi-head attention and 2048 dimension for the feed-forward network. For the NetVLAD module, we used 64 clusters, and a fully-connected layer with 1024 output dimensions and Layer Normalization (Ba et al., 2016). For the fine-grained students' training, we employed the similarity regularization loss from Kordopatis-Zilos et al. (2019b), weighted with 10^{-3} , arriving at marginal performance improvements.

For the selector, we used the same feature extraction scheme that was used for the students. It was trained with a batch size of 64 video pairs for 100 epochs, using only the extracted video features. At each epoch, we sampled 5000 video pairs from each class. We employed Adam optimization (Kingma & Ba, 2015) with learning rate 10^{-4} . For the fully-connected layers of the MLP, we used 100 hidden units and 0.5 dropout rate. For the training of the selector model, the similarities of the fine-grained student were rescaled to $[0, 1]$ to match similarities calculated from the coarse-grained student. Finally, we used a threshold of $t = 0.2$ for the class separation, unless stated otherwise.

5 Experiments

In this section, the experimental results of the proposed approach are provided. First, a comprehensive ablation study on the FIVR-5K dataset is presented, evaluating the proposed students and the overall approach under different configurations to gain better insight into its behaviour (Sect. 5.1). Then, we compare the performance and requirements of the developed solutions against several methods from the literature on the four benchmark datasets (Sect. 5.2).

5.1 Ablation Study

5.1.1 Retrieval Performance of the Individual Networks

In Table 1, we show the performance and storage/time requirements of the teacher \mathbf{T} and the three proposed student networks, namely, \mathbf{S}_A^f , \mathbf{S}_B^f , and \mathbf{S}^c , trained with the proposed scheme. The fine-grained attention student \mathbf{S}_A^f achieves the best results on all evaluation tasks, outperforming the teacher \mathbf{T} by a large margin. Also, the fine-grained binarization student \mathbf{S}_B^f reports performance very close to the teacher's on the DSVR and CSVr tasks, and it outperforms the teacher on the ISVR task, using only quantized features with lower dimen-

sionality than the ones used by the teacher and therefore requiring up to 240 times less storage space. This highlights the effectiveness of the proposed training scheme and the high quality of the collected dataset. Furthermore, both fine-grained students have similar time requirements, and they are three times faster than the teacher because they process lower dimensionality features. Finally, as expected, the coarse-grained student \mathbf{S}^c results in the worst performance compared to the other networks, but it has the lowest requirements in terms of both storage space and computational time.

5.1.2 Distillation Versus Supervision

In Table 2, we show the performance of the teacher \mathbf{T} trained with supervision on VCDB (as proposed in Kordopatis-Zilos et al. (2019b) and used for our teacher training) and the three proposed students, namely, \mathbf{S}_A^f , \mathbf{S}_B^f , and \mathbf{S}^c , trained under various combinations: (i) with supervision on VCDB (same as the original teacher), (ii) with distillation on VCDB, and (iii) with distillation on the DnS-100K dataset. It is evident

Table 1 Comparison of the teacher \mathbf{T} and the students \mathbf{S}_A^f , \mathbf{S}_B^f , and \mathbf{S}^c , in terms of mAP on FIVR-5K and computational requirements, i.e., storage space in KiloBytes (KB) per video and computational time in seconds (s) per query

Net.	DSVR	CSVr	ISVR	KB	Sec
\mathbf{T}	0.882	0.872	0.783	15124	10.10
\mathbf{S}_A^f	0.893	0.882	0.803	2016	3.300
\mathbf{S}_B^f	0.879	0.868	0.788	63	3.250
\mathbf{S}^c	0.634	0.647	0.608	4	0.018

Bold values indicate the best performing method in the corresponding dataset/subset

Table 2 Comparison of the teacher \mathbf{T} and the students \mathbf{S}_A^f , \mathbf{S}_B^f , and \mathbf{S}^c trained on different datasets and training schemes in terms of mAP on FIVR-5K

Task	Net.	VCDB w/ supervision	VCDB w/ distillation	DnS-100K w/ distillation
DSVR	\mathbf{T}	0.882	–	–
	\mathbf{S}_A^f	0.821	0.873	0.893
	\mathbf{S}_B^f	0.846	0.868	0.879
	\mathbf{S}^c	0.444	0.510	0.634
CSVr	\mathbf{T}	0.872	–	–
	\mathbf{S}_A^f	0.801	0.861	0.882
	\mathbf{S}_B^f	0.837	0.859	0.868
	\mathbf{S}^c	0.443	0.520	0.647
ISVR	\mathbf{T}	0.783	–	–
	\mathbf{S}_A^f	0.727	0.787	0.803
	\mathbf{S}_B^f	0.769	0.781	0.788
	\mathbf{S}^c	0.400	0.491	0.608

Table 3 Comparison of students S_A^f , S_B^f , and S^c , in terms of mAP, trained with different amount of training data on FIVR-5K

Task	Student	Dataset (%)			
		25	50	75	100
DSVR	S_A^f	0.888	0.894	0.893	0.893
	S_B^f	0.868	0.871	0.873	0.879
	S^c	0.573	0.608	0.621	0.634
CSVR	S_A^f	0.877	0.883	0.882	0.883
	S_B^f	0.858	0.861	0.863	0.868
	S^c	0.586	0.621	0.633	0.647
ISVR	S_A^f	0.801	0.803	0.803	0.805
	S_B^f	0.776	0.779	0.783	0.788
	S^c	0.564	0.583	0.594	0.608

Bold values indicate the best performing method in the corresponding dataset/subset

that the proposed training scheme using a large unlabelled dataset leads to considerably better retrieval performance compared to the other setups for all students. Also, it is noteworthy that training students with supervision, same as the teacher, results in a considerable drop in performance compared to distillation on either dataset. The students achieve better results when trained with DnS-100K instead of VCDB. An explanation for this is that our dataset contains various video relations (not only near-duplicates as in VCDB) and represents a very broad and diverse domain (by contrast to VCDB, which consists of randomly selected videos), resulting in better retrieval performance for the students.

5.1.3 Impact of Dataset Size

In Table 3, we show the performance of the proposed students, namely, S_A^f , S_B^f , and S^c , in terms of mAP, when they are trained with different percentages of the collected DnS-100K dataset (that is, 25%, 50%, 75%, and 100%). We report large differences in performance for the fine-grained binarization student S_B^f and the coarse-grained student S^c . We note that the more data is used for training, the better their retrieval results are. On the other hand, the fine-grained attention student’s S_A^f performance remains relatively steady, regardless of the amount used for training. We attribute this behaviour to the fact that S_A^f learns to weigh the input features without transforming them; hence, a smaller dataset with real video pairs with diverse relations, as in our collected dataset, is adequate for its robust performance.

5.1.4 Student Performance with Different Teachers

In Table 4, we show the performance of the proposed students, namely, S_A^f , S_B^f , and S^c , in terms of mAP, when they are trained/distilled using different teachers. More specifi-

Table 4 Comparison of students S_A^f , S_B^f , and S^c , in terms of mAP, trained with different teachers on FIVR-5K

Task	Student	Teacher		
		T	$S_A^{f(1)}$	$S_A^{f(2)}$
DSVR	S_A^f	0.893	0.896	0.899
	S_B^f	0.879	0.888	0.895
	S^c	0.634	0.631	0.632
CSVR	S_A^f	0.882	0.884	0.887
	S_B^f	0.868	0.878	0.883
	S^c	0.647	0.643	0.645
ISVR	S_A^f	0.803	0.807	0.810
	S_B^f	0.788	0.797	0.804
	S^c	0.608	0.603	0.607

Bold values indicate the best performing method in the corresponding dataset/subset

cally, using as a teacher: (i) the original teacher T , leading to the student $S_A^{f(1)}$, (ii) the fine-grained attention student $S_A^{f(1)}$, leading to the student $S_A^{f(2)}$ (first iteration), and (iii) the fine-grained attention student $S_A^{f(2)}$ (second iteration). In the case of fine-grained students, training with the $S_A^{f(1)}$ and $S_A^{f(2)}$ yields large performance boost in comparison to original teacher T . More precisely, the fine-grained attention student S_A^f exhibits a total improvement of about 0.006 mAP comparing its results trained with the teacher T (i.e., 0.893 mAP on DSVR task) and the $S_A^{f(2)}$ (i.e., 0.899 mAP on DSVR task). A very considerable improvement has the fine-grained binarization student, i.e., training with $S_A^{f(1)}$ gives a performance increase of almost 0.01 mAP on DSVR task, which further improves when trained with $S_A^{f(2)}$ by 0.007. On the other hand, using a better teacher does not improve the performance of the coarse-grained student S^c .

5.1.5 Student Performance with Different Settings

In this section, the retrieval performance of the proposed students is evaluated under different design choices.

Fine-Grained Attention Student In Table 5, we show how the adopted attention scheme (ℓ^2 -attention—Sect. 3.2.1, or h -attention—Sect. 3.2.2) affects the performance of the student S_A^f . Using h -attention leads to considerably better results compared to the ℓ^2 -attention, that was originally used in ViSiL (Kordopatis-Zilos et al., 2019b).

Fine-Grained Binarization Student In Table 6, we report the retrieval results of the fine-grained binarization student S_B^f implemented with different activation functions in the binarization layer, i.e., $\text{sgn}(\mathbf{x})$ which is not differentiable so the layer weights remain fixed, $\tanh(\beta\mathbf{x})$, as proposed in Cao et al. (2017) with $\beta = 10^3$, and the proposed $E[\text{sgn}(\mathbf{x})]$ (Sect. 3.2.3). The binarization student with the proposed

function achieves notably better results on all tasks, especially on ISVR. Moreover, we experimented with different number of bits for the region vectors and report results in Table 7. As expected, larger region hash codes lead to better retrieval performance. Nevertheless, the student achieves competitive retrieval performance even with low number of bits per region vector.

Coarse-Grained Student In Table 8, we report the performance of the coarse-grained student S^c implemented under various combinations of its components. The proposed setup with all three components achieves the best results compared to the other configurations. The single component that provides the best results is the transformer network, followed by NetVLAD. Also, the attention mechanism provides a considerable boost in performance when applied. The second-best performance is achieved with the combination of the transformer module with the NetVLAD.

5.1.6 Selector Network Performance

In this section, the performance of the proposed selector network is evaluated in comparison with the following approaches: (i) a selection mechanism that applies naive similarity thresholding for choosing between the coarse-grained and the fine-grained student, (ii) an oracle selector, where the similarity difference between the fine-grained and coarse-grained student is known and used for the re-ranking of video pairs, and (iii) a random selector that sends with a fixed probability videos to either the coarse-grained or the fine-grained student. Figure 9 illustrates the performance of the DnS approach in terms of mAP with respect to the percentage of video pairs from the evaluation dataset sent to the fine-grained student. We consider that the closer the curves are to the upper left corner, the better their performance. For

Table 5 Comparison of different attention schemes of the fine-grained attention student S_A^f , in terms of mAP on FIVR-5K

attention	DSVR	CSV	ISVR
ℓ^2 -attention (Kordopatis-Zilos et al., 2019b)	0.888	0.873	0.787
h -attention (Yang et al., 2016)	0.893	0.882	0.803

Bold values indicate the best performing method in the corresponding dataset/subset

Table 6 Comparison of different activation functions for fine-grained binarization student S_B^f , in terms of mAP on FIVR-5K

activation	DSVR	CSV	ISVR
sgn(x)	0.876	0.863	0.776
tanh(βx) (Cao et al., 2017)	0.875	0.861	0.781
$\mathbb{E}[\text{sgn}(x)]$ (ours)	0.879	0.868	0.788

Bold values indicate the best performing method in the corresponding dataset/subset

Table 7 Comparison of different fine-grained binarization student S_B^f implemented with different number of bits per region vector, in terms of mAP on FIVR-5K

bits	DSVR	CSV	ISVR
64	0.845	0.835	0.748
128	0.862	0.849	0.766
256	0.870	0.857	0.779
512	0.879	0.868	0.788

Bold values indicate the best performing method in the corresponding dataset/subset

Table 8 Comparison of different design choices of our coarse-grained student S^c , in terms of mAP on FIVR-5K

Att.	Trans.	NetVLAD	DSVR	CSV	ISVR
✓	–	–	0.595	0.600	0.564
–	✓	–	0.612	0.622	0.590
–	–	✓	0.600	0.610	0.578
✓	✓	–	0.620	0.628	0.591
✓	–	✓	0.609	0.618	0.584
–	✓	✓	0.630	0.637	0.600
✓	✓	✓	0.634	0.647	0.608

Bold values indicate the best performing method in the corresponding dataset/subset

this experiment, we used the proposed fine-grained attention student S_A^f and the coarse-grained student S^c . All three runs outperform the performance of the random selector by a large margin on all dataset percentages. The oracle selector performs the best with considerable margin, highlighting that using the similarity difference between the two students (Sect. 3.3.3) is a good optimization criterion. Furthermore, the proposed selector network outperforms the one with similarity thresholding on all tasks and percentages, i.e., in lower dataset percentages ($< 25\%$) with a large margin. It achieves more than 0.85 mAP on the DSVR task with only 10% of the video pairs in FIVR-5k sent to the fine-grained student.

5.1.7 Impact of Threshold on the Selector Performance

In this section, we assess the impact of the threshold parameter t that is used to obtain binary labels for the selector network [see Sect. 3.3.3, Eq. (12)], on the retrieval performance. To do so, we report the mAP as a function of the dataset percentage sent to the fine-grained student for re-ranking—we do so for selectors trained with different values of t in order to compare the curves. The results are shown in Fig. 10. The best results are obtained for $t = 0.2$; however, the performance is rather stable for thresholds between 0.1 and 0.4, as well. For threshold values > 0.4 , the performance drops considerably on all evaluation tasks.

Table 9 mAP comparison of our proposed students and re-ranking method against several video retrieval methods on four evaluation datasets

Approach	FIVR-200K				CC_WEB_VIDEO				SVD	EVVE
	DSVR	CSVR	ISVR	cc_web	cc_web*	cc_web _c	cc_web _c	cc_web _c *		
Coarse-grained	ITQ† (Gong et al., 2012)	0.491	0.472	0.402	0.967	0.941	0.976	0.954	0.842	0.606
	MFH† (Song et al., 2013)	0.525	0.507	0.424	0.972	0.950	0.981	0.967	0.849	0.604
	CSQ (Yuan et al., 2020)	0.267	0.252	0.219	0.948	0.899	0.954	0.909	0.364	0.400
	BoW† (Cai et al., 2011)	0.699	0.674	0.581	0.975	0.958	0.985	0.977	0.568	0.539
	LBoW (Kordopatis-Zilos et al., 2017a)	0.700	0.666	0.566	0.976	0.960	0.984	0.975	0.756	0.500
	DML (Kordopatis-Zilos et al., 2017b)	0.411	0.392	0.321	0.971	0.941	0.979	0.959	0.785	0.531
	DML† (Kordopatis-Zilos et al., 2017b)	0.503	0.487	0.410	0.971	0.951	0.979	0.965	0.850	0.611
	R-UTS-GV (Liang et al., 2019)	0.509	0.498	0.432	–	–	–	–	–	–
	TCA _c † (Shao et al., 2021)	0.570	0.553	0.473	0.973	0.947	0.983	0.965	–	0.598*
	Coarse-grained student-S^c (ours)	0.574	0.558	0.476	0.972	0.952	0.980	0.967	0.868	0.636
	TMK† (Poullot et al., 2015)	0.524	0.507	0.425	0.977	0.959	0.986	0.975	0.863	0.618
	LAMV (Baraldi et al., 2018)	0.496	0.466	0.371	0.975	0.956	0.986	0.975	0.781	0.531
	LAMV† (Baraldi et al., 2018)	0.619	0.587	0.479	0.978	0.964	0.988	0.982	0.880	0.620
	TN† (Tan et al., 2009)	0.844	0.804	0.660	0.982	0.970	0.993	0.989	0.894	0.471
DP† (Chou et al., 2015)	0.827	0.783	0.642	0.980	0.966	0.991	0.987	0.880	0.580	
R-UTS-FRP (Liang et al., 2019)	0.769	0.724	0.611	–	–	–	–	–	–	
A-DML (Wang et al., 2021)	0.627	–	–	0.964	0.949	–	–	0.885*	–	
TCA _r † (Shao et al., 2021)	0.877	0.830	0.703	0.983	0.969	0.994	0.990	–	0.603*	
ViSIL (Kordopatis-Zilos et al., 2019b)	0.899	0.854	0.723	0.985	0.973	0.995	0.992	0.881	0.658	
Fine-grained att. student-S^f_A (ours)	0.921	0.875	0.741	0.984	0.973	0.995	0.992	0.902	0.651	
Fine-grained bin. student-S^f_B (ours)	0.909	0.863	0.729	0.984	0.974	0.995	0.991	0.891	0.640	
PPT (Chou et al., 2015)	–	–	–	0.959	–	–	–	–	–	
HM (Liang & Wang, 2020)	–	–	–	0.977	–	–	–	–	–	
TMK†+QE (Poullot et al., 2015)	0.580	0.564	0.480	0.977	0.960	0.986	0.976	0.774	0.648	
LAMV†+QE (Baraldi et al., 2018)	0.659	0.629	0.520	0.979	0.964	0.990	0.984	0.786	0.653	
DnS _A ^{5%} (ours)	0.874	0.829	0.699	0.972	0.951	0.983	0.969	0.895	0.594	
DnS _B ^{5%} (ours)	0.862	0.817	0.687	0.970	0.948	0.981	0.967	0.884	0.584	
DnS _A ^{30%} (ours)	0.913	0.868	0.733	0.978	0.958	0.990	0.977	0.902	0.646	
DnS _B ^{30%} (ours)	0.900	0.854	0.720	0.977	0.954	0.988	0.974	0.894	0.634	

Bold values indicate the best performing method in the corresponding dataset/subset

†Runs are implemented with the same features extracted with the same process as ours

*Corresponding results are on different dataset split

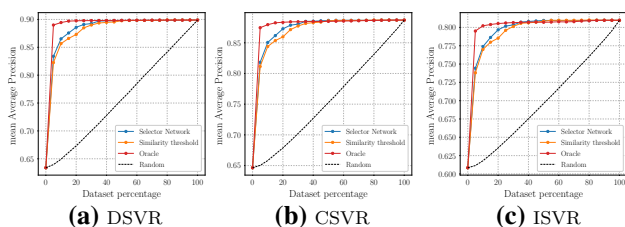


Fig. 9 mAP with respect to the dataset percentage sent to the fine-grained student for re-ranking based on four selectors: (i) the proposed selector network, (ii) a selector with naive similarity thresholding, (iii) an oracle selector, ranking videos based on the similarity difference between the two students, and (iv) a random selector

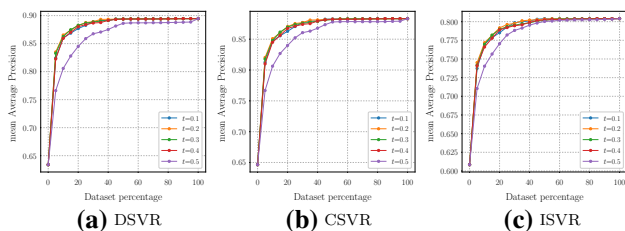


Fig. 10 mAP with respect to the dataset percentage sent to the fine-grained student for re-ranking based on our selector network trained with different values for the threshold t

5.2 Comparison with State of the Art

In this section, the proposed approach is compared with several methods from the literature on four datasets. In all experiments, the fine-grained attention student $S_A^{f(2)}$ is used as the teacher. We report the results of our re-ranking DnS scheme using both fine-grained students and sending the 5% and 30% of the dataset videos per query for re-ranking based on our selector score. We compare its performance with several coarse-grained, fine-grained, and re-ranking approaches: **ITQ** (Gong et al., 2012) and **MFH** (Song et al., 2013) are two unsupervised and **CSQ** (Yuan et al., 2020) a supervised video hashing methods using Hamming distance for video ranking, **BoW** (Cai et al., 2011) and **LBoW** (Kordopatis-Zilos et al., 2017a) extracts video representations based on BoW schemes with tf-idf weighting, **DML** (Kordopatis-Zilos et al., 2017b) extract a video embedding based on a network trained with DML, **R-UTS-GV** and **R-UTS-FRP** (Liang et al., 2019) are a coarse- and fine-grained methods trained with a teacher–student setup distilling feature representations, **TCA_c** and **TCA_f** (Shao et al., 2021) are a coarse- and fine-grained methods using a transformer-based architecture trained with contrastive learning, **TMK** (Poullot et al., 2015) and **LAMV** (Baraldi et al., 2018) extracts spatio-temporal video representations based on Fourier transform, which are also combined with **QE** (Douze et al., 2013), **TN** (Tan et al., 2009) employs a temporal network to find video segments with large similarity, **DP** (Chou et al., 2015) is a dynamic programming scheme for similarity

calculation, **A-DML** (Wang et al., 2021) assess video similarity extracting multiple video representations based on a multi-head attention network, **PPT** (Chou et al., 2015) is a re-ranking method with a BoW-based indexing scheme combined with DP for reranking, and **HM** (Liang & Wang, 2020) is also a re-ranking method using a concept-based similarity and a BoW-based method for refinement, and our re-implementation of **ViSiL** (Kordopatis-Zilos et al., 2019b). From the aforementioned methods, we have re-implemented BoW, TN, and DP, and we use the publicly available implementations for ITQ, MFH, CSQ, DML, TMK, and LAMV. For the rest, we provide the results reported in the original papers. Also, for fair comparison, we have implemented (if possible) the publicly available methods using our extracted features.

In Table 9, the mAP of the proposed method in comparison to the video retrieval methods from the literature is reported. The proposed students achieve very competitive performance achieving state-of-the-art results in several cases. First, the fine-grained attention student achieves the best results on the two large-scale datasets, i.e., FIVR-200K and SVD, outperforming ViSiL (our teacher network) by a large margin, i.e., 0.022 and 0.021 mAP, respectively. It reports almost the same performance as ViSiL on the CC_WEB_VIDEO dataset, and it is slightly outperformed on the EVVE dataset. Additionally, it is noteworthy that the fine-grained binarization student demonstrates very competitive performance on all datasets. It achieves similar performance with ViSiL and the fine-grained attention student on the CC_WEB_VIDEO, the second-best results on all three tasks of FIVR-200K, and the third-best on SVD with a small margin from the second-best. However, its performance is lower than the teacher’s on the EVVE dataset, highlighting that feature reduction and hashing have considerable impact on the student’s retrieval performance on this dataset. Also, another possible explanation for this performance difference could be that the training dataset does not cover the included events sufficiently.

Second, the coarse-grained student exhibits very competitive performance among coarse-grained approaches on all datasets. It achieves the best mAP on two out of four evaluation datasets, i.e., on SVD and EVVE, reporting performance close or even better than several fine-grained methods. On FIVR-200K and CC_WEB_VIDEO, it is outperformed by the BoW-based approaches, which are trained with samples from the evaluation sets. However, when they are built with video corpora other than the evaluation (which simulates more realistic scenarios), their performance drops considerably (Kordopatis-Zilos et al., 2017b, 2019a). Also, their performance on the SVD and EVVE datasets is considerably lower.

Third, our DnS runs maintain competitive performance. It improves the performance of the coarse-grained student by more than 0.2 on FIVR-200K and 0.02 on SVD by re-ranking

Table 10 Performance in mAP, storage in KiloBytes (KB) and time in seconds (S) requirements of our proposed students and re-ranking method and several video retrieval implemented with the same features

Approach	FIVR-200K			CC_WEB_VIDEO			SVD			EVVE			
	mAP	KB	Sec	mAP	KB	Sec	mAP	KB	Sec	mAP	KB	Sec	
Coarse-grained	ITQ (Gong et al., 2012)	0.491	0.1	0.733	0.954	0.1	0.045	0.842	0.1	1.793	0.606	0.1	0.005
	MFH (Song et al., 2013)	0.525	0.1	0.733	0.967	0.1	0.045	0.849	0.1	1.793	0.604	0.1	0.005
	BoW (Cai et al., 2011)	0.699	0.3	1.540	0.977	0.3	0.053	0.568	0.1	3.308	0.539	0.2	0.005
	DML (Kordopatis-Zilos et al., 2017b)	0.503	2	0.769	0.965	2	0.047	0.850	2	1.915	0.611	2	0.006
	TCA _c (Shao et al., 2021)	0.570	4	0.812	0.965	4	0.047	–	–	–	0.598*	4	0.006
Fine-grained	Coarse-grained student - S^c (ours)	0.574	4	0.812	0.967	4	0.047	0.868	4	1.920	0.636	4	0.006
	TMK (Poullot et al., 2015)	0.524	256	119.8	0.975	256	6.949	0.863	256	282.5	0.618	256	1.010
	LAMV (Baraldi et al., 2018)	0.619	256	167.2	0.975	256	9.703	0.880	256	394.5	0.620	256	1.410
	TCA _f (Shao et al., 2021)	0.877	438	36.15	0.990	596	2.097	–	–	–	0.603*	932	0.228
	ViSIL (Kordopatis-Zilos et al., 2019b)	0.899	15124	451.9	0.992	20111	24.26	0.881	2308	319.8	0.658	31457	5.718
Re-ranking	Fine-grained att. student - S_A^f (ours)	0.921	2016	149.1	0.992	2682	8.260	0.902	308	271.8	0.651	4194	1.506
	Fine-grained bin. student - S_B^f (ours)	0.909	63	146.9	0.991	84	8.129	0.891	10	266.5	0.640	131	1.487
	TMK+QE (Poullot et al., 2015)	0.580	256	239.6	0.976	256	13.90	0.774	256	576.0	0.648	256	2.020
	LAMV+QE (Baraldi et al., 2018)	0.659	256	334.4	0.984	256	19.41	0.786	256	766.0	0.653	256	2.820
	DnS _A ^{5%} (ours)	0.874	2020	8.267	0.969	2686	0.463	0.895	312	15.41	0.594	4198	0.081
	DnS _B ^{5%} (ours)	0.862	67	8.154	0.967	88	0.456	0.884	14	15.14	0.584	135	0.080
	DnS _A ^{30%} (ours)	0.913	2020	45.55	0.974	2686	2.528	0.902	312	83.36	0.646	4198	0.458
	DnS _B ^{30%} (ours)	0.900	67	44.87	0.974	88	2.489	0.894	14	81.76	0.634	135	0.452

Bold values indicate the best performing method in the corresponding dataset/subset

*Corresponding results are on different dataset split

only 5% of the dataset with the fine-grained students. However, on the other two datasets, i.e., CC_WEB_VIDEO and EVVE, the re-ranking has negative effects on performance. A possible explanation for this might be that the performance of the coarse- and fine-grained students is very close, especially on the EVVE dataset. Also, this dataset consists of longer videos than the rest, which may impact the selection process. Nevertheless, the performance drop on these two datasets is mitigated when 30% of the dataset is sent to the fine-grained students for re-ranking; while on the FIVR-200K and SVD, the DnS method reaches the performance of the corresponding fine-grained students, or it even outperforms them, i.e., $\text{DnS}_B^{30\%}$ outperforms S_B^f on SVD dataset.

Additionally, Table 10 displays the storage and time requirements and the reference performance of the proposed method on each dataset. In comparison, we include the video retrieval methods that are implemented with the same features and run on GPU. For FIVR-200K and CC_WEB_VIDEO datasets, we display the DSVR and cc_web_c^* runs, respectively. We have excluded the TN and DP methods, as they have been implemented on CPU and their transfer to GPU is non-trivial. Also, the requirements of the TCA runs from Shao et al. (2021) are approximated based on features of the same dimensionality. All times are measured on a Linux machine with the Intel i9-7900X CPU and an Nvidia 2080Ti GPU.

First, the individual students are compared against the competing methods in their corresponding category. The fine-grained binarization student has the lowest storage requirements among the fine-grained approaches on all datasets, having 240 times lower storage requirements than the ViSiL teacher. The fine-grained attention student needs the second-highest requirements in terms of space, but still, it needs 7.5 times less than ViSiL, achieving considerably better retrieval performance on two out of four evaluation datasets. However, the required retrieval time is high for all fine-grained approaches, especially in comparison with the coarse-grained ones. The coarse-grained student, which employs global vectors, has high storage requirements compared to the hashing and BoW methods that need notably lower storage space. In terms of time, all coarse-grained methods need approximately the same on all datasets, which is several orders of magnitude faster than the fine-grained ones.

Second, we benchmark our DnS approach with the two fine-grained students and two dataset percentages sent for refinement. An excellent trade-off between time and performance comes with the $\text{DnS}_B^{5\%}$ offering an acceleration of more than 17 times in comparison to the fine-grained students, at a small cost in terms of performance when 5% is used. Combined with the fine-grained binarization student, on FIVR-200K, it offers 55 times faster retrieval and 240 times lower storage requirements compared to the original

ViSiL teacher providing comparable retrieval performance, i.e., 0.041 relative drop in terms of mAP. The performance of the DnS increases considerably when 30% of the video pairs are sent for re-ranking, outperforming the ViSiL on two datasets with considerable margins. However, this performance improvement comes with a corresponding increase in the retrieval time.

6 Conclusion

In this paper, we proposed a video retrieval framework based on Knowledge Distillation that addresses the problem of performance-efficiency trade-off focused on large-scale datasets. In contrast to typical video retrieval methods that rely on either a high-performance but resource demanding fine-grained approach or a computationally efficient but low-performance coarse-grained one, we introduced a Distill-and-Select approach. Several student networks were trained via a Teacher–Student setup at different performance-efficiency trade-offs. We experimented with two fine-grained students, one with a more elaborate attention mechanism that achieves better performance and one using a binarization layer offering very high performance with significantly lower storage requirements. Additionally, we trained a coarse-grained student that provides very fast retrieval with low storage requirements but at a high cost in performance. Once the students were trained, we combined them using a selector network that directs samples to the appropriate student in order to achieve high performance with high efficiency. It was trained based on the similarity difference between a coarse-grained and a fine-grained student so as to decide at query-time whether the similarity calculated by the coarse-grained one is reliable or the fine-grained one needs to be applied. The proposed method has been benchmarked to a number of content-based video retrieval datasets, where it improved the state-of-art in several cases and achieved very competitive performance with a remarkable reduction of the computational requirements.

The proposed scheme can be employed with several setups based on the requirements of the application. For example, when small-scale databases are involved, with no strict storage space and computational time restrictions, the fine-grained attention student could be employed since it achieves the best retrieval performance. On the other hand, for mid-scale databases, where the storage requirements increase, the fine-grained binarization student would be a reasonable option since it achieves very high retrieval performance with remarkable reduction of storage space requirements. Finally, for large-scale databases, where both storage space and computation time are an issue, the combination of fine-grained binarization student and the coarse-grained student with the

selector network would be an appropriate solution that offers high retrieval performance and high efficiency.

In the future, we plan to investigate alternatives for the better selection and re-ranking of video pairs based on our selector network by exploiting the ranking of videos derived from the two students. Also, we will explore better architectural choices for the development of the coarse-grained student to further improve the system's scalability with little compromises in retrieval performance.

Acknowledgements This work has been supported by the projects MediaVerse and AI4Media, partially funded by the European Commission under Contract Nos. 957252 and 951911, respectively, and DECSTER funded by EPSRC under Contract No. EP/R025290/1.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., & Sivic, J. (2016). NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*
- Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv preprint [arXiv:1607.06450](https://arxiv.org/abs/1607.06450)
- Baraldi, L., Douze, M., Cucchiara, R., & Jégou, H. (2018). LAMV: Learning to align and match videos with kernelized temporal layers. In *Proceedings of the IEEE conference on computer vision and pattern recognition*
- Bhardwaj, S., Srinivasan, M., & Khapra, M. M. (2019). Efficient video classification using fewer frames. In *Proceedings of the IEEE conference on computer vision and pattern recognition*
- Bishay, M., Zoumpourlis, G., & Patras, I. (2019). TARN: Temporal attentive relation network for few-shot and zero-shot action recognition. In *Proceedings of the British machine vision conference*
- Cai, Y., Yang, L., Ping, W., Wang, F., Mei, T., Hua, X. S., & Li, S. (2011). Million-scale near-duplicate video retrieval system. In *Proceedings of the ACM international conference on multimedia*. ACM.
- Cao, Z., Long, M., Wang, J., & Yu, P. S. (2017). HashNet: Deep learning to hash by continuation. In *Proceedings of the IEEE international conference on computer vision*
- Chou, C. L., Chen, H. T., & Lee, S. Y. (2015). Pattern-based near-duplicate video retrieval and localization on web-scale videos. *IEEE Transactions on Multimedia*, 17(3), 382–395.
- Chum, O., Philbin, J., Sivic, J., Isard, M., & Zisserman, A. (2007). Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proceedings of the IEEE international conference on computer vision*
- Crasto, N., Weinzaepfel, P., Alahari, K., & Schmid, C. (2019). Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*
- Deng, J., Pan, Y., Yao, T., Zhou, W., Li, H., & Mei, T. (2019). Relation distillation networks for video object detection. In *Proceedings of the IEEE international conference on computer vision*
- Douze, M., Revaud, J., Schmid, C., & Jégou, H. (2013). Stable hyperpooling and query expansion for event detection. In *Proceedings of the IEEE international conference on computer vision* (pp 1825–1832).
- Douze, M., Jégou, H., & Schmid, C. (2010). An image-based approach to video copy detection with spatio-temporal post-filtering. *IEEE Transactions on Multimedia*, 12(4), 257–266.
- Feng, Y., Ma, L., Liu, W., Zhang, T., & Luo, J. (2018). Video re-localization. In *Proceedings of the European conference on computer vision*
- Gao, Z., Hua, G., Zhang, D., Jojic, N., Wang, L., Xue, J., & Zheng, N. (2017). ER3: A unified framework for event retrieval, recognition and recounting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*
- Garcia, N. C., Morerio, P., & Murino, V. (2018). Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European conference on computer vision*
- Gong, Y., Lazebnik, S., Gordo, A., & Perronnin, F. (2012). Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12), 2916–2929.
- Gordo, A., Radenovic, F., & Berg, T. (2020). Attention-based query expansion learning. In *Proceedings of the European conference on computer vision*
- Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, 129, 1789–1819.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. In *Proceedings of the international conference on neural information processing systems*
- Huang, Z., Shen, H. T., Shao, J., Cui, B., & Zhou, X. (2010). Practical online near-duplicate subsequence detection for continuous video streams. *IEEE Transactions on Multimedia*, 12(5), 386–398.
- Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the international conference on machine learning*
- Jégou, H., & Chum, O. (2012). Negative evidences and co-occurrences in image retrieval: The benefit of pca and whitening. In *Proceedings of the European conference on computer vision* (pp 774–787). Springer.
- Jiang, Q. Y., He, Y., Li, G., Lin, J., Li, L., & Li, W. J. (2019). SVD: A large-scale short video dataset for near-duplicate video retrieval. In *Proceedings of the IEEE international conference on computer vision*
- Jiang, Y. G., Jiang, Y., & Wang, J. (2014). VCDB: A large-scale database for partial copy detection in videos. In *Proceedings of the European conference on computer vision* (pp. 357–371). Springer.
- Jiang, Y. G., & Wang, J. (2016). Partial copy detection in videos: A benchmark and an evaluation of popular methods. *IEEE Transactions on Big Data*, 2(1), 32–42.
- Kingma, D., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the international conference on learning representations*.
- Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., & Kompatsiaris, I. (2017a). Near-duplicate video retrieval by aggregating intermedi-

- ate cnn layers. In *Proceedings of the international conference on multimedia modeling* (pp. 251–263). Springer.
- Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., & Kompatsiaris, I. (2017b). Near-duplicate video retrieval with deep metric learning. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 347–356). IEEE.
- Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., & Kompatsiaris, I. (2019a). FIVR: Fine-grained incident video retrieval. *IEEE Transactions on Multimedia*, 21, 2638–2652.
- Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., & Kompatsiaris, I. (2019b). ViSiL: Fine-grained spatio-temporal video similarity learning. In *Proceedings of the IEEE international conference on computer vision*.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the international conference on neural information processing systems*.
- Lassance, C., Bontou, M., Hacene, G. B., Gripon, V., Tang, J., & Ortega, A. (2020). Deep geometric knowledge distillation with graphs. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing*.
- Lee, J., Abu-El-Haija, S., Varadarajan, B., & Natsev, A. (2018). Collaborative deep metric learning for video understanding. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery & data mining*.
- Lee, H., Lee, J., Ng, J. Y. H., & Natsev, P. (2020). Large scale video representation learning via relational graph clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Li, Q., Jin, S., & Yan, J. (2017). Mimicking very efficient network for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Liang, D., Lin, L., Wang, R., Shao, J., Wang, C., & Chen, Y. W. (2019). Unsupervised teacher–student model for large-scale video retrieval. In *Proceedings of the IEEE international conference on computer vision workshops*.
- Liang, S., & Wang, P. (2020). An efficient hierarchical near-duplicate video detection algorithm based on deep semantic features. In *Proceedings of the international conference on multimedia modeling*.
- Liao, K., Lei, H., Zheng, Y., Lin, G., Cao, C., Zhang, M., & Ding, J. (2018). IR feature embedded bof indexing method for near-duplicate video retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(12), 3743–3753.
- Liong, V. E., Lu, J., Tan, Y. P., & Zhou, J. (2017). Deep video hashing. *IEEE Transactions on Multimedia*, 19(6), 1209–1219.
- Liu, Y., Cao, J., Li, B., Yuan, C., Hu, W., Li, Y., & Duan, Y. (2019). Knowledge distillation via instance relationship graph. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Liu, H., Zhao, Q., Wang, H., Lv, P., & Chen, Y. (2017). An image-based near-duplicate video retrieval and localization using improved edit distance. *Multimedia Tools and Applications*, 76(22), 24435–24456.
- Luo, Z., Hsieh, J. T., Jiang, L., Niebles, J. C., & Fei-Fei, L. (2018). Graph distillation for action detection with privileged modalities. In *Proceedings of the European conference on computer vision*.
- Markatopoulou, F., Galanopoulos, D., Mezaris, V., & Patras, I. (2017). Query and keyframe representations for ad-hoc video search. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval* (pp. 407–411).
- Markatopoulou, F., Mezaris, V., & Patras, I. (2018). Implicit and explicit concept relations in deep neural networks for multi-label video/image annotation. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(6), 1631–1644.
- Miech, A., Laptev, I., & Sivic, J. (2017). Learnable pooling with context gating for video classification. arXiv preprint [arXiv:1706.06905](https://arxiv.org/abs/1706.06905)
- Pan, B., Cai, H., Huang, D. A., Lee, K. H., Gaidon, A., Adeli, E., & Niebles, J. C. (2020). Spatio-temporal graph for video captioning with knowledge distillation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Park, W., Kim, D., Lu, Y., & Cho, M. (2019). Relational knowledge distillation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., & Antiga, L., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of the international conference on neural information processing systems*.
- Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., & Zhang, Z. (2019). Correlation congruence for knowledge distillation. In *Proceedings of the IEEE international conference on computer vision*.
- Piergiorgio, A., Angelova, A., & Ryoo, M.S. (2020). Evolving losses for unsupervised video representation learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Pouillot, S., Tsukatani, S., Phuong Nguyen, A., Jégou, H., & Satoh, S. (2015). Temporal matching kernel with explicit feature maps. In *Proceedings of the ACM international conference on multimedia*.
- Revaud, J., Douze, M., Schmid, C., & Jégou, H. (2013). Event retrieval in large video collections with circulant temporal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2459–2466). IEEE.
- Shao, J., Wen, X., Zhao, B., & Xue, X. (2021). Temporal context aggregation for video retrieval with contrastive learning. In *Proceedings of the IEEE winter conference on applications of computer vision*.
- Shmelkov, K., Schmid, C., & Alahari, K. (2017). Incremental learning of object detectors without catastrophic forgetting. In *Proceedings of the IEEE international conference on computer vision*.
- Sivic, J., & Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Song, J., Yang, Y., Huang, Z., Shen, H. T., & Hong, R. (2011). Multiple feature hashing for real-time large scale near-duplicate video retrieval. In *Proceedings of the 19th ACM international conference on multimedia*.
- Song, J., Yang, Y., Huang, Z., Shen, H. T., & Luo, J. (2013). Effective multiple feature hashing for large-scale near-duplicate video retrieval. *IEEE Transactions on Multimedia*, 15(8), 1997–2008.
- Song, J., Zhang, H., Li, X., Gao, L., Wang, M., & Hong, R. (2018). Self-supervised video hashing with hierarchical binary auto-encoder. *IEEE Transactions on Image Processing*, 27(7), 3210–3221.
- Stroud, J., Ross, D., Sun, C., Deng, J., & Sukthankar, R. (2020). D3d: Distilled 3d networks for video action recognition. In *Proceedings of the IEEE winter conference on applications of computer vision*.
- Tan, H. K., Ngo, C. W., Hong, R., & Chua, T. S. (2009). Scalable detection of partial near-duplicate videos by visual-temporal consistency. In *Proceedings of the ACM international conference on Multimedia*.
- Tavakolian, M., Tavakoli, H. R., & Hadid, A. (2019). AWSd: Adaptive weighted spatiotemporal distillation for video representation. In *Proceedings of the IEEE international conference on computer vision*.
- Thoker, F. M., & Gall, J. (2019). Cross-modal knowledge distillation for action recognition. In *Proceedings of the IEEE international conference on image processing*.
- Tolias, G., Sicre, R., & Jégou, H. (2016). Particular object retrieval with integral max-pooling of cnn activations. In *Proceedings of the international conference on learning representations*.
- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2020). Training data-efficient image transformers & distillation through attention. arXiv preprint [arXiv:2012.12877](https://arxiv.org/abs/2012.12877)

- Tung, F., & Mori, G. (2019). Similarity-preserving knowledge distillation. In *Proceedings of the IEEE international conference on computer vision*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the international conference on neural information processing systems*
- Wang, L., Bao, Y., Li, H., Fan, X., & Luo, Z. (2017). Compact cnn based video representation for efficient video copy detection. In *Proceedings of the international conference on multimedia modeling*.
- Wang, K. H., Cheng, C. C., Chen, Y. L., Song, Y., & Lai, S. H. (2021). Attention-based deep metric learning for near-duplicate video retrieval. In *Proceedings of the international conference on pattern recognition*.
- Wu, X., Hauptmann, A. G., & Ngo, C. W. (2007). Practical elimination of near-duplicates from web video search. In *Proceedings of the ACM international conference on multimedia*.
- Xie, Q., Luong, M. T., Hovy, E., & Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*
- Yalniz, I. Z., Jegou, H., Chen, K., Paluri, M., & Mahajan, D. (2019). Billion-scale semi-supervised learning for image classification. arXiv preprint [arXiv:1905.00546](https://arxiv.org/abs/1905.00546)
- Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the conference of the North American chapter of the association for computational linguistics: Human language technologies*.
- Yang, Y., Tian, Y., & Huang, T. (2019). Multiscale video sequence matching for near-duplicate detection and retrieval. *Multimedia Tools and Applications*, 78(1), 311–336.
- Yuan, L., Wang, T., Zhang, X., Tay, F. E., Jie, Z., Liu, W., & Feng, J. (2020). Central similarity quantization for efficient image and video retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Zhang, Z., Shi, Y., Yuan, C., Li, B., Wang, P., Hu, W., & Zha, Z. J. (2020). Object relational graph with teacher-recommended learning for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Zhang, C., & Peng, Y. (2018). Better and faster: Knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification. In *Proceedings of the international joint conference on artificial intelligence*.
- Zhao, Z., Chen, G., Chen, C., Li, X., Xiang, X., Zhao, Y., & Su, F. (2019). Instance-based video search via multi-task retrieval and re-ranking. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.