International Journal of Multimedia Information Retrieval (2023) 12:32
https://doi.org/10.1007/s13735-023-00298-1

REGULAR PAPER

SPSD: Similarity-preserving self-distillation for video–text retrieval

Jiachen Wang[1] · Yan Hua[2] · Yingyun Yang[2] · Hongwei Kou[2]

Received: 8 March 2023 / Revised: 25 June 2023 / Accepted: 1 August 2023 / Published online: 1 September 2023

Abstract

Most of existing methods solve cross-modal video and text retrieval via coarse-grained similarity computation based on global representations or fine-grained cross-modal interaction. The former misses sufficient information, while the latter suffers from inferior efficiency in inference. Furthermore, hierarchical features of transformer have not been fully utilized in cross-modal contrastive learning. In this paper, we propose similarity-preserving self-distillation method (SPSD) to achieve video and text alignment by cross-granularity and cross-layer ways. For cross-granularity self-distillation, fine-grained cross-modal similarity based on video and text token-wise interaction is transferred to coarse-grained similarity based on global video and text representations. To utilize hierarchical features of deep video and text transformer encoders, we propose cross-layer self-distillation by regarding cross-modal similarity based on semantic features as teacher to provide soft label for the similarity learning based on low-level features. Besides, we construct hierarchical contrastive loss and cross-granularity self-distillation loss at both feature and semantic levels for training transformer-based video and text encoders. SPSD sufficiently utilizes the fine-grained cross-modal interaction and hierarchical transformer features by generating distillation signals through network itself in training stage. In retrieval inference, cross-modal similarity computation between video and text is based on semantic-level global embeddings. Our SPSD achieves outstanding performance for video–text retrieval on MSRVTT, ActivityNet and LSMDC datasets. Our code is available at https://github.com/Macro-1998/SPSD/.

Keywords Video–text alignment · Cross-modality · Knowledge distillation · Transformer · Contrastive learning

1 Introduction

As the development of short video applications, video has become one of the most important media forms for people to obtain information. Video–text retrieval has attracted increasing attention. Due to the success of Transformer [1] and Bert [2] in natural language processing field, there has been a lot of transformer-based vision-language alignment models for cross-modal retrieval [3–7]. Existing approaches could be roughly categorized as global embedding-based [3, 4, 8, 9] and fine-grained interaction-based methods [5, 7, 10–12]. Embedding-based methods usually lie in global contrastive alignment of videos and texts. The embedding learning for two modalities can be decoupled, and the representations for test data can be pre-computed offline. Thus, embedding-based methods are efficient when retrieval inference is carried out. These methods model coarse cross-modal interaction via the similarity of the global representations of video and text.

To explore fine-grained interaction between heterogeneous data, a lot of studies [7, 11, 13–17] are proposed. Most of them [7, 11, 13–15] fed visual and linguistic elements (usually patches from image and words from sentence) simultaneously into a transformer-based network for cross-modal interaction learning. This way could granularly align and aggregate visual and linguistic clues. In retrieval inference, pairwise video and text are required as input to network for computing their relevance score. In addition, a late interaction architecture is proposed to firstly compute the similarities between tokens of video and text elements, and then, the summation [16] or mean [17] of token-wise similarities is calculated as the relevance score for the video and text. These methods suffer from inferior efficiency in inference.

In this paper, we aim to achieve global embedding-based inference; meanwhile, we expect the feature embeddings

✉ Yan Hua
huayan@cuc.edu.cn

[1] Zhihu, A5 Xueyuan Road, Haidian District, Beijing 100083, China

[2] School of Information and Communication Engineering, Communication University of China, 1 Dingfuzhuang East St, Chaoyang District, Beijing 100024, China

could obtain the characteristic of cross-modal fine-grained interaction. Inspired by knowledge distillation [18–21], we propose cross-granularity self-distillation method by distilling the token-wise fine-grained similarity of video and text into coarse-grained similarity relationship based on global embeddings. The fine-grained cross-modal similarity is considered as soft label to guide the learning of global embeddings, and thus, global features are actually enforced to obtain the performance of fine-grained interaction. In retrieval stage, we utilize global embeddings of video and text for similarity computation and ranking to achieve efficient retrieval.

According to the attention allocated characteristics of different transformer layers, the features in different layers focus on different views [4, 7, 22–25]. For example, local syntax is encoded at the lower layers and longer range semantics at the upper layers [26]. A recent visual-language learning method [4] explored hierarchical features by adding the feature-level (the first layer) and semantic-level (the last layer) contrastive loss to learn the transformer-based encoders. We consider discriminating the binary relationship (similar and dissimilar) between cross-modal data in contrastive learning may be too strict and difficult to low-level features. To alleviate this problem and further explore hierarchical features, we propose cross-layer self-distillation method by regarding semantic-level similarity between video and text as soft label and distilling it to the cross-modal similarity based on low-level features. In this way, the model could learn similarity-oriented low-level features for cross-modal retrieval.

In this paper, we propose similarity-preserving self-distillation (SPSD) method for video and text alignment with cross-granularity and cross-layer self-distillation ways. Figure 1 shows the framework. Two transformer-based encoding modules are used to extract video and text features. The global embeddings of video and text are used to compute coarse-grained similarity. Meanwhile, the token features of video and text are utilized to get token-wise fine-grained similarity by late interaction way. Specifically, we design a token screening module to adaptively select important tokens for fine-grained similarity computation. To mine hierarchical capacity of transformer encoders, we perform cross-granularity self-distillation with semantic-level and feature-level representations. The cross-granularity and cross-layer self-distillation losses are all based on KL divergence. Together with the self-distillation losses, we employ InfoNCE [27] to construct contrastive loss with hierarchical features for training the model.

The cross-granularity self-distillation and cross-layer self-distillation both generate distillation signals through the network itself to help the encoders of video and text learn better. They are applied in the training stage, so they will not cause additional computational overhead in retrieval inference. Experiments on three public datasets show the effectiveness of SPSD.

2 Related work

2.1 Cross-modal interaction learning

Existing approaches for cross-modal retrieval address fine-grained interaction between video and text generally by two ways, feeding video and text together into a single stream network [7, 10–15, 28–30] or modeling the interplay based on dual stream network [5, 6, 16, 17, 31–35]. Our method is based on dual stream network. SCAN [32] discovers the latent alignments using both image regions and words in a sentence as context and infers image-text similarity. T2VLAD [31] aggregates the multi-modal video sequences and text features with a set of shared semantic centers, and then, the local cross-modal similarities are computed between the video feature and text feature within the same center. MMT [5] computes the video-caption similarity as a weighted sum of each expert's video-caption similarity. FILIP [17] achieves a cross-modal late interaction mechanism with token-wise maximum similarity between visual and textual tokens. In CRET [33], the text and video embeddings are aligned by learned transformer decoder centers. In recent CMMT model [34], each raw video denotes a pseudo-video class and a cross-modal fine-grained classification task is conducted where the text queries are classified with pseudo-video class prototypes. X-pool [35] utilized a scaled dot product attention for a text to attend to its most semantically similar frames, and then, an aggregated video representation is generated conditioned on the text's attention weights over the frames. Jin et al. [36] used coarse-fine-grained parallel attention model and feature fusion module to learn effective video feature representation for video–text retrieval task.

Different from all these methods, we make the similarity of global representations have the ability of fine-grained interaction characteristic by self-distillation learning. The most related work to ours is FILIP [17] and MMT [5]. We adopt the same expert features as MMT and the token-wise fine-grained similarity proposed by FILIP as teacher in our cross-granularity self-distillation learning.

2.2 Hierarchical alignment

A lot of studies have researched the different level features of deep network [22–26] for cross-modal alignment [4, 6, 7, 37, 38] since deep architecture can learn representations that vary with network depth from local syntax encoded at the lower layers to longer range semantics at the upper layers. COOT [38] proposes to align the representations at

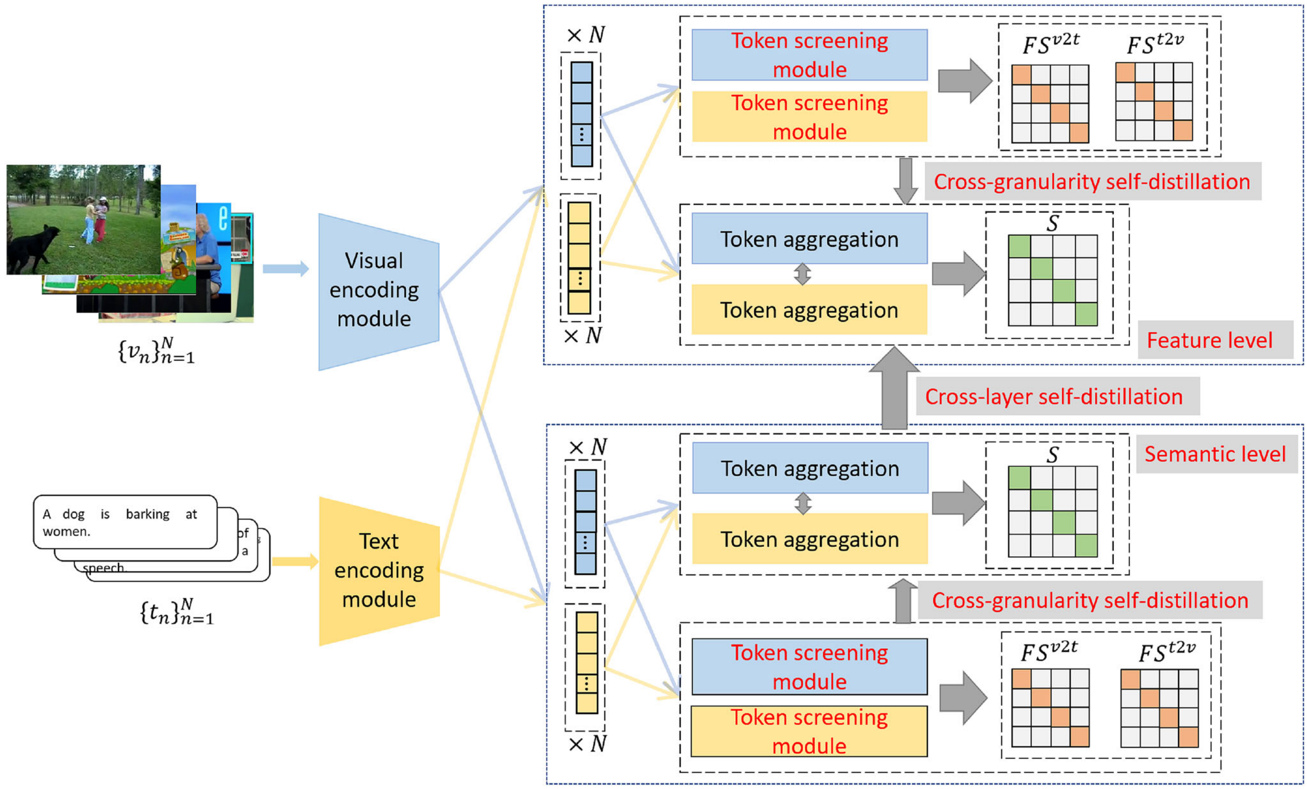

Fig. 1 The framework of SPSD. We propose a similarity-preserving self-distillation method to align video and text. The different layer outputs of visual encoding module and text encoding module are, respectively, utilized to compute feature-level and semantic-level related loss. To get fine-grained similarity in a token-wise interaction way, we design a token screening module to select important tokens for video and text modalities. The fine-grained similarity is then distilled to coarse-grained similarity which is based on global embeddings of video and text. This operation goes on at both feature level and semantic level to form hierarchical cross-granularity self-distillation loss. Besides, cross-layer self-distillation loss is proposed by distilling the semantic-level similarity to feature-level similarity. In this way, the learned semantic representations for video and texts are utilized to compute distances then ranked for cross-modal retrieval

frame–word, clip–sentence and video–paragraph three levels. TACo [6] proposes to construct hierarchical contrastive loss including token-level and sentence-level loss with the output of individual video and text encoders before multi-modal fusion network, and another sentence-level loss after the multi-model fusion network. CrossCLR [37] also utilizes a two-level hierarchy of transformers, where the loss is applied at the clip–sentence level and video–paragraph level. HiT [4] proposes to add feature-level and semantic-level contrastive loss to learn the video and text encoders based on transformer architecture. Ji et al. [39] proposed a step-wise hierarchical alignment network (SHAN) that decomposes image–text matching into multi-step cross-modal reasoning process including local-to-local alignment at fragment level, global-to-local and global-to-global alignment at context level. Jiang et al. [40] explored multi-level cross-modal relationships among video–sentence, clip–phrase, and frame–word for text–video retrieval based on the pre-trained CLIP.

They all utilize the different layer features to construct cross-modal correlation for learning multi-modal encoders. Besides the hierarchical correlation, we propose the cross-layer self-distillation way to take advantage of hierarchical features, i.e., the semantic-level similarity based on the last output of transformer encoders is transferred to the low-level feature learning.

2.3 Knowledge distillation

Knowledge distillation [41] is proposed to transfer the activation of individual example representation from a large teacher network to a small student network. Some studies have shown that transferring the mutual similarity instead of actual representation is beneficial to student representation learning [19, 20, 42–44]. Park et al. [43] proposed to transfer the relational information from teacher to student by distance-wise and angle-wise distillation losses. Tung et al. [19] proposed to guide the training of a student network such that input pairs that produce similar (dissimilar) activations in the teacher network produce similar (dissimilar) activations in the student network. Zhu et al. [20] selected a neighbor example from the teacher space as anchor and encouraged the anchor–student relation to be consistent with the anchor–teacher relation. Tian et al. [44] encouraged the teacher and student to map the same input to close representations and different inputs to distant representations. Li et al. [45] explore the merit of the student model in each time step to guide the training process of the teacher model.

Another line of work is self-knowledge distillation through distilling knowledge within network itself [18, 21, 46]. Zhang et al. [18] proposed to distill the classifier's representations in the deeper portion of the CNN networks into the shallow ones. Hou et al. [46] exploited the activation-based CNN attention maps from its own layers as the distillation targets for its lower layers. Ji et al. [21] introduced an auxiliary self-teacher network to enable the transfer of a refined knowledge to the classifier network. Different from all these methods, our SPSD transfers the fine-grained similarity relationship between video and text to coarse-grained similarity based on global features, which is intra a transformer layer, and transfers high-level features' similarity of video and text to low-level features' similarity, which is cross-two transformer layers.

2.4 Others

Recently, a lot CLIP pre-trained-based models and contrastive learning are studied for video and text retrieval. CLIP4Clip [47] transferred the knowledge of the CLIP model to video-language retrieval in an end-to-end manner. CLIP-ViP [48] utilized a video proxy mechanism to transfer CLIP model to video domain and introduced an omnisource cross-modal learning method to reduce the domain gap between pre-training data and downstream data. Cross-modal adapter [49] is proposed for parameter efficient fine-tuning with a few parameterization layers. Huang et al. [50] proposed the text–video cooperative prompt tuning model to efficient tune the pre-trained CLIP for text–video cross-modal retrieval. Wang et al. [51] proposed a diversity-sensitive contrastive learning loss by adaptive negative pair weighting to capture the fine-grained discrepancies among negative pairs. Better pre-trained model and contrastive learning loss achieve better performance on cross-modal retrieval. Nevertheless, they are not the focus of this paper. For the sake of fairness, we compare with the models using expert features for video as MMT [5] to validate our proposed cross-granularity and cross-layer self-distillation method.

3 Method

As shown in Fig. 1, our model has several key components, cross-granularity self-distillation, token screening module and cross-layer self-distillation. We first introduce the preliminaries and then describe the innovative points.

3.1 Preliminaries

3.1.1 Video encoding module

In our paper, video encoder is implemented by a stack of 4 self-attention layers and fully connected layers as the architecture of the transformer encoder presented in [1, 5]. Inspired by recent work [4, 5, 9, 37], some expert features are firstly extracted for video using pre-trained models such as motion features from S3D trained on Kinetics, audio features extracted using VGGish model trained on YT8M and appearance features from SENet-154 trained on ImageNet. The input of video encoder contains the expert features, embeddings of the expert type and the embeddings of the time in the video when the feature was extracted [5]. Expert features are firstly, respectively, projected to the same dimension 512 by fully connected layers and L_2 normalization.

For a video v, the n-th-type expert feature at k time is denoted as F_k^n, where $n \in [1, N]$ and $N = 7$ is the total kinds of experts as [5]. The global feature for a kind of expert is obtained by max pooling on all times. The expert feature is then represented as,

$$F_v = [F_{agg}^1, F_1^1, \ldots, F_K^1, \ldots, F_{agg}^N, F_1^N, \ldots, F_K^N]. \tag{1}$$

To distinguish different types of expert and the time of the extracted feature, 512-dimensional embeddings of expert type and temporal information are learned as video encoder inputs. They are denoted as,

$$E_v = [E^1, E^1, \ldots, E^1, \ldots, E^N, E^N, \ldots, E^N], \tag{2}$$

$$T_v = [T_{\text{agg}}, T_1, \ldots, T_D, \ldots, T_{\text{agg}}, T_1, \ldots\ldots\ldots\ldots, T_D]. \tag{3}$$

The summation of F_v, E_v and T_v is fed into a 4-layer transformer-based video encoder to learn the video token representations.

3.1.2 Text encoding module

We employ a pre-trained Bert-base-uncased model [2] as the text encoder and fine-tune it. Each word in a text t is embedded into a vector as token embeddings F_t. [CLS] and [END] are placed on the first and last positions. Text Segment Mask M_t is used to indicate the id of input sequence, which is meaningless in our method since only one text is processed every time. Text Position Embedding P_t is used to encode the indexes of word in the text sequence. The final input for text encoder is the sum of F_t, M_t and P_t, which is fed into the pre-trained Bert model to get text token representations.

3.1.3 Token aggregation

For video modality, we utilize mean pooling on all output tokens of video encoding module to obtain the global video representation and apply a linear fully connected layer to project the representation into a vector with the same dimension d_{rep} with text data.

For text modality, the text representation is got by applying mean pooling on all word tokens, the output of text encoding module. A linear fully connected layer is used to project the text representation into the same dimension d_{rep} with video data.

Then, we design a shared linear layer to project video and text to a d-dimensional common space. It should be noted that the expert models of video are fixed, the parameters of text encoding module are fine-tuned, and other parameters are learned from scratch in training stage.

3.1.4 Contrastive loss

For video–text retrieval task, the target of our method is to obtain the global visual and textual embeddings by learning the model parameters. We employ contrastive loss [6, 27, 27, 37] to make the pairwise video and text similar and unmatched samples dissimilar. Given a mini-batch of N video and text pairs $B = \{v_n, t_n\}_{n=1}^{N}$, where v_n and t_n are pairwise video and its text description, we get their common space embeddings by video and text encoders as $\{r_n^v\}_{n=1}^{N}$ and $\{r_n^t\}_{n=1}^{N}$, respectively. All pairs $\{v_i, t_j\}$ with $i \neq j$ are regarded as negative pairs. The similarity matrix $S \in R^{N\times N}$ between a mini-batch examples is computed by inner product of the embeddings, that is,

$$S_{i,j} = (r_i^v)^T r_j^t, \tag{4}$$

which is the similarity of the i-th video and j-th text. The contrastive loss InfoNCE [27] for video–text retrieval on a mini-batch is,

$$L_c^{v2t} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(S_{i,i}/\tau)}{\sum_{j=1}^{N}\exp(S_{i,j}/\tau)}, \tag{5}$$

where τ is a temperature hyper-parameter [52]. Similarly, the loss for text–video retrieval on a mini-batch is,

$$L_c^{t2v} = -\frac{1}{N}\sum_{j=1}^{N}\log\frac{\exp(S_{j,j}/\tau)}{\sum_{i=1}^{N}\exp(S_{i,j}/\tau)}. \tag{6}$$

The two losses are combined as,

$$L_c = \frac{1}{2}(L_c^{v2t} + L_c^{t2v}). \tag{7}$$

By optimizing the contrastive loss, the similarities between pairwise video and text embeddings in a mini-batch are maximized and that of the unmatched sample embeddings are minimized.

3.2 Cross-granularity self-distillation

The every layer outputs of video and text transformer encoders contain token features. For video, output tokens contain the information of expert feature at a certain time. For text, output tokens contain the information of every word in the text. We employ token-wise late interaction [16, 17] to obtain the fine-grained cross-modal similarity values. To guarantee the effectiveness of late interaction, we propose token screening module to select important tokens for video and text alignment. At the same time, the coarse-grained similarity between video and text is obtained according to inner product of global embeddings, as Eq. (4). In training stage, the fine-grained similarity is used as soft label to optimize coarse-grained similarity by the way of knowledge distillation. In retrieval stage, cross-modal matching depends on the inner product of global embeddings. In the way of cross-granularity self-distillation, fine-grained global representations of video and text are learned to achieve retrieval.

3.2.1 Token screening module

In [16, 17], all tokens are participated in fine-grained similarity computation. It is time-consuming and sensitive to noisy token. For example, given text “a dog is barking at women,” three words “dog,” “barking” and “women” are obviously critical to cross-modal matching, and thus, they should be focused on rather than others. Nonsignificant tokens will hinder the reliability of cross-modal alignment. Therefore, we propose token screening networks for video and text, respectively, to adaptively determine which tokens would participate in cross-modal fine-grained interaction according to the token features of video and text encoders.

Figure 2 shows the structure of our proposed token screening module. We denote the output of transformer encoders (before mean pooling) as $X = \{x_i | i \in [1, n]\}$, which could be video or text tokens, and n is the number of tokens. The figure shows 5 tokens as example. The n token vectors are successively fed into linear layer, ReLU, linear layer and softmax layer to get normalized n probabilities, which are regarded as the tokens’ importance measurement. Denote the ratio of screening token as r, then the number of selected tokens is $k = \lfloor n \times r \rfloor$, where $0 \le r \le 1$ and $\lfloor\cdot\rfloor$ represents taking the integer portion. Our token screening module selects the most important k tokens according to the adaptive probabilities, and the selected token features are used to compute the fine-grained cross-modal similarity.

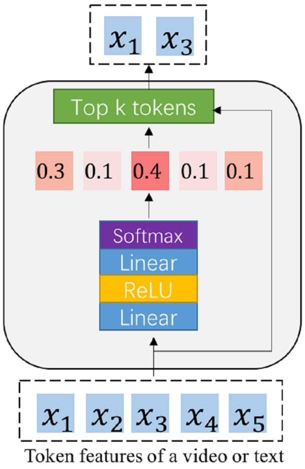

Fig. 2 The illustration of token screening module. Top k token features are adaptively selected according to input token features themselves

3.2.2 Fine-grained similarity matrix

After token screening module, the selected token features of video and text are projected to same dimension d_{rep} by a linear layer, respectively. Then, they are together projected to a d-dimensional common subspace by a shared linear layer, which is the same way as the processing of global embeddings. The features of i-th video and j-th text are denoted as $R^v_i \in R^{n_1 \times d}$ and $R^t_j \in R^{n_2 \times d}$, respectively, where n_1 and n_2 are the selected number of tokens by video and text token screening module, respectively. As [17], for a visual token $[R^v_i]_k$, its similarity with text is the largest one among the token with all textual tokens $[R^t_j]^{n_2}_{r=1}$. The token-wise fine-grained similarity between the video and text is the average on all video tokens. Then, the similarity value of the i-th video to j-th text is formulated as,

$$FS^{v2t}_{i,j} = \frac{1}{n_1}\sum_{k=1}^{n_1} [R^v_i]_k{}^T [R^t_j]_{m^v_k}, \tag{8}$$

where $m^v_k = argmax_{0 \leq r \leq n_2}[R^v_i]_k{}^T[R^t_j]_r$. Similarity, the similarity of the j-th text to i-th video is,

$$FS^{t2v}_{i,j} = \frac{1}{n_2}\sum_{k=1}^{n_2} [R^v_i]_{m^t_k}{}^T [R^t_j]_k, \tag{9}$$

where $m^t_k = argmax_{0 \leq r \leq n_1}[R^v_i]_r{}^T[R^t_j]_k$. In this way, we can obtain the fine-grained similarity matrix $FS^{v2t} \in R^{N \times N}$ and $FS^{t2v} \in R^{N \times N}$ for video–text and text–video retrieval with a batch cross-modal samples. It should be noted that $FS^{v2t} \neq FS^{t2v}$.

3.2.3 Cross-granularity loss

In the retrieval based on token-wise interaction method [17], all token features need to be stored and the fine-grained similarity is computed as above. We expect retrieval is achieved by inner product of vectors yet has the effectiveness of fine-grained cross-modal alignment. In this paper, we propose novel similarity-preserving self-distillation approach. The fine-grained similarity matrix is regarded as teacher and coarse-grained similarity matrix as student. The knowledge is transferred from the teacher to the student by minimizing their difference with KL divergence. Given two distributions $P = \{p_i \mid i \in [1, m]\}$ and $Q = \{q_i \mid i \in [1, m]\}$, the formulation of KL divergence is as follows,

$$D_{KL}[P||Q] = \sum_{i=1}^{m} p_i[log(p_i) - log(q_i)]. \tag{10}$$

The student similarity for a batch data is S, as computed in Eq. (4). Teacher similarity matrices for video–text and text–video retrieval are FS^{v2t} and FS^{t2v}, respectively, as shown in Eqs. (8) and (9). The cross-granularity self-distillation loss for video–text retrieval is defined as,

$$L^{v2t}_{cg} = \frac{1}{N}\sum_{i=1}^{N} D_{KL}[s(FS^{v2t}_i)/\tau||s(S_i)], \tag{11}$$

where FS^{v2t}_i and S_i are, respectively, the i-th row of the similarity matrix that is the similarity values between the i-th video and all texts. τ is the temperature scaling parameter. The s operation means softmax, used to normalize the row of similarity matrix. Similarly, the distillation loss for text–video retrieval is defined as,

$$L^{t2v}_{cg} = \frac{1}{N}\sum_{i=1}^{N} D_{KL}[s(FS^{t2v}_i)/\tau||s(S_{:,i})], \tag{12}$$

where FS^{t2v}_i is the i-th row of the fine-grained similarity matrix. $S_{:,i}$ is the i-th column of the coarse-grained similarity matrix, which represents the similarities between all videos with the i-th text. The whole cross-granularity self-distillation is then formulated as,

$$L_{cg} = L^{v2t}_{cg} + L^{t2v}_{cg}. \tag{13}$$

By optimizing L_{cg}, the student coarse-grained similarity is preserved consistent with the teacher fine-gained similarity. Thus, the global embeddings for coarse-grained video and text alignment could learn the fine-grained interaction by the similarity-preserving self-distillation way. The similarity computations of two granularities are based on the representations from the same transformer layer, and the cross-granularity loss is intra-layer self-distillation.

3.3 Cross-layer self-distillation

Different layers of deep network usually focus on features with different degrees of abstraction [4, 6, 37, 38]. For example, low-level layer tends to encode local visual content and basic syntax, while high-level layer tends to capture more complex semantics and obtain more abstract representation. In other words, high-level features are more appropriate for semantic task than low-level features. It is too difficult for low-level features to achieve strict pairwise judgment. In this paper, we propose cross-layer self-distillation to explore hierarchical features by using semantic layer similarity providing soft label for feature layer alignment.

Specially, we employ the last-layer representations of video and text encoding modules to compute the similarity as teacher, and the similarity of the first-layer representations as student. The computation is as in Eq. (4). For a mini-batch B, the teacher similarity matrix is denoted as S^h and student similarity matrix as S^l. We obtain the distillation loss by KL divergence for video–text retrieval as follows,

$$L_{cl}^{v2t} = \frac{1}{N}\sum_{i=1}^{N} D_{KL}[s(S_i^h/\tau)||s(S_i^l)], \tag{14}$$

where S_i^h and S_i^l are the i-th row of similarity matrix, respectively, representing the similarities between the i-th video and all texts. The distillation loss for text–video retrieval is defined as,

$$L_{cl}^{t2v} = \frac{1}{N}\sum_{i=1}^{N} D_{KL}[s(S_{:,i}^h/\tau)||s(S_{:,i}^l)], \tag{15}$$

where $S_{:,i}^h$ and $S_{:,i}^l$ are the i-th column of matrix, respectively, representing the similarities between all videos with the i-th text. And the whole cross-layer self-distillation loss is then formulated as,

$$L_{cl} = L_{cl}^{v2t} + L_{cl}^{t2v}. \tag{16}$$

By optimizing the loss function, the student similarity matrix is preserved consistent with the teacher. That is, semantic layer relationship provides soft label (similarity) for low-level feature alignment, which helps make the learned hierarchical features more suitable for video and text retrieval.

3.4 Objective function

Our objective function consists of three components, feature level, semantic level and cross-layer. The third one is the above cross-layer self-distillation L_{cl}, as Eq. (16). Feature-level loss includes two parts, contrastive loss and cross-granularity self-distillation, respectively, computed by Eqs. (7) and (13) based on the first-layer outputs of video and text encoders. They are, respectively, denoted as L_c^f and L_{cg}^f, and then, the formulation of feature-level loss is,

$$L_f = L_c^f + \lambda L_{cg}^f. \tag{17}$$

Semantic-level loss also includes two parts, contrastive loss L_c^s and cross-granularity self-distillation L_{cg}^s based on the last-layer outputs of video and text encoders, and the formulation is,

$$L_s = L_c^s + \lambda L_{cg}^s, \tag{18}$$

where λ is the trade-off parameter for contrastive loss and cross-granularity self-distillation loss. Our final objective function is calculated as follows,

$$L = L_s + \alpha L_f + \gamma L_{cl}, \tag{19}$$

where α and γ are the trade-off parameters for semantic-level, feature-level and cross-layer losses. By optimizing the loss, our method SPSD could adequately take advantage of hierarchical features and fine-grained interactions between video and text tokens to alignment cross-modal data.

4 Experiments

4.1 Datasets and settings

We compare SPSD with state of the arts on three datasets MSRVTT [53], LSMDC [54] and ActivityNet Captions [55]. Ablation experiments are conducted on MSRVTT.

MSRVTT dataset consists of 10000 videos collected from YouTube with 257 queries. The length of each video is about 10–30 s, and each video has 20 manually tagged English sentence descriptions. For the 10000 videos, we refer [30] and divide this dataset into training set with 9000 videos and test set with 1000 videos.

LSMDC dataset contains 118081 short videos truncated from 202 movies. The length of each short video is about 45 s, and each video is equipped with a text caption from the movie script or audio description. The test set consists of 1000 videos, from movies not present in the training set.

ActivityNet Captions dataset consists of 20K YouTube videos temporally annotated with sentence descriptions. We follow the approach of [4], where all the descriptions of a video are concatenated to form a paragraph. The training set has 10,009 videos. We evaluate our video–paragraph retrieval on the “val1” split (4917 videos).

Table 1 Comparison with SOTA on MSRVTT (The bold font indicates the best results)

Methods	Video–text retrieval				Text–video retrieval				Rsum
	R@1	R@5	R1@0	MedR	R@1	R@5	R@10	MedR	
Random	0.2	0.7	1.3	507.0	0.1	0.5	0.8	504.5	3.6
JSFusion [30]	9.5	28.6	40.2	18.0	9.6	29.8	42.1	20.0	159.8
CE [9]	20.9	48.8	62.4	6.0	20.6	50.3	64.0	5.3	267.0
MMT [5]	24.4	56.0	67.8	4.0	24.6	54.0	67.1	4.0	293.9
Support-set [8]	26.6	55.1	67.5	**3.0**	27.4	56.3	67.7	**3.0**	300.6
TACo [6]	–	–	–	–	26.7	54.5	68.2	4.0	–
HiT [4]	28.8	60.3	72.3	**3.0**	27.7	59.2	72.0	**3.0**	320.3
Jin [36]	–	58.3	–	4.0	–	56.5	–	4.0	
SPSD	**29.7**	**61.1**	**73.8**	**3.0**	**29.3**	**60.9**	**72.8**	4.0	**327.6**

Evaluation metrics include R@1, R@5, R@10, R@50, MedR and Rsum. R@K is the percentage of test queries that at least one relevant item is found among the top-K retrieved results. The MedR measures the median rank of correct items in the retrieved ranking list. We also take the sum of all R@K as Rsum to reflect the overall retrieval performance. Larger R@K and Rsum and smaller MedR indicate better retrieval performance.

In training stage, AdamW [56] optimizer is used, the initialization learning rate is set to 5×10^{-5}, and the weight decay is set to 0.01. The learning rate is decayed by a multiplicative factor 0.95 every epoch, and the network is trained for 60 epochs. The size of mini-batch is fixed to 128. In terms of the hyper-parameters, the dimension of video and text representation is $d_{rep} = 512$, and the dimension of shared space for similarity computation is $d = 1024$. Temperature hyper-parameter τ is set to 0.07.

4.2 Comparison with state-of-arts

For a fair comparison, we compare with the similar state-of-the-art methods which also fuse multiple expert features for video. The state-of-art methods include JSFusion [30], CE [9], MMT [5], support-set [8], TACo [6], HiT [4], CrossCLR [37] and Jin [36]. CE and support-set achieve retrieval based on global representations. JSFusion, MMT and Jin are fine-grained alignment methods for video and text. TACo, HiT and CrossCLR are hierarchical contrastive learning methods. The performances of these methods are from their papers.

The results on MSRVTT dataset are shown in Table 1. We can see that SPSD achieves the best performances at all metrics except that SPSD gets the second place on text-to-video task with MedR. Our method achieves the performance $Rsum = 327.6\%$, which is 7.4% higher than the second place HiT with $Rsum = 320.3\%$. In practical applications, people tend to pay more attention on the top retrieval results. The R@1 performance of SPSD is 1.1% and 1.6% higher than the second place HiT on video-to-text and text-to-video retrieval, respectively. HiT [4] performs hierarchical cross-modal contrastive matching with global features from feature level and semantic level. In comparison, token-wised fine-grained similarity and cross-layer interaction for self-distillation learning explored in our method help our method outperform HiT with most of the evaluation metrics. Our method outperforms other global representation-based methods CE and support-set, fine-grained alignment methods JSFusion, MMT and Jin, and hierarchical contrastive learning method TACo. This further proves that it is effective to align video and text with cross-granularity and cross-layer self-distillation losses.

Table 2 Comparison with SOTA on LSMDC (The bold font indicates the best results)

Methods	Text–video retrieval				Rsum
	R@1	R@5	R@10	MedR	
Random	0.0	0.3	0.9	491.0	1.2
JSFusion [30]	9.1	21.2	34.1	36.0	64.4
CE [9]	11.2	26.9	34.8	25.3	72.9
MMT [5]	13.2	29.2	38.8	21.0	81.2
CrossCLR [37]	15.0	32.5	42.0	18.0	89.5
HiT [4]	14.0	31.2	41.6	18.5	86.8
Jin [36]	–	30.8	–	18.1	–
SPSD	**15.3**	**32.9**	**43.4**	**17.0**	**91.6**

On LSMDC dataset, we only conduct the text-to-video retrieval since most compared methods only provide the results on this retrieval task. The performances are shown in Table 2. We can see that SPSD achieves the performances $R@1 = 15.3\%$, $R@5 = 32.9\%$, $R@10 = 43.4\%$ and $MedR = 17.0$, which are all the best performances among the compared methods. The Rsum performance of SPSD is 2.1% higher than the second place CrossCLR and 4.8% higher than the third place HiT on video-to-text retrieval. HiT [4] performs hierarchical cross-modal contrastive matching with global features from feature level and semantic level. In

Table 3 Comparison with SOTA on ActivityNet (The bold font indicates the best results)

Methods	Video–text retrieval				Text–video retrieval				Rsum
	R@1	R@5	R@50	MedR	R@1	R@5	R@50	MedR	
Random	0.01	0.1	1.02	2548	0.02	0.1	1.02	2458	2.26
CE [9]	17.7	46.4	90.9	6.0	18.2	47.7	91.4	6.0	312.3
Support-set [8]	25.5	57.3	93.5	**3.0**	26.8	58.1	93.5	**3.0**	354.7
MMT [5]	22.9	54.8	93.1	4.0	22.7	54.2	93.2	5.0	340.9
HiT [4]	–	–	–	–	**27.7**	**58.6**	94.7	4.0	–
TACo [6]	–	–	-	–	25.8	56.3	93.8	4.0	–
Jin [36]	–	57.5	–	4.0	-	56.5	–	4.0	–
SPSD	**26.6**	**59.9**	**97.0**	4.0	23.7	56.7	**96.8**	4.0	**360.7**

Table 4 The performances of cross-granularity self-distillation on MSRVTT (The bold font indicates the best results)

λ	Video–text retrieval				Text–video retrieval				Rsum
	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
0	28.3	57.7	70.1	4.0	26.7	56.8	68.3	**4.0**	307.9
3	28.9	57.5	70.2	4.0	27.5	56.5	68.4	**4.0**	309.0
30	27.8	58.7	69.3	4.0	27.4	**57.2**	67.3	**4.0**	307.7
300	**29.0**	**59.6**	**71.9**	**3.0**	**28.1**	56.8	**68.9**	**4.0**	**314.3**
600	28.6	58.1	71.8	4.0	24.4	56.5	68.2	**4.0**	307.6
900	27.2	56.9	68.6	4.0	24.8	54.6	67.3	5.0	299.4
1200	26.0	56.2	68.4	5.0	23.8	52.9	66.1	5.0	293.4

comparison, token-wised fine-grained similarity and cross-layer interaction for self-distillation learning explored in our method help our method outperform HiT. CrossCLR [37] utilizes a two-level hierarchy of transformers, where the loss is applied at the clip/sentence level and at the video/paragraph level. In comparison, fine-grained interaction and hierarchical features are both explored in CrossCLR and our method, and thus, cross-layer learning may be the reason why our model outperforms CrossCLR. With cross-layer self-distillation loss, the low-level features could be better with the soft label provided by semantic layer features. Our method outperforms other global representation-based methods CE, fine-grained alignment methods JSFusion, MMT and Jin. This further proves that it is effective to align video and text with cross-granularity and cross-layer self-distillation.

On ActivityNet dataset, we report the performances with R@1, R@5, R@50 and MedR in Table 3. We can see that the R@1, R@5 and R@50 performances of our method are 26.6%, 59.9% and 97.0% on video-to-text retrieval, respectively, which are much better than others. The R@50 performance of our method is 96.8% for text-to-video retrieval, which is better than 94.7% of the second place HiT. But R@1 and R@5 performances of our method are worse than that of HiT. Further optimization of our method is needed for text-to-video retrieval on ActivityNet dataset. The Rsum of our method is 360.7%, which is better than 354.7% of the second place support-set and 340.9% of the third place MMT. With Rum, our method outperforms all the methods who realized video-to-text and text-to-video retrieval.

Table 5 The performances of token screening module on MSRVTT (The bold font indicates the best results)

r_t	r_v			
	0.25	0.5	0.75	1
0.25	315.1	315.0	312.6	315.7
0.5	313.8	312.0	**318.5**	317.0
0.75	315.0	315.7	314.7	316.4
1	315.2	313.1	313.9	314.3

The computation burden of our method's retrieval process is related to the embedding extraction for query, the embedding dimension and the size of database. The average retrieval time for a query is 0.1447s on MSRVTT, 0.1992s on LSMDC and 0.4428s on ActivityNet dataset, which are test with CPU. It is noted that we set the trade-off hyper-parameters $\lambda = 30, \gamma = 10, \alpha = 1$ of our model on MSRVTT, LSMDC and ActivityNet for comparison with others. The ratios for video and text token screening module are, respectively, $r_v = 0.75$ and $r_t = 0.5$. It shows that the hyper-parameter setting of our model is generalized.

Table 6 The performances of hierarchical contrastive loss on MSRVTT (The bold font indicates the best results)

α	Video–text retrieval				Text–video retrieval				Rsum
	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
1	25.4	57.4	69.5	4.0	26.1	56.7	68.3	4.0	303.4
0.1	27.4	**58.4**	69.8	4.0	27.3	56.8	**69.7**	4.0	**309.4**
0.01	27.4	57.9	69.8	4.0	**27.9**	56.9	69.4	4.0	309.3
0.001	**28.8**	57.4	69.6	4.0	**27.9**	**57.0**	68.2	4.0	308.9
0	28.3	57.7	**70.1**	4.0	26.7	56.8	68.3	4.0	307.9

Table 7 The performances of cross-layer self-distillation on MSRVTT (The bold font indicates the best results)

γ	Video–text retrieval				Text–video retrieval				Rsum
	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
$\gamma = 0$	27.4	58.4	69.8	4.0	27.3	56.8	69.7	4.0	309.4
$\gamma = 0.3$	28.0	59.5	70.4	4.0	28.0	56.4	70.1	4.0	312.4
$\gamma = 3$	**28.6**	**60.7**	70.7	4.0	27.6	**57.8**	69.9	4.0	**315.3**
$\gamma = 30$	27.5	58.2	**71.5**	4.0	**29.2**	56.5	**71.2**	4.0	314.1

Table 8 The performances of SPSD with different parameter settings on MSRVTT (The bold font indicates the best results)

λ	γ	α	Video–text retrieval				Text–video retrieval				Rsum
			R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
30	10	0.1	28.1	59.2	71.1	4.0	28.2	57.2	70.4	4.0	314.2
30	10	0.5	28.0	**61.5**	**73.9**	**3.0**	**30.2**	59.6	72.4	4.0	325.6
3	10	1	28.0	59.0	70.2	4.0	27.1	57.2	69.2	4.0	310.7
300	10	1	28.8	60.5	73.2	**3.0**	28.2	57.3	71.0	4.0	319.0
30	100	1	27.8	58.8	72.1	4.0	26.8	57.1	69.9	4.0	312.5
30	1	1	28.5	58.8	72.3	4.0	29.0	58.3	72.1	4.0	319.0
30	10	1	**29.7**	61.1	73.8	**3.0**	29.3	**60.9**	**72.8**	4.0	**327.6**

4.3 Ablation studies

4.3.1 Cross-granularity self-distillation

To evaluate the effectiveness of cross-granularity self-distillation, we adopt the loss as shown in Eq. (18), which consists of the cross-granularity self-distillation and contrastive loss computed only on semantic-level features. The other hierarchical losses are ignored in this experiment. The ratio of token screening r is set to 1. We vary parameter λ, the trade-off parameter of contrastive part and cross-granularity self-distillation part in the loss. With $\lambda = 0$, only contrastive loss is used for the model. The result is shown in Table 4. We can see that the model with $\lambda = 300$ achieves the best overall performance $Rsum = 314.3\%$ of video-to-text and text-to-video retrieval, which is better than the model with $\lambda = 0$ ($Rsum = 307.9\%$). And $MedR = 3$ obtained with $\lambda = 300$ is better than $MedR = 4$ with $\lambda = 0$ on video-to-text retrieval. This validates the effectiveness of cross-granularity self-distillation loss. In other words, the global representations could obtain information from fine-grained interaction of video and text tokens.

4.3.2 Token screening module

To validate token screening module, which selects important tokens for computing fine-grained similarity, we vary the screening ratio based on the above experiment and λ is set to 300. Since we have two token screening modules, respectively, for video and text, r_v for video and r_t for text are both varied, and the performances are shown in Table 5. $r_v = 1, r_t = 1$ represents using all tokens for similarity computation, with which the overall performance obtained is $Rsum = 314.3\%$. The model with $r_v = 0.75, r_t = 0.5$ obtains the best performance $Rsum = 318.5\%$. This validates that selecting 75% important visual tokens and 50% important textual tokens are optimal for fine-grained similarity computation. In following experiments, we fix $r_v = 0.75$ and $r_t = 0.5$.

4.3.3 Hierarchical contrastive loss

To validate the hierarchical contrastive learning, we ignore the two similarity-preserving self-distillation losses and vary the trade-off parameter α between contrastive loss of semantic level and feature level. The result is shown in Table 6. When $\alpha = 0$, only the semantic-level contrastive loss is considered in the model, which has the performance $Rsum = 307.9\%$. When $\alpha = 1$, the weights for semantic level and feature level are the same, with which the model gets $Rum = 303.4\%$. We can see that the model with $\alpha = 0.1$ achieves the best overall performance $Rsum = 309.4\%$. This explains that the low-level and high-level features of encoders both contribute to retrieval performance, but improper weight will hinder the performance. Relative to the low-level features, the high-level features are more suitable for retrieval task.

4.3.4 Cross-layer self-distillation

To further validate the cross-layer self-distillation, we set the model without cross-granularity loss and $\alpha = 0.1$ for low-level contrastive loss. We conduct the experiments on hyper-parameter γ for cross-layer self-distillation loss as Eq. (19). The result is shown in Table 7, which shows that the model with $\gamma = 3$ obtains the best overall performance $Rsum = 315.3\%$. Without cross-layer self-distillation, i.e., setting $\gamma = 0$, the model has the overall performance $Rsum = 309.4\%$. This declares it is effective to construct cross-layer self-distillation loss by utilizing high-level similarity to provide soft label to guide the learning of the cross-modal similarity based on low-level transformer features.

4.3.5 The trade-off parameters

The three components in the whole objective function in Eq. (19) influence each other. We conduct the experiments on the trade-off hyper-parameters of the function. The experimental result in Table 8 shows that the model with $\lambda = 30, \gamma = 10, \alpha = 1$ gets the best overall performance $Rsum = 327.6\%$ and the best $MedR = 3.0$ on video-to-text retrieval. The result states that the proposed two kinds of similarity-preserving self-distillation and hierarchical loss are effective to cross-modal retrieval.

5 Conclusion

In this paper, we introduce a similarity-preserving self-distillation method for fine-grained video–text alignment and hierarchical feature learning. The proposed cross-granularity self-distillation can make the global representations of video and text encoders obtain the fine-grained cross-modal interaction. Cross-layer self-distillation demonstrates that the similarity learning based on low-level features benefits from the soft label provided by the similarity of high-level features. The hierarchical losses including hierarchical cross-granularity self-distillation loss, hierarchical contrastive loss and cross-layer self-distillation loss improve the performances of video-to-text and text-to-video retrieval tasks. Our method achieves outstanding performances on MSRVTT, LSMDC and ActivityNet.

Author Contributions Yan Hua and Jiachen Wang wrote the main manuscript text, Jiachen Wang and Hongwei Kou did the coding work, and Yingyun Yang managed the whole project and provided funding. And all authors reviewed the manuscript.

Funding This work was funded by National Key Research and Development Program of China (Grant No.2020YFB1406800).

Research Data Policy and Data Availability Statements The three dataset features used in this paper are available and obtained in http://pascal.inrialpes.fr/data2/vgabeur/video-features/MSRVTT.tar.gz and http://pascal.inrialpes.fr/data2/vgabeur/video-features/LSMDC.tar.gz and http://pascal.inrialpes.fr/data2/vgabeur/video-features/activity-net.tar.gz.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: Proceedings of the 31st international conference on neural information processing systems, pp 6000—6010
2. Devlin J, Chang MW, Lee K, et al (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
3. Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR, pp 8748–8763, arXiv:1609.08124
4. Liu S, Fan H, Qian S, et al (2021) Hit: hierarchical transformer with momentum contrast for video-text retrieval. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 11,915–11,925, https://doi.org/10.1109/ICCV48922.2021.01170
5. Gabeur V, Sun C, Alahari K, et al (2020) Multi-modal transformer for video retrieval. In: European conference on computer vision. Springer, pp 214–229, https://doi.org/10.1007/978-3-030-58548-8_13
6. Yang J, Bisk Y, Gao J (2021) Taco: token-aware cascade contrastive learning for video-text alignment. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 11,562–11,572, https://doi.org/10.1109/ICCV48922.2021.01136
7. Li LH, Yatskar M, Yin D, et al (2019) Visualbert: a simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557
8. Patrick M, Huang PY, Asano Y, et al (2021) Support-set bottlenecks for video-text representation learning. In: International conference on learning representations, Vienna, Austria
9. Liu Y, Albanie S, Nagrani A, et al (2019) Use what you have: video retrieval using representations from collaborative experts. arXiv preprint arXiv:1907.13487
10. Huang Z, Zeng Z, Liu B, et al (2020) Pixel-bert: aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849v2

11. Li G, Duan N, Fang Y, et al (2020) Unicoder-vl: a universal encoder for vision and language by cross-modal pre-training. In: Proceeding of the AAAI conference on artificial intelligence, pp 11,336–11,344, https://doi.org/10.1609/aaai.v34i07.6795
12. Su W, Zhu X, Cao Y, et al (2020) Vl-bert: pre-training of generic visual-linguistic representations. In: Proceedings of the international conference on learning representation, Addis Ababa, Ethiopia
13. Lample G, Conneau A (2019) Cross-lingual language model pretraining. arXiv preprint arXiv: 1901.07291
14. Huang H, Liang Y, Duan N, et al (2019) Unicoder: a universal language encoder by pre-training with multiple cross-lingual tasks. arXiv preprint https://doi.org/10.48550/arXiv.1909.00964
15. Tan H, Bansal M (2019) Lxmert: Leaning cross-modality encoder representations from transformers. In: Conference on empirical methods in natural language processing and 9th international joint conference on natural language processing, https://doi.org/10.18653/v1/D19-1514
16. Khattab O, Zaharia M (2020) Colbert: efficient and effective passage search via contextualized late interaction over Bert. In: Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, pp 39–48, https://doi.org/10.1145/3397271.3401075
17. Yao L, Huang R, Hou L, et al (2021) Filip: Fine-grained interactive language-image pre-training. arXiv preprint https://doi.org/10.48550/arXiv.2111.07783
18. Zhang L, Song J, Gao A, et al (2019) Be your own teacher: improve the performance of convolutional neural networks via self distillation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3713–3722, https://doi.org/10.1109/ICCV.2019.00381
19. Tung F, Mori G (2019) Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1365–1374, https://doi.org/10.1109/ICCV.2019.00145
20. Zhu J, Tang S, Chen D, et al (2021) Complementary relation contrastive distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9260–9269, https://doi.org/10.1109/CVPR46437.2021.00914
21. Ji M, Shin S, Hwang S, et al (2021a) Refine myself by teaching myself: feature refinement via self-knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10,664–10,673, https://doi.org/10.1109/CVPR46437.2021.01052
22. Tenney I, Das D, Pavlick E (2019) Bert rediscovers the classical NLP pipeline. In: The 57th annual meeting of the association for computational linguistics (ACL), https://doi.org/10.18653/v1/P19-1452
23. Hao Y, Dong L, Wei F, et al (2019) Visualizing and understanding the effectiveness of Bert. In: conference on empirical methods in natural language processing and 9th international joint conference on natural language processing, https://doi.org/10.18653/v1/D19-1424
24. Qiao Y, Xiong C, Liu Z, et al (2019) Understanding the behaviors of Bert in ranking. arXiv preprint. arXiv: 1904.07531
25. Vig J (2019) A multiscale visualization of attention in the transformer model. In: The 57th annual meeting of the association for computational linguistics (ACL), p 37, https://doi.org/10.18653/v1/P19-3007
26. Peters M, Neumann M, Zettlemoyer L, et al (2018) Dissecting contextual word embeddings: Architecture and representation. In: Proceedings of the 2018 conference on empirical methods in natural language processing (EMNLP), arXiv:1808.08949
27. Van den Oord A, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv preprint 2(3):4. arXiv:1807.03748
28. Zhu L, Yang Y (2020) Actbert: learning global-local video-text representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8746–8755, https://doi.org/10.1109/CVPR42600.2020.00877
29. Lu J, Batra D, Parikh D, et al (2019) Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Proceedings of the 33rd international conference on neural information processing systems, p 13–23
30. Yu Y, Kim J, Kim G (2018) A joint sequence fusion model for video question answering and retrieval. In: Proceedings of the European conference on computer vision (ECCV), pp 471–487, https://doi.org/10.1007/978-3-030-01234-2_29
31. Wang X, Zhu L, Yang Y (2021) T2vlad: global-local sequence alignment for text-video retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5079–5088, https://doi.org/10.1109/CVPR46437.2021.00504
32. Lee KH, Chen X, Hua G, et al (2018) Stacked cross attention for image-text matching. In: Proceedings of the European conference on computer vision (ECCV), pp 201–216, https://doi.org/10.1007/978-3-030-01225-0_13
33. Ji K, Liu J, Hong W, et al (2022) Cret: cross-modal retrieval transformer for efficient text-video retrieval. In: Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval, pp 949–959
34. Gao Y, Lu Z (2023) Cmmt: cross-modal meta-transformer for video-text retrieval. In: Proceedings of the 2023 ACM international conference on multimedia retrieval, pp 76–84
35. Gorti SK, Vouitsis N, Ma J, et al (2022) X-pool: cross-modal language-video attention for text-video retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5006–5015
36. Jin M, Zhang H, Zhu L et al (2022) Coarse-to-fine dual-level attention for video-text cross modal retrieval. Knowl Syst 242(108):354
37. Zolfaghari M, Zhu Y, Gehler P, et al (2021) Crossclr: cross-modal contrastive learning for multi-modal video representations. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1450–1459, https://doi.org/10.1109/ICCV48922.2021.00148
38. Ging S, Zolfaghari M, Pirsiavash H, et al (2020) Coot: cooperative hierarchical transformer for video-text representation learning. In: 34th conference on neural information processing systems (NeurIPS 2020), pp 22,605–22,618
39. Ji Z, Chen K, Wang H (2021b) Step-wise hierarchical alignment network for image-text matching. In: Proceedings of the thirtieth international joint conference on artificial intelligence, IJCAI-21. international joint conferences on artificial intelligence organization, pp 765–771, https://doi.org/10.24963/ijcai.2021/106
40. Jiang J, Min S, Kong W, et al (2022b) Tencent text-video retrieval: hierarchical cross-modal interactions with multi-level representations. IEEE Access
41. Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. In: NIPS deep learning workshop, https://doi.org/10.4140/TCP.n.2015.249
42. Wang L, Yoon KJ (2022) Knowledge distillation and student-teacher learning for visual intelligence: a review and new outlooks. IEEE Trans Pattern Anal Mach Intell 44(6):3048–3068. https://doi.org/10.1109/TPAMI.2021.3055564
43. Park W, Kim D, Lu Y, et al (2019) Relational knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3967–3976, https://doi.org/10.1109/CVPR.2019.00409
44. Tian Y, Krishnan D, Isola P (2020) Contrastive representation distillation. In: International conference on learning representations
45. Li J, Ji Z, Wang G, et al (2022) Learning from students: online contrastive distillation network for general continual learning. In: Proc 31st Int Joint Conf Artif Intell, pp 3215–3221

46. Hou Y, Ma Z, Liu C, et al (2019) Learning lightweight lane detection cnns by self attention distillation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1013–1021, https://doi.org/10.1109/ICCV.2019.00110
47. Luo H, Ji L, Zhong M et al (2022) Clip4clip: an empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing 508:293–304
48. Xue H, Sun Y, Liu B, et al (2023) Clip-vip: adapting pre-trained image-text model to video-language representation alignment. In: International conference on learning representations
49. Jiang H, Zhang J, Huang R, et al (2022a) Cross-modal adapter for text-video retrieval. arXiv preprint arXiv:2211.09623
50. Huang S, Gong B, Pan Y, et al (2023) Vop: text-video co-operative prompt tuning for cross-modal retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6565–6574
51. Wang H, He D, Wu W, et al (2022) Coder: coupled diversity-sensitive momentum contrastive learning for image-text retrieval. In: 17th European conference on computer vision–ECCV 2022, Springer, pp 700–716
52. Wu Z, Xiong Y, Yu SX, et al (2018) Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3733–3742, https://doi.org/10.1109/CVPR.2018.00393
53. Xu J, Mei T, Yao T, et al (2016) Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5288–5296, https://doi.org/10.1109/CVPR.2016.571
54. Rohrbach A, Rohrbach M, Tandon N, et al (2015) A dataset for movie description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3202–3212, https://doi.org/10.1109/CVPR.2015.7298940
55. Krishna R, Hata K, Ren F, et al (2017) Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision, pp 706–715
56. Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. In: International conference on learning representations

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