electronics

Article

Revisiting Hard Negative Mining in Contrastive Learning for Visual Understanding

Hao Zhang, Zheng Li, Jiahui Yang, Xin Wang, Caili Guo and Chunyan Feng *

Beijing Key Laboratory of Network System Architecture and Convergence, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China; zhanghao0215@bupt.edu.cn (H.Z.); lizhengzachary@bupt.edu.cn (Z.L.); yangjh@bupt.edu.cn (J.Y.); wangxin1999@bupt.edu.cn (X.W.); guocaili@bupt.edu.cn (C.G.)
* Correspondence: cyfeng@bupt.edu.cn

Abstract: Efficiently mining and distinguishing hard negatives is the key to Contrastive Learning (CL) in various visual understanding tasks. By properly emphasizing the penalty of hard negatives, Hard Negative Mining (HNM) can improve the CL performance. However, there is no method to quantitatively analyze the penalty strength of hard negatives, which makes training difficult to converge. In this paper, we propose a method for measuring and controlling the penalty strength. We first define a penalty strength metric to provides a quantitative analysis tool for HNM. Then, we propose a Triplet loss with Penalty Strength Control (T-PSC), which can balance the penalty strength of hard negatives and the difficulty of model optimization. In order to verify the effectiveness of the proposed T-PSC method in different modalities, we applied it to two visual understanding tasks: Image–Text Retrieval (ITR) for multi-model processing, and Temporal Action Localization (TAL) for video processing. T-PSC can be applied to existing ITR and TAL models in a plug-and-play manner without any changes. Experiments combined with existing models show that a reasonable control of the penalty strength can speed up training and improve the performance on higher-level tasks.

Keywords: contrastive learning; hard negative mining; Image–Text Retrieval; Temporal Action Localization; visual understanding

Citation: Zhang, H.; Li, Z.; Yang, J.; Wang, X.; Guo, C.; Feng, C. Revisiting Hard Negative Mining in Contrastive Learning for Visual Understanding. *Electronics* **2023**, *12*, 4884. https://doi.org/10.3390/electronics12234884

Academic Editor: Dah-Jye Lee

Received: 13 October 2023
Revised: 26 November 2023
Accepted: 29 November 2023
Published: 4 December 2023

1. Introduction

Contrastive learning (CL) has achieved great success in visual understanding in recent years [1–6]. It is widely applied in cross-modal retrieval [7–9], action recognition [10], instance segmentation [11], and other fields [12–14]. Through contrastive loss, the purpose of CL is to bring positive pairings together and push negative pairs apart in the feature embedding space. In other words, given the positive and negative pairings, we can build a meaningful feature embedding space using a contrastive loss.

The challenges that affect the contrastive learning performance are mainly reflected in the following two aspects: (1) how to define positive and negative sample pairs; and (2) how to design appropriate contrastive loss. The first aspect is used to design supervised learning or self-supervised learning paradigms. A positive sample pair in supervised contrastive learning is made up of samples from the same category, whereas a negative sample pair is made up of samples from separate categories. A positive pair in self-supervised contrastive learning is frequently generated via two perspectives (e.g., distinct data augmentations) of the same sample, whereas a negative sample pair is built of a sample and additional samples of other categories or their augmented samples, according to [15]. The effectiveness of the second element strongly influences the performance on high-level contrastive learning tasks, which is a major difficulty for CL. Broadly speaking, contrastive loss is the most significant element influencing the contrastive learning performance.

The triplet loss [1] is one of the most widely used contrastive loss functions for visual understanding tasks. A triplet consists of three parts: an anchor, a positive, and a negative. Taking image-to-text retrieval as an example, we take each image as an anchor. Captions that are relevant to the anchor image are positive, while those that are irrelevant are negative. The case where the anchor is closer to the negative than the positive is penalized via the triplet loss.

However, in visual understanding tasks, in addition to negative samples that are completely opposite to positive samples, there are also many negative samples that are very similar to positive samples in semantics or pixels. Specifically, each anchor has one positive and many negatives in a single batch. A large proportion of these negatives are further away from the anchor than the positives. Therefore, these negatives are redundant and are often called easy negatives. Negatives that are closer to the anchor than the positive are defined as hard negatives [1]. Ignoring the contrastive learning loss of hard negative samples will prevent the network from learning discriminative features. In practice, the performance of triplet loss is highly dependent on Hard Negative Mining (HNM). HNM mines use hard negatives for triplet loss. Many state-of-the-art visual understanding models [16–21] employ Triplet loss with Hard Negative Mining (T-HNM) [22] as the optimization objective. T-HNM enables models to mine hard negative samples on various visual tasks, improving the performance of high-level tasks.

Nevertheless, some studies observe that HNM can make training difficult to converge [23]. HNM essentially increases the penalty strength for hard negatives. HNM provides a large gradient-to-hard negatives, which are optimized emphatically. Easy negatives are either barely or not at all optimized. Focusing on optimizing hard negatives can help the model learn discriminative features [21]. However, is it true that the stronger the penalty, the better? Existing studies mainly design HNM strategies based on intuition and lack quantitative analysis. What the appropriate level of penalty strength is for hard negatives has not been studied.

In order to solve the above-mentioned problems, we revisit hard negative mining in contrastive learning and propose a method for measuring and controlling the penalty strength of negatives. We first define a metric for the penalty strength of negatives. Then, we perform a quantitative analysis of common loss functions. The penalty strength of hard negatives and the complexity of model optimization are shown to be conflicting. Too large a penalty strength can lead to difficulties in optimizing. As a result, training is difficult to converge. To this end, we further propose a Triplet loss with Penalty Strength Control (T-PSC). A temperature coefficient τ is introduced to control the penalty strength. We can balance these two contradictory properties by controlling τ. Balancing the two properties can speed up model convergence and improve the retrieval performance. The major contributions of this paper are summarized as follows:

- We define a metric for the penalty strength of negatives, which provides a quantitative analysis tool for HNM.
- We find that the penalty strength of hard negatives and the difficulty of model optimization are contradictory. The design of loss functions needs to balance the two items.
- Experiments on two visual understanding tasks, i.e., Image–Text Retrieval (ITR) and Temporal Action Localization (TAL), with different modal data as research objects have verified that T-PSC can accelerate model training and improve the performance of current visual understanding models. T-PSC can be applied to existing ITR and TAL models in a plug-and-play manner without any changes.

2. Related Work

2.1. Contrastive Learning

In recent years, contrastive learning has made great progress in visual understanding, and a meaningful feature embedding space is generated via the contrastive loss [24]. SimCLR [25] proposes a simple framework based on contrastive learning, which learns effective visual representations by minimizing the distance between differently augmented views of

the same sample via a contrastive loss. Supervised contrastive learning [2] aims to leverage label information more effectively than cross entropy, forcing normalized embeddings from the same class to be closer than embeddings from different classes. MoCo [4] enables unsupervised contrastive learning to provide competitive results under the common linear protocol on downstream tasks in a dynamic dictionary look-up manner. MoCo is a general mechanism for using contrastive losses that can outperform its supervised pre-training counterpart in some visual understanding tasks. On the basis of MoCo, MoCo v2 [26] combines an MLP head and the stronger data augmentation proposed in SimCLR. MoCo v2 establishes stronger baselines that outperform SimCLR and do not require large training batches. Some recent work explores contrastive learning in broader visual understanding tasks, such as anomaly detection [27], keypoint detection [28], depth prediction [29], and so on.

2.2. Image–Text Retrieval

Image–Text Retrieval (ITR) is the main application of vision–language retrieval. ITR is defined as retrieving relevant items across images and captions [30–32]. Given a query image, the goal is to find the most relevant caption from the text gallery. The challenge of ITR is the heterogeneous gap between images and captions. The mainstream approach to ITR is to learn a model to measure the similarities between images and captions. Then, the retrieval results are obtained by ranking the similarities.

Existing ITR methods can be divided into two categories according to the image–text matching methods, i.e., global-level matching methods [17,33,34] and local-level matching methods [18,30,35]. The global-level matching methods embed the whole images and sentences into a joint embedding space, and the matching score between the embeddings of images and sentences can be calculated via a simple similarity metric (e.g., cosine similarity). DeViSE [36] proposes the first global-level matching model, which employs CNN and Skip-Gram to project images and sentences into a joint embedding space. The local-level matching methods obtain the matching score by calculating the cross attention between the image regions and words. SCAN [16] is known as a stacked cross-attention network, which measures image–text similarity by aligning the image regions and words.

2.3. Temporal Action Localization

The goal of Temporal Action Localization (TAL) is to find the categories and temporal boundaries of actions in an untrimmed video. The existing TAL methods can be divided into two categories: one-stage methods and two-stage methods.

The one-stage TAL method directly extracts features and performs action classification and regression for video segments. SSAD [37] is inspired by single-shot detection methods (SSD [38] and YOLO [39]), which abandons the process of generating proposals in the two-stage TAL and directly predicts the action score and timing boundary of the action. GTAN [40] introduces Gaussian kernels to dynamically optimize the temporal scale of each action proposal, which can generate multiple feature maps in different temporal resolutions. MGG [41] devises two temporal convolutional layers and a bilinear matching model to obtain segment proposals and frame actionness simultaneously on the RGB frames. In order to simplify the complexity, ActionFormer [42] proposed an anchor-free model based on transformers, using feature pyramids and local self-attention to model a long-term temporal context. TriDet [43] alleviates the problem of boundary prediction by modeling action boundaries via Trident-head in a way that estimates the relative probability distribution around the boundary.

The two-stage temporal action positioning method requires first generating proposals and then performing action recognition and temporal regression. BSN [44] uses a local-to-global method and uses a three-layer temporal convolutional neural network to generate all starting and ending timestamps and action probabilities. In order to solve the problem that existing methods cannot efficiently generate adequately reliable confidence scores for retrieving proposals, BMN [45] is proposed to generate action confidence for all

proposals simultaneously. On the premise of BSN, PGCN [46] applies GCN to do message aggregation among proposals to improve the performance. ContextLoc [47] proposes to use three networks (L-Net, G-Net, and P-Net) to respectively utilize local context information, global context information, and context-aware inter-proposal relations to obtain rich feature representation.

3. Methodology

3.1. Preliminaries

3.1.1. Contrastive Learning

The goal of CL is to learn a representation of data such that similar instances are close together in the representation space, while dissimilar instances are far apart [24]. We denote a sample as I_i and a text as T_i. Take I_i as an anchor; (I_i, T_i) is a positive pair; and $(I_i, T_{j,i\neq j})$ is a negative pair. The mainstream approach to CL is to learn a model to measure the similarity $s_{i,j}$ between I_i and T_j.

3.1.2. Hardness of Negatives

The hardness of the negative pair (I_i, T_j) is defined as $h_{i,j} = s_{i,j} - s_{i,i}$, where $s_{i,j}$ denotes the similarity of the negative pair (I_i, T_j), and $s_{i,i}$ denotes the similarity of the positive pair (I_i, T_i). The larger the hardness, the harder it is for the negatives to be distinguished correctly.

3.1.3. Triplet Loss

Triplet loss [1] is a ranking loss widely used in various visual understanding tasks [48–51]. Given the positive similarity $s_{i,i}$ and the negative pair similarity $s_{i,j}$, triplet loss in the sample-to-text direction can be formulated as follows:

$$\mathcal{L}_{\text{Triplet},i2t} = \sum_{i=1}^{B} \sum_{j=1,i\neq j}^{B} \left[s_{i,j} - s_{i,i} + \lambda\right]_+ \tag{1}$$

Similarly, the text-to-sample direction $\mathcal{L}_{\text{Triplet},t2i}$ is symmetrical to $\mathcal{L}_{\text{Triplet},i2t}$:

$$\mathcal{L}_{\text{Triplet},t2i} = \sum_{i=1}^{B} \sum_{j=1,i\neq j}^{B} \left[s_{j,i} - s_{i,i} + \lambda\right]_+, \tag{2}$$

where B is the batch size, λ is the margin for similarity separation, $[x]_+ \equiv \max(x, 0)$. Triplet loss encourages increasing the similarity of positive pairs. All negatives with $\mathcal{L}_{\text{Triplet}} > 0$ are penalized equally, regardless of their hardness.

3.1.4. Hard Negative Mining

Triplet loss with Hard Negative Mining (T-HNM) [22] yields significant performance gains on visual understanding tasks. Most visual understanding models [52–54] adopt T-HNM as the optimization objective. T-HNM takes the form of:

$$\mathcal{L}_{\text{T-HNM},i2t} = \sum_{i=1}^{B} \max_{j=1,i\neq j}^{B} \left[s_{i,j} - s_{i,i} + \lambda\right]_+. \tag{3}$$

Only the negative with maximum hardness is penalized. Despite the performance gain, some studies observe that $\mathcal{L}_{\text{T-HNM}}$ makes training difficult to converge [1].

3.2. Metric for the Penalty Strength of Negatives

HNM essentially increases the penalty strength for hard negatives. HNM provides a large gradient-to-hard negatives. Focusing on optimizing hard negatives can help the model learn discriminative features. However, is it true that the stronger the penalty, the better?

To clear up the above doubt, we first define a metric for the penalty strength of negatives. For dissimilar loss functions, the gradients with respect to the negative pairs are different. When using gradient descent for optimization, pairs with large gradients are optimized emphatically. Therefore, the gradient can directly reflect the penalty strength. We denote the gradient of the loss $\mathcal{L}$ with respect to $s_{i,j}$ as $g_{i,j} = \partial\mathcal{L}/\partial s_{i,j}$. The penalty strength of the negative pair (I_i, T_j) is defined as:

$$p_{i,j} = \frac{g_{i,j}}{\sum_{k=1,i\neq k}^{B} g_{i,k}}, \tag{4}$$

where $p_{i,j}$ is the ratio of the gradient of $s_{i,j}$ to the total gradient.

The gradient of $\mathcal{L}_{\text{Triplet},i2t}$ with respect to $s_{i,j}$ is:

$$g_{i,j}^{\text{Triplet},i2t} = \frac{\partial\mathcal{L}_{\text{Triplet},i2t}}{\partial s_{i,j}} = \mathbb{I}\{h_{i,j} > -\lambda\}, \tag{5}$$

where $\mathbb{I}\{x\}$ is an indicator function; if x is true, $\mathbb{I}\{x\} = 1$, otherwise $\mathbb{I}\{x\} = 0$. The gradient with respect to all $s_{i,j}$ satisfying $h_{i,j} > -\lambda$ is the same. Figure 1 left shows the optimization processes of $\mathcal{L}_{\text{Triplet}}$. All negatives satisfying $h_{i,j} > -\lambda$ are penalized equally. Based on Equations (4) and (5), the relationship between $p_{i,j}$ and $h_{i,j}$ can be obtained, as shown in Figure 2a. $h_{i,j}$ is uniformly sampled in $[-1, 1]$. For $\mathcal{L}_{\text{Triplet}}$, when $h_{i,j} > -\lambda$, $p_{i,j}$ of all negatives is equal.

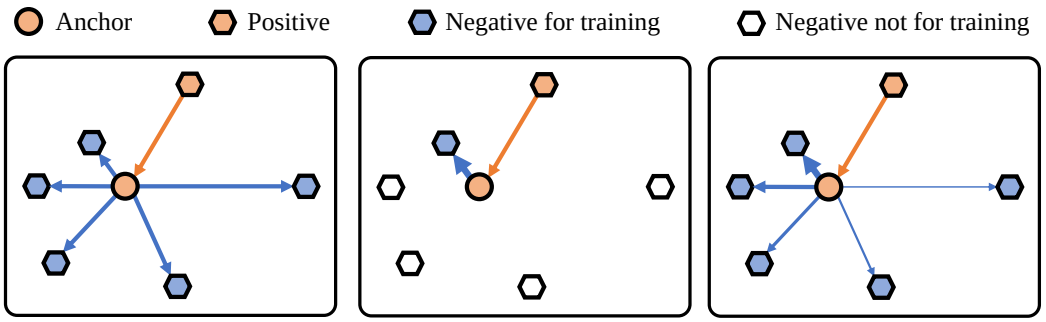
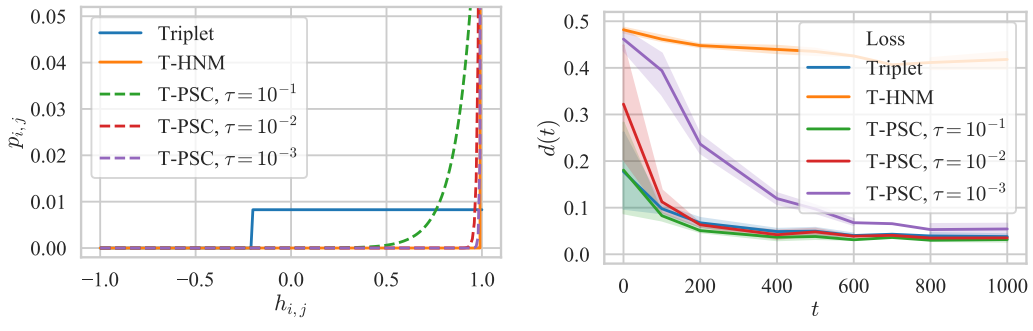


Figure 1. The orange arrow indicates that the positive is pulled closer to the anchor. From left, middle and right: Triplet, T-HNM and T-PSC. The blue arrow indicates that the negative is pushed away from the anchor. The thickness of the blue line indicates the penalty strength.



(**a**) (**b**)

Figure 2. The penalty strength of hard negatives and model optimization difficulty are contradictory. (**a**) Penalty strength of negatives. (**b**) Model optimization difficulty.

The gradient of $\mathcal{L}_{\text{T-HNM},i2t}$ with respect to $s_{i,j}$ is:

$$g_{i,j}^{\text{T-HNM},i2t} = \frac{\partial\mathcal{L}_{\text{T-HNM},i2t}}{\partial s_{i,j}} = \begin{cases} 1, & h_{i,j} = \hat{h}_i, h_{i,j} > -\lambda, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

where $\hat{h}_i = \max_{j=1,i\neq j}^{B} h_{i,j}$. $\mathcal{L}_{\text{T-HNM}}$ only provides the gradient for the hardest negative. As shown in Figure 1 middle, $\mathcal{L}_{\text{T-HNM}}$ only penalizes the negative most similar to the anchor. As shown in Figure 2b, $p_{i,j} > 0$ only when $h_{i,j}$ is the largest. The steeper the curve, the stronger the penalty for hard negatives. $\mathcal{L}_{\text{T-HNM}}$ has the largest penalty strength for hard negatives.

3.3. Model Optimization Difficulty

Most models [35,55–57] adopt $\mathcal{L}_{\text{T-HNM}}$ as the optimization objective. However, $\mathcal{L}_{\text{T-HNM}}$ can make model training difficult to converge. Various studies show that optimizing with hard negatives leads to optimization difficulties [1]. To quantitatively analyze the optimization process of the model, we define the optimization difficulty as follows:

$$d(t) = \frac{\sum_{i=1}^{B}\sum_{j=1,i\neq j}^{B}\mathbb{I}\{h_{i,j} > 0\}}{B\cdot(B-1)}, \tag{7}$$

where t is the number of iterations of training. Negative pairs satisfying $h_{i,j} > 0$ are hard negative pairs. $d(t)$ is the ratio of the number of hard negative pairs to the total number of negative pairs. A batch contains $B\cdot(B-1)$ negative pairs. $d(t)$ can reflect the optimization difficulty of the current batch. The more hard negatives, the more difficult it is to optimize.

We quantitatively analyze the optimization difficulty of different loss functions in the early stages of training, as shown in Figure 2b. We conduct experiments on a classic ITR model, VSE++ [22]. The optimization difficulty of $\mathcal{L}_{\text{Triplet}}$ is quickly reduced to a lower level. In contrast, the optimization difficulty of $\mathcal{L}_{\text{T-HNM}}$ is always large. Nearly half of the negative pairs are hard negative pairs. $\mathcal{L}_{\text{T-HNM}}$ with maximum penalty strength leads to a large number of hard negatives in training. This shows that the penalty strength and model optimization difficulty are contradictory. Too large a penalty strength leads to difficulties in optimizing the model. A small penalty strength can reduce the optimization difficulty, but it also reduces the performance.

3.4. Penalty Strength Control

To balance the two contradictory properties, we propose T-PSC, which is inspired by the contrastive loss [58] used in pre-training models [59]. Contrastive loss takes the form of:

$$\mathcal{L}_{\text{Contrastive},i2t} = -\sum_{i=1}^{B}\log\frac{\exp(s_{i,i}/\tau)}{\sum_{j=1}^{B}\exp(s_{i,j}/\tau)}, \tag{8}$$

where τ is a temperature coefficient. $\mathcal{L}_{\text{Contrastive}}$ can be transformed into a form of pair similarity optimization:

$$\mathcal{L}_{\text{Contrastive},i2t} = \sum_{i=1}^{B}\log\left(1+\sum_{j=1,i\neq j}^{B}\exp\left(\left(s_{i,j}-s_{i,i}\right)/\tau\right)\right). \tag{9}$$

From Equation (9), we can see that $\mathcal{L}_{\text{Contrastive},i2t}$, like $\mathcal{L}_{\text{Triplet},i2t}$, is optimizing $(s_{i,j} - s_{i,i})$. τ can control the penalty strength of hard negatives. The penalty strength increases as τ decreases.

Although $\mathcal{L}_{\text{Contrastive}}$ can control the penalty strength, it is not directly suitable for visual understanding tasks that require high feature discriminability, such as ITR and TAL. Both ITR and TAL require the model to correctly distinguish between positive and negative samples to achieve a high performance, including effectively identifying positive samples and hard negative samples. However, $\mathcal{L}_{\text{Contrastive}}$ does not increase $(s_{i,i} - s_{i,j})$ sufficiently. As a result, $\mathcal{L}_{\text{Contrastive}}$ does not provide the strong discriminative information required for the visual understanding task. The results on the training set do not generalize well

to the test set. Following $\mathcal{L}_{\text{Triplet}}$ that uses a margin λ for better similarity separation, we introduce the margin λ into $\mathcal{L}_{\text{Contrastive}}$ and propose our T-PSC:

$$\mathcal{L}_{\text{T-PSC},i2t} = \tau \sum_{i=1}^{B} \log\left(1 + \sum_{j=1, i\neq j}^{B} \exp\left(\left(s_{i,j} - s_{i,i} + \lambda\right)/\tau\right)\right). \tag{10}$$

In particular, $\mathcal{L}_{\text{T-HNM}}$ is a special form of $\mathcal{L}_{\text{T-PSC}}$:

$$\mathcal{L}_{\text{T-HNM}} = \lim_{\tau \to 0^{+}} \mathcal{L}_{\text{T-PSC}}. \tag{11}$$

The gradient of $\mathcal{L}_{\text{T-PSC},i2t}$ with respect to $s_{i,j}$ is:

$$g_{i,j}^{\text{T-PSC},i2t} = \frac{\partial \mathcal{L}_{\text{T-PSC},i2t}}{\partial s_{i,j}} = \frac{\exp\left(\left(h_{i,j} + \lambda\right)/\tau\right)}{1 + \sum_{j=1, i\neq j}^{B} \exp\left(\left(h_{i,j} + \lambda\right)/\tau\right)}. \tag{12}$$

Based on Equations (4) and (12), the relationship between $p_{i,j}$ and $h_{i,j}$ can be obtained. As shown in Figure 2a, $\mathcal{L}_{\text{T-PSC}}$ assigns a large penalty strength to hard negatives and a small one to easy negatives. The penalty strength can be controlled by τ. As shown in Figure 2b, as τ decreases, the penalty strength increases and the optimization difficulty increases accordingly. Choosing an appropriate τ, such as $\tau = 10^{-2}$, can achieve a large penalty strength and a small optimization difficulty. Two contradictory properties are balanced.

4. Experiments

We utilize the proposed T-PSC to conduct experiments on visual understanding tasks in a plug-and-play manner. In order to verify the effectiveness of T-PSC in different modalities, we applied it to two tasks: ITR and TAL. ITR focuses on the matching degree of image–text pairs, while TAL tends to the semantic similarity of similar video frames. From the perspective of contrastive learning, the purpose of both tasks is to pull positive pairs together and push negative pairs apart in the feature embedding space. We first introduce our training details and evaluation metrics, then perform extensive ablation studies on different aspects of the ITR task and provide a better understanding of how T-PSC measures and controls the penalty strength of negatives. Finally, we apply T-PSC to existing ITR and TAL models and obtain performance improvements.

4.1. Datasets and Experiment Settings

Image–Text Retrieval (ITR). Our method is evaluated using two benchmarks: Flickr30K [60] contains 31,000 images; each image is annotated with five sentences. We use 29,000 images for training, 1000 images for validation, and 1000 images for testing. MS-COCO [61] contains 123,287 images; each image is annotated with five sentences. We use 113,287 images for training, 5000 images for validation, and 5000 images for testing. For the performance evaluation of ITR, we use Recall@K (R@K) and Average Recall@K (Avg.), with $K = \{1, 5, 10\}$ as the evaluation metric. R@K indicates the percentage of queries for which the model returns the correct item in its top K results. Avg. represents the mean value of R@K.

Temporal Action Localization (TAL). Our method is evaluated using two benchmarks: THUMOS14 [62] and ActivityNet-1.3 [63]. THUMOS14 consists of 200 validation videos and 213 test videos for TAL. Without a loss of generality, we apply training on the validation subset and evaluate the model performance on the test subset [64]. ActivityNet-1.3 contains 10,024 videos and 15,410 action instances for training, 4926 videos and 7654 action instances for validation, and 5044 videos for testing. Following the standard practice [65], we train our method on the training subset and test it on the validation subset.

Experiment Settings. For T-PSC, the hyperparameters are set to $\lambda = 0.2$ and $\tau = 10^{-2}$ for all the datasets used in this paper.

4.2. Ablation Studies

We conducted ablation experiments on the ITR model to verify the impact of different parameters on the cross-modal performance of T-PSC.

4.2.1. Impact of Hyperparameters

There are two hyperparameters, i.e., margin λ and temperature coefficient τ, in T-PSC that can be tuned. We experiment with several combinations of hyperparameters on Flickr30K using VSE++. All experiments on the effects of hyperparameters are shown in Figure 3.

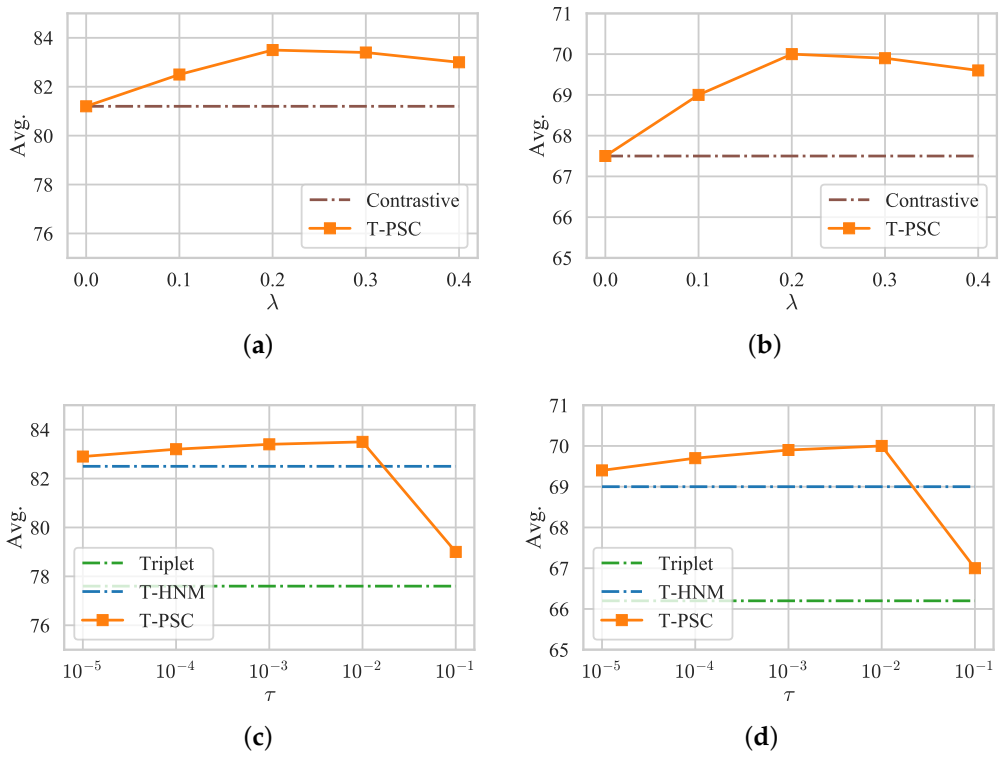


Figure 3. Impact of hyperparameters on Flickr30K. (**a**) The impact of Margin λ on the image-to-text task ($\tau = 10^{-2}$). (**b**) The impact of Margin λ on the text-to-image task ($\tau = 10^{-2}$). (**c**) The impact of Temperature coefficient τ on the image-to-text task ($\lambda = 0.2$). (**d**) The impact of Temperature coefficient τ on the text-to-image task ($\lambda = 0.2$).

We test the impact of λ by fixing τ to 10^{-2}. Figure 3a,b are the performance impact curves of hyperparameter λ on image-to-text and text-to-image, respectively. It can be seen from these two pictures that when $\lambda = 0.2$, the ITR model achieves the best performance. $\mathcal{L}_{\text{T-PSC}}$ degenerates into $\mathcal{L}_{\text{Contrastive}}$ when $\lambda = 0$. When $\lambda > 0$, the performance of $\mathcal{L}_{\text{T-PSC}}$ is always better than $\mathcal{L}_{\text{Contrastive}}$. It shows the significance of the margin for ITR.

We test the impact of τ by fixing λ to 0.2. Figure 3c,d are the performance impact curves of hyperparameter τ on image-to-text and text-to-image, respectively. As shown in these two pictures, when $\tau \leq 10^{-2}$, the performance of $\mathcal{L}_{\text{T-PSC}}$ is always higher than $\mathcal{L}_{\text{T-HNM}}$. According to Figure 2, when $\tau \leq 10^{-2}$, the penalty strength of $\mathcal{L}_{\text{T-PSC}}$ for hard negatives is comparable to that of $\mathcal{L}_{\text{T-HNM}}$. At the same time, the optimization difficulty of $\mathcal{L}_{\text{T-PSC}}$ is lower than $\mathcal{L}_{\text{T-HNM}}$. When $\tau = 10^{-1}$, the performance of $\mathcal{L}_{\text{T-PSC}}$ is lower than $\mathcal{L}_{\text{T-HNM}}$. According to Figure 2, when $\tau = 10^{-1}$, the penalty strength of $\mathcal{L}_{\text{T-PSC}}$ for hard negatives is not large enough. In particular, when $\tau = 10^{-2}$, $\mathcal{L}_{\text{T-PSC}}$ can achieve a large penalty strength and a small optimization difficulty. Two contradictory properties are balanced. At this point, the best performance is achieved.

4.2.2. Impact of Loss Functions

T-PSC is designed based on triplet loss and contrastive loss. Therefore, we compare $\mathcal{L}_{\text{T-PSC}}$ with $\mathcal{L}_{\text{Triplet}}$, $\mathcal{L}_{\text{T-HNM}}$, and $\mathcal{L}_{\text{Contrastive}}$. We conduct experiments on VSE++ [22]. We reproduce VSE++ with all experimental settings identical to [34]. Experimental results are shown in Table 1. $\mathcal{L}_{\text{T-PSC}}$ outperforms all three loss functions. As for Avg., $\mathcal{L}_{\text{T-PSC}}$ is 5.9%p, 2.3%p, and 1%p ahead of $\mathcal{L}_{\text{Triplet}}$, $\mathcal{L}_{\text{Contrastive}}$, and $\mathcal{L}_{\text{T-HNM}}$ in the image-to-text sub-task, respectively. At the same time, the performance of $\mathcal{L}_{\text{T-PSC}}$is 3.8%p, 2.5%p, and 1%p higher than $\mathcal{L}_{\text{Triplet}}$, $\mathcal{L}_{\text{Contrastive}}$, and $\mathcal{L}_{\text{T-HNM}}$ in the text-to-image sub-task, respectively. Both $\mathcal{L}_{\text{T-HNM}}$ and $\mathcal{L}_{\text{Contrastive}}$ are special forms of $\mathcal{L}_{\text{T-PSC}}$. As a more flexible form, $\mathcal{L}_{\text{T-PSC}}$ exhibits an optimal performance.

Table 1. Comparisons with related loss functions on Flickr30K.

Eval Task	**Image-to-Text (%)**				**Text-to-Image (%)**			
Loss	**R@1**	**R@5**	**R@10**	**Avg.**	**R@1**	**R@5**	**R@10**	**Avg.**
$\mathcal{L}_{\text{Triplet}}$	56.3	85.1	91.4	77.6	43.4	72.9	82.4	66.2
$\mathcal{L}_{\text{T-HNM}}$	64.0	88.8	94.6	82.5	47.0	76.0	83.9	69.0
$\mathcal{L}_{\text{Contrastive}}$	61.4	88.4	93.9	81.2	44.6	74.5	83.5	67.5
$\mathcal{L}_{\textbf{T-PSC}}$	66.5	90.0	94.0	**83.5**	48.4	76.9	84.6	**70.0**

4.2.3. Comparisons with Existing Loss Functions

There are several loss functions proposed for ITR: SSP [66], Meta-SPN [67], AOQ [68], and NCR [69]. We compare T-PSC with these losses on Flickr30K. The experimental results are shown in Table 2. Compared with these losses, T-PSC improves most evaluation metrics. T-PSC does not need to introduce many hyperparameters like SSP. Compared with Meta-SPN and NCR, T-PSC does not need to train an additional weight assignment network for loss functions. T-PSC also does not need to mine hard negatives on the entire dataset like AOQ. Overall, T-PSC achieves an impressive performance with a simple and easy-to-implement modification.

Table 2. Comparisons with existing ITR loss functions on Flickr30K.

Eval Task	**Image-to-Text (%)**				**Text-to-Image (%)**			
Model	**R@1**	**R@5**	**R@10**	**Avg.**	**R@1**	**R@5**	**R@10**	**Avg.**
BFAN [70]	68.1	91.4	95.9	85.1	50.8	78.4	85.8	71.7
+ SSP [66]	71.3	92.6	96.2	86.7	52.5	79.5	86.6	72.9
+ AOQ [68]	73.2	94.5	97.0	88.2	54.0	80.3	87.7	74.0
+ Meta-SPN [67]	72.5	93.2	96.7	87.5	53.3	80.2	87.2	73.6
+ **T-PSC**	74.3	93.8	96.7	**88.3**	54.5	80.8	87.5	**74.3**
SGRAF [71]	77.8	94.1	97.4	89.8	58.5	83.0	88.8	76.8
+ NCR [69]	77.3	94.0	97.5	89.6	59.6	84.4	89.9	78.0
+ **T-PSC**	78.3	95.0	97.4	**90.2**	60.4	85.0	90.6	**78.7**

4.3. Comparisons with Existing ITR and TAL Models

4.3.1. Improvements to Existing ITR Models

T-PSC can plug-and-play to improve the performance of existing ITR models. We conduct experiments on three classic ITR models: VSE++ [22], BFAN [70] and SGRAF [71]. Except for replacing the loss function, the other experimental settings are the same. Table 3 shows the improvements to these models on Flickr30K. In the image-to-text sub-task, the Avg. of T-PSC is 0.5%p, 3.2%p and 0.4%p higher than VSE++, BFAN and SGRAF, respectively, while the improvement in the text-to-image sub-task is 0.9%p, 2.6%p and 1.9%p. As shown in Table 4, on MS-COCO, applying T-PSC to VSE++, BFAN and SGRAF

can improve Avg. by 0.4%p, 0.7%p and 0.5%p in image-to-text sub-task. While the improvement in the text-to-image sub-task is 0.3%p, 0.5%p and 0.7%p. T-PSC can be easily integrated into the existing ITR model and improve the retrieval performance.

Table 3. Experimental Results on Flickr30K.

Eval Task	**Image-to-Text (%)**				**Text-to-Image (%)**			
Model	**R@1**	**R@5**	**R@10**	**Avg.**	**R@1**	**R@5**	**R@10**	**Avg.**
SCAN [16]	67.4	90.3	95.8	84.5	48.6	77.7	85.2	70.5
VSRN [17]	71.3	90.6	96.0	86.0	54.7	81.8	88.2	74.9
VSE++ [22]	69.4	90.7	95.4	85.2	52.1	79.0	85.5	72.2
+ T-PSC	70.7	90.8	95.6	**85.7**	52.9	79.8	86.7	**73.1**
BFAN [70]	68.1	91.4	95.9	85.1	50.8	78.4	85.8	71.7
+ T-PSC	74.3	93.8	96.7	**88.3**	54.5	80.8	87.5	**74.3**
SGRAF [71]	77.8	94.1	97.4	89.8	58.5	83.0	88.8	76.8
+ T-PSC	78.3	95.0	97.4	**90.2**	60.4	85.0	90.6	**78.7**

Table 4. Experimental Results on MS-COCO.

Eval Task	**Image-to-Text (%)**				**Text-to-Image (%)**			
Model	**R@1**	**R@5**	**R@10**	**Avg.**	**R@1**	**R@5**	**R@10**	**Avg.**
SCAN [16]	72.7	94.8	98.4	88.6	58.8	88.4	94.8	80.7
VSRN [17]	76.2	94.8	98.2	89.7	62.8	89.7	95.1	82.5
VSE++ [22]	73.0	94.5	98.2	88.6	58.3	88.1	94.4	80.3
+ T-PSC	74.0	94.6	98.4	**89.0**	58.7	88.6	94.4	**80.6**
BFAN [70]	74.9	95.2	98.3	89.5	59.4	88.4	94.5	80.8
+ T-PSC	76.2	95.8	98.7	**90.2**	60.7	88.6	94.7	**81.3**
SGRAF [71]	79.6	96.2	98.5	91.4	63.2	90.7	96.1	83.3
+ T-PSC	79.9	97.0	98.8	**91.9**	65.1	90.8	96.0	**84.0**

4.3.2. Improvements to Existing TAL Models

To maintain the plug-and-play nature, we introduce our earlier work [72] on generating boundary-aware proposals in TAL, which is the same as T-PSC and uses contrastive learning with the hard negative mining strategy. Instead of using cosine similarity, we conduct the experiments on T-PSC to verify the contrastive loss performance on video. Specifically, the main approach is to use the T-PSC loss function proposed in this paper to replace the cosine similarity loss function used via BAPG. The experiment results of our T-PSC with the state-of-the-art method in THUMOS14 are shown in Table 5.

As can be seen from Table 5, after replacing the cosine similarity loss function in BAPG with T-PSC, the model performance is improved compared to both the original model and BAPG. Especially when tIoU = 0.7, taking TriDet as an example, T-PSC can obtain a gain of 0.1% based on BAPG, which is 1.05% compared to the original TriDet. The experiment results of our T-PSC with the state-of-the-art method in ActivityNet-1.3 are shown in Table 6. It can be seen from the data in Table 6 that T-PSC comprehensively improves the performance of the existing model. Although the videos in ActivityNet-1.3 are more complicated and variable than THUMOS14, T-PSC can still improve the original TriDet model and BAPG model by 0.3% and 0.1% at tIoU = 0.8, respectively. The experiment results show that T-PSC is effective and can improve the TAL performance.

Table 5. Performance comparison on THUMOS14 in terms of mAP at different tIoU thresholds. The "Avg" column denotes the average mAP in [0.3:0.1:0.7].

Model	THUMOS14 (%)					
	0.3	**0.4**	**0.5**	**0.6**	**0.7**	**Avg.**
BMN [45]	56.00	47.40	38.80	29.70	20.50	38.50
G-TAD [73]	54.50	47.60	40.30	30.80	23.40	39.30
A2Net [74]	58.60	54.10	45.50	32.50	17.20	41.60
VSGN [75]	66.70	60.40	52.40	41.00	30.40	50.20
ContextLoc [47]	68.30	63.80	54.30	41.80	26.20	50.90
AFSD [76]	67.30	62.40	55.50	43.70	31.10	52.00
TadTR [77]	74.80	69.10	60.10	46.60	32.80	56.70
Actionformer [42]	81.96	77.41	71.08	58.91	43.74	66.62
+ BAPG [72]	82.09	77.66	71.45	59.51	44.35	67.01
+ **T-PSC**	82.19	77.70	71.48	59.62	44.65	**67.13**
TriDet [43]	83.53	79.60	72.12	60.76	45.40	68.28
+ BAPG [72]	83.60	79.72	72.49	61.49	46.35	68.73
+ **T-PSC**	83.72	79.81	72.52	61.56	46.45	**68.81**

Table 6. Performance comparison on ActivityNet v1.3 in terms of mAP at different tIoU thresholds. The "Avg" columns denote the average mAP in [0.5:0.05:0.95].

Model	ActivityNet v1.3 (%)				
	0.5	**0.6**	**0.7**	**0.8**	**Avg.**
BMN [45]	50.11	44.21	37.61	28.83	33.91
G-TAD [73]	50.42	43.98	38.13	29.42	34.13
Actionformer [42]	54.67	48.25	41.85	32.91	36.56
+ BAPG [72]	54.80	48.38	41.97	32.93	36.61
+ **T-PSC**	54.89	48.40	42.01	32.95	**36.65**
TriDet [43]	54.71	48.54	42.34	32.93	36.76
+ BAPG [72]	54.85	48.58	42.50	33.15	36.83
+ **T-PSC**	54.95	48.62	42.58	33.26	**36.89**

4.4. Convergence Analysis

Figure 4 compares the performance of $\mathcal{L}_{\text{T-HNM}}$ and $\mathcal{L}_{\text{T-PSC}}$ during training. As shown in Figure 4a, $\mathcal{L}_{\text{T-PSC}}$ has a better convergence than $\mathcal{L}_{\text{T-HNM}}$. $\mathcal{L}_{\text{T-PSC}}$ decreases rapidly in the early phase of training. On the contrary, $\mathcal{L}_{\text{T-HNM}}$ decreases slowly since the optimization of $\mathcal{L}_{\text{T-HNM}}$ is too difficult. Figure 4b,c shows that $\mathcal{L}_{\text{T-PSC}}$ can achieve a higher performance faster. On the one hand, $\mathcal{L}_{\text{T-PSC}}$ reduces the model optimization difficulty by controlling the penalty strength. Thus, model training is accelerated. On the other hand, $\mathcal{L}_{\text{T-PSC}}$ still provides a relatively large penalty strength for hard negatives. Coupled with better training behavior, the final retrieval performance is also improved.

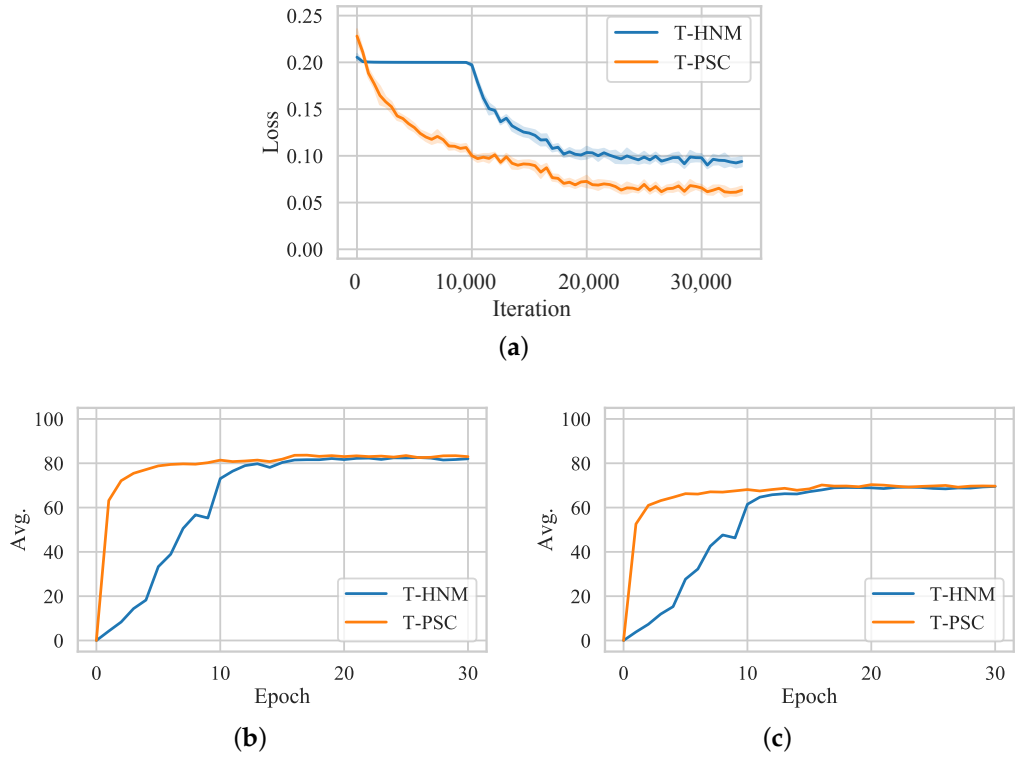


Figure 4. Training behavior on Flickr30K. (**a**) Loss. (**b**) Avg. of image-to-text. (**c**) Avg. of text-to-image.

5. Conclusions

By revisiting hard negative mining in contrastive learning, this paper proposes T-PSC to effectively distinguish hard negative samples in visual understanding tasks. In order to overcome the side effects of convergence difficulties caused by traditional hard negative mining methods, we define a metric for the penalty strength of negatives. We can use the penalty strength of hard negatives to quantitatively analyze and find the appropriate level for visual understanding models. Moreover, we can employ T-PSC to balance the penalty strength of hard negatives and the difficulty of model optimization. We find that a reasonable control of the penalty strength can speed up training and obtain discriminative visual presentations. Our T-PSC is flexible and can seamlessly combine with the current visual understanding models in a plug-and-play manner. In order to confirm that the characteristics of T-PSC can be generally applied to various tasks of visual understanding, we conduct extensive experiments. By combining it with models in the field of Image–Text Retrieval, we verify the feature representation capabilities of T-PSC in both the image and text modalities. By combining it with models in the field of video temporal localization, we discover the effectiveness of T-PSC in the video modality. In future work, we will explore the adaptive control of the penalty strength to avoid complicated parameter tuning and find the optimal penalty intensity for different visual understanding tasks.

Author Contributions: Conceptualization, H.Z. and Z.L.; Methodology, H.Z. and Z.L.; Software, H.Z. and Z.L.; Validation, H.Z., J.Y. and X.W.; Investigation, H.Z.; Writing—original draft, H.Z. and Z.L.; Writing—review & editing, H.Z. and C.G.; Visualization, H.Z. and J.Y.; Supervision, C.G. and C.F.; Project administration, C.G. and C.F.; Funding acquisition, C.G. and C.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Program of National Natural Science Foundation of China (No. 92067202).

Data Availability Statement: All data included in this study may be made available upon request via contacting the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 815–823.
2. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; Volume 33, pp. 18661–18673.
3. Luo, D.; Liu, C.; Zhou, Y.; Yang, D.; Ma, C.; Ye, Q.; Wang, W. Video cloze procedure for self-supervised spatio-temporal learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11701–11708.
4. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 9729–9738.
5. Fang, Z.; Wang, J.; Wang, L.; Zhang, L.; Yang, Y.; Liu, Z. Seed: Self-supervised distillation for visual representation. *arXiv* **2021**, arXiv:2101.04731.
6. Xiang, N.; Chen, L.; Liang, L.; Rao, X.; Gong, Z. Semantic-Enhanced Cross-Modal Fusion for Improved Unsupervised Image Captioning. *Electronics* **2023**, *12*, 3549. [CrossRef]
7. Xu, T.; Liu, X.; Huang, Z.; Guo, D.; Hong, R.; Wang, M. Early-Learning Regularized Contrastive Learning for Cross-Modal Retrieval with Noisy Labels. In Proceedings of the 30th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2022; pp. 629–637.
8. Qian, S.; Xue, D.; Fang, Q.; Xu, C. Integrating Multi-Label Contrastive Learning with Dual Adversarial Graph Neural Networks for Cross-Modal Retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4794–4811. [CrossRef] [PubMed]
9. Liu, Y.; Wu, J.; Qu, L.; Gan, T.; Yin, J.; Nie, L. Self-Supervised Correlation Learning for Cross-Modal Retrieval. *IEEE Trans. Multimed.* **2023**, *25*, 2851–2863. [CrossRef]
10. Park, J.; Lee, J.; Kim, I.J.; Sohn, K. Probabilistic Representations for Video Contrastive Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 14711–14721.
11. Wang, X.; Zhao, K.; Zhang, R.; Ding, S.; Wang, Y.; Shen, W. ContrastMask: Contrastive Learning To Segment Every Thing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 11604–11613.
12. Mohamed, M. Empowering deep learning based organizational decision making: A Survey. *Sustain. Mach. Intell. J.* **2023**, *3* . [CrossRef]
13. Mohamed, M. Agricultural Sustainability in the Age of Deep Learning: Current Trends, Challenges, and Future Trajectories. *Sustain. Mach. Intell. J.* **2023**, *4*. [CrossRef]
14. Sleem, A. Empowering Smart Farming with Machine Intelligence: An Approach for Plant Leaf Disease Recognition. *Sustain. Mach. Intell. J.* **2022**, *1*. [CrossRef]
15. Yang, C.; An, Z.; Cai, L.; Xu, Y. Mutual Contrastive Learning for Visual Representation Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 3045–3053.
16. Lee, K.H.; Chen, X.; Hua, G.; Hu, H.; He, X. Stacked cross attention for image-text matching. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 201–216.
17. Li, K.; Zhang, Y.; Li, K.; Li, Y.; Fu, Y. Visual semantic reasoning for image-text matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4654–4662.
18. Liu, C.; Mao, Z.; Zhang, T.; Xie, H.; Wang, B.; Zhang, Y. Graph structured network for image-text matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020; pp. 10921–10930.
19. Zhang, K.; Mao, Z.; Wang, Q.; Zhang, Y. Negative-Aware Attention Framework for Image-Text Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 15661–15670.
20. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737.
21. Xuan, H.; Stylianou, A.; Liu, X.; Pless, R. Hard negative examples are hard, but useful. In Proceedings of the European Conference on Computer Vision (ECCV), Online, 23–28 August 2020; pp. 126–142.
22. Faghri, F.; Fleet, D.J.; Kiros, J.R.; Fidler, S. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv* **2017**, arXiv:1707.05612.
23. Oh Song, H.; Xiang, Y.; Jegelka, S.; Savarese, S. Deep metric learning via lifted structured feature embedding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4004–4012.
24. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality Reduction by Learning an Invariant Mapping. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1735–1742.
25. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 1597–1607.

26. Chen, X.; Fan, H.; Girshick, R.; He, K. Improved baselines with momentum contrastive learning. *arXiv* **2020**, arXiv:2003.04297.
27. Reiss, T.; Hoshen, Y. Mean-shifted contrastive loss for anomaly detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 2155–2162.
28. Bai, Y.; Wang, A.; Kortylewski, A.; Yuille, A. CoKe: Contrastive Learning for Robust Keypoint Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–7 January 2023; pp. 65–74.
29. Fan, R.; Poggi, M.; Mattoccia, S. Contrastive Learning for Depth Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Vancouver, BC, Canada, 18–22 June 2023; pp. 3225–3236.
30. Chen, H.; Ding, G.; Liu, X.; Lin, Z.; Liu, J.; Han, J. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020; pp. 12655–12663.
31. Lan, H.; Zhang, P. Learning and Integrating Multi-Level Matching Features for Image-Text Retrieval. *IEEE Signal Process. Lett.* **2021**, *29*, 374–378. [CrossRef]
32. Li, S.; Lu, A.; Huang, Y.; Li, C.; Wang, L. Joint Token and Feature Alignment Framework for Text-Based Person Search. *IEEE Signal Process. Lett.* **2022**, *29*, 2238–2242. [CrossRef]
33. Wang, H.; Zhang, Y.; Ji, Z.; Pang, Y.; Ma, L. Consensus-aware visual-semantic embedding for image-text matching. In Proceedings of the European Conference on Computer Vision (ECCV), Online, 23–28 August 2020; pp. 18–34.
34. Chen, J.; Hu, H.; Wu, H.; Jiang, Y.; Wang, C. Learning the Best Pooling Strategy for Visual Semantic Embedding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 15789–15798.
35. Zhang, Q.; Lei, Z.; Zhang, Z.; Li, S.Z. Context-Aware Attention Network for Image-Text Retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020; pp. 3536–3545.
36. Frome, A.; Corrado, G.S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; Mikolov, T. Devise: A deep visual-semantic embedding model. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, 5–10 December 2013; pp. 2121–2129.
37. Lin, T.; Zhao, X.; Shou, Z. Single shot temporal action detection. In Proceedings of the 25th ACM International Conference on Multimedia, New York, NY, USA, 23–27 October 2017; pp. 988–996.
38. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the 2016 European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
39. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 June 2017; pp. 7263–7271.
40. Long, F.; Yao, T.; Qiu, Z.; Tian, X.; Luo, J.; Mei, T. Gaussian temporal awareness networks for action localization. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 344–353.
41. Liu, Y.; Ma, L.; Zhang, Y.; Liu, W.; Chang, S.F. Multi-Granularity Generator for Temporal Action Proposal. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 3604–3613.
42. Zhang, C.L.; Wu, J.; Li, Y. Actionformer: Localizing moments of actions with transformers. In Proceedings of the 2022 European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 492–510.
43. Shi, D.; Zhong, Y.; Cao, Q.; Ma, L.; Li, J.; Tao, D. TriDet: Temporal Action Detection with Relative Boundary Modeling. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 18857–18866.
44. Lin, T.; Zhao, X.; Su, H.; Wang, C.; Yang, M. Bsn: Boundary sensitive network for temporal action proposal generation. In Proceedings of the 2018 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
45. Lin, T.; Liu, X.; Li, X.; Ding, E.; Wen, S. BMN: Boundary-Matching Network for Temporal Action Proposal Generation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October– 2 November 2019; pp. 3888–3897.
46. Zeng, R.; Huang, W.; Tan, M.; Rong, Y.; Zhao, P.; Huang, J.; Gan, C. Graph Convolutional Networks for Temporal Action Localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7094–7103.
47. Zhu, Z.; Tang, W.; Wang, L.; Zheng, N.; Hua, G. Enriching Local and Global Contexts for Temporal Action Localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; pp. 13516–13525.
48. Chen, S.; Zhao, Y.; Jin, Q.; Wu, Q. Fine-grained video-text retrieval with hierarchical graph reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020; pp. 10638–10647.
49. Croitoru, I.; Bogolin, S.V.; Leordeanu, M.; Jin, H.; Zisserman, A.; Albanie, S.; Liu, Y. TeachText: CrossModal Generalized Distillation for Text-Video Retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; pp. 11583–11593.
50. Bueno-Benito, E.; Vecino, B.T.; Dimiccoli, M. Leveraging Triplet Loss for Unsupervised Action Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Vancouver, BC, Canada, 18–22 June 2023; pp. 4921–4929.

51. Tang, Z.; Huang, J. Harmonious Multi-Branch Network for Person Re-Identification with Harder Triplet Loss. *ACM Trans. Multimed. Comput. Commun. Appl.* **2022**, *18*, 1–21 [CrossRef]
52. Tian, M.; Wu, X.; Jia, Y. Adaptive Latent Graph Representation Learning for Image-Text Matching. *IEEE Trans. Image Process.* **2023**, *32*, 471–482. [CrossRef]
53. Zhang, K.; Mao, Z.; Liu, A.A.; Zhang, Y. Unified Adaptive Relevance Distinguishable Attention Network for Image-Text Matching. *IEEE Trans. Multimed.* **2023**, *25*, 1320–1332. [CrossRef]
54. Liu, Y.; Liu, H.; Wang, H.; Liu, M. Regularizing Visual Semantic Embedding With Contrastive Learning for Image-Text Matching. *IEEE Signal Process. Lett.* **2022**, *29*, 1332–1336. [CrossRef]
55. Zhang, H.; Mao, Z.; Zhang, K.; Zhang, Y. Show your faith: Cross-modal confidence-aware network for image-text matching. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Virtual, 22 Febrary–1 March 2022; Volume 36, pp. 3262–3270.
56. Wang, Y.; Yang, H.; Qian, X.; Ma, L.; Lu, J.; Li, B.; Fan, X. Position focused attention network for image-text matching. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 10–16 August 2019; pp. 3792–3798.
57. Li, Z.; Guo, C.; Feng, Z.; Hwang, J.N.; Xue, X. Multi-View Visual Semantic Embedding. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI), Vienna, Austria, 23–29 July 2022; pp. 1130–1136.
58. van den Oord, A.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
59. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning, PMLR, Virtual, 6–7 August 2021; pp. 8748–8763.
60. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [CrossRef]
61. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
62. Idrees, H.; Zamir, A.R.; Jiang, Y.G.; Gorban, A.; Laptev, I.; Sukthankar, R.; Shah, M. The THUMOS challenge on action recognition for videos "in the wild". *Comput. Vis. Image Underst.* **2017**, *155*, 1–23. [CrossRef]
63. Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; Carlos Niebles, J. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 961–970.
64. Zeng, R.; Huang, W.; Tan, M.; Rong, Y.; Zhao, P.; Huang, J.; Gan, C. Graph Convolutional Module for Temporal Action Localization in Videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 6209–6223. [CrossRef]
65. Liu, X.; Hu, Y.; Bai, S.; Ding, F.; Bai, X.; Torr, P.H. Multi-shot temporal event localization: A benchmark. In Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 12596–12606.
66. Wei, J.; Xu, X.; Yang, Y.; Ji, Y.; Wang, Z.; Shen, H.T. Universal Weighting Metric Learning for Cross-Modal Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020; pp. 13005–13014.
67. Wei, J.; Xu, X.; Wang, Z.; Wang, G. Meta Self-Paced Learning for Cross-Modal Matching. In Proceedings of the 29th ACM International Conference on Multimedia (ACM MM), Chengdu, China, 20–24 October 2021; pp. 3835–3843.
68. Chen, T.; Deng, J.; Luo, J. Adaptive Offline Quintuplet Loss for Image-Text Matching. In Proceedings of the European Conference on Computer Vision (ECCV), Online, 23–28 August 2020; pp. 549–565.
69. Huang, Z.; Niu, G.; Liu, X.; Ding, W.; Xiao, X.; Wu, H.; Peng, X. Learning with Noisy Correspondence for Cross-modal Matching. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–14 December 2021; Volume 34, pp. 29406–29419.
70. Liu, C.; Mao, Z.; Liu, A.A.; Zhang, T.; Wang, B.; Zhang, Y. Focus your attention: A bidirectional focal attention network for image-text matching. In Proceedings of the 27th ACM International Conference on Multimedia (ACM MM), Nice, France, 21–25 October 2019; pp. 3–11.
71. Diao, H.; Zhang, Y.; Ma, L.; Lu, H. Similarity Reasoning and Filtration for Image-Text Matching. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 4–6 November 2021; Volume 35, pp. 1218–1226.
72. Zhang, H.; Feng, C.; Yang, J.; Li, Z.; Guo, C. Boundary-Aware Proposal Generation Method for Temporal Action Localization. *arXiv* **2023**, arXiv:2309.13810.
73. Xu, M.; Zhao, C.; Rojas, D.S.; Thabet, A.; Ghanem, B. G-TAD: Sub-Graph Localization for Temporal Action Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020.
74. Yang, L.; Peng, H.; Zhang, D.; Fu, J.; Han, J. Revisiting Anchor Mechanisms for Temporal Action Localization. *IEEE Trans. Image Process.* **2020**, *29*, 8535–8548. [CrossRef] [PubMed]
75. Zhao, C.; Thabet, A.K.; Ghanem, B. Video Self-Stitching Graph Network for Temporal Action Localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; pp. 13658–13667.

76. Lin, C.; Xu, C.; Luo, D.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; Fu, Y. Learning Salient Boundary Feature for Anchor-free Temporal Action Localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 3320–3329.
77. Liu, X.; Wang, Q.; Hu, Y.; Tang, X.; Zhang, S.; Bai, S.; Bai, X. End-to-End Temporal Action Detection with Transformer. *IEEE Trans. Image Process.* **2022**, *31*, 5427–5441. [CrossRef] [PubMed]