Contrastive Losses Are Natural Criteria for Unsupervised Video Summarization

Zongshang Pang
Osaka University
pangzs@is.ids.osaka-u.ac.jp

Yuta Nakashima
Osaka University
n-yuta@ids.osaka-u.ac.jp

Mayu Otani
CyberAgent, Inc.
otani-mayu@cyberagent.co.jp

Hajime Nagahara
Osaka University
nagahara@ids.osaka-u.ac.jp

Abstract

Video summarization aims to select the most informative subset of frames in a video to facilitate efficient video browsing. Unsupervised methods usually rely on heuristic training objectives such as diversity and representativeness. However, such methods need to bootstrap the online-generated summaries to compute the objectives for importance score regression. We consider such a pipeline inefficient and seek to directly quantify the frame-level importance with the help of contrastive losses in the representation learning literature. Leveraging the contrastive losses, we propose three metrics featuring a desirable key frame: local dissimilarity, global consistency, and uniqueness. With features pre-trained on the image classification task, the metrics can already yield high-quality importance scores, demonstrating competitive or better performance than past heavily-trained methods. We show that by refining the pre-trained features with a lightweight contrastively learned projection module, the frame-level importance scores can be further improved, and the model can also leverage a large number of random videos and generalize to test videos with decent performance.

1. Introduction

Recently, deep neural networks have significantly advanced the development of efficient video summarization tools. The supervised workflow and evaluation protocols proposed by Zhang *et al*. [56] have become a cornerstone for most of the subsequent deep-learning-based supervised methods. Unsupervised methods avoid using annotated summaries by utilizing heuristic training objectives such as *diversity* and *representativeness* [35, 42, 33, 41, 62, 21, 22]. The diversity objective aims to enforce the dissimilarity between the key frame candidates, and the representativeness objective guarantees that the generated summaries can well reflect the major information in the original video.

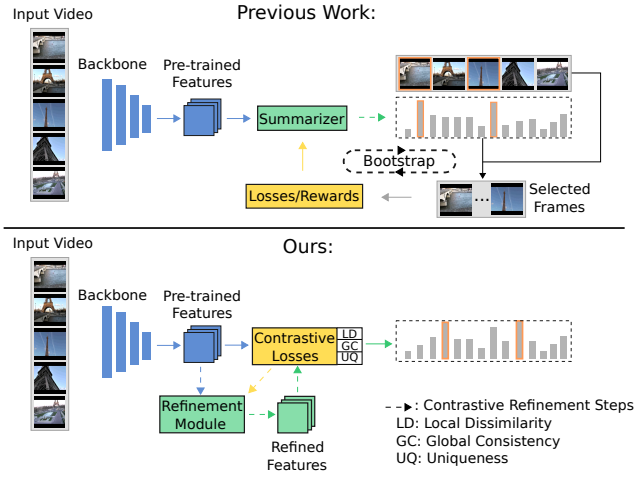

Figure 1: A comparison between our method and previous work.

The past unsupervised approaches focus on bootstrapping the summaries generated on the fly during training to evaluate their diversity and representativeness and then utilizing the resulting loss terms to train models. However, the basis for these algorithms is, by all means, the more fundamental elements of the summaries, *i.e*., the frames. The premise for producing a decent summary is that the correct frames are selected. Bootstrapping the online-generated summaries with poor quality seems less straightforward, if not redundant. Here we pose a question: *How do we directly quantify how much each frame contributes to the quality of the final summary?*

To answer this question, we start by defining two desirable properties for key frame candidates: *local dissimilarity* and *global consistency*. Inspired by the diversity objective, if a frame is excessively similar to its semantically close neighbors in the feature space, then this frame, to-

gether with its neighbors, lacks *local* dissimilarity, where the locality is defined in the feature space based on cosine similarity [63]. The information in such frames is often monotonous, as they appear multiple times in the video but rarely demonstrate any variations. Thus, we risk introducing redundancy to the final summary if such frames are considered key frames. On the other hand, merely selecting frames based on dissimilarity may wrongly incorporate noisy frames with very few semantically meaningful neighbors, thus not indicative of the video theme. Inspired by the representativeness objective, we consider frames consistent with the majority of frames in a video as being related to the central video theme, *i.e*., they are globally consistent. Eventually, we would like to select frames with a desirable level of local dissimilarity and global consistency so that the resulting summary can have well-balanced diversity and representativeness.

Interestingly, the above two metrics can be readily calculated by utilizing contrastive losses for image representation learning, *i.e*., alignment and uniformity [52] losses. The alignment loss calculates the distance between an image and a semantically related sample, *e.g*., an augmented version of the image. With a pool of semantically related samples, Zhuang *et al*. [63] defines the aggregation of these samples as a local neighborhood. The alignment loss can readily measure the local dissimilarity of each frame in such a neighborhood. The uniformity loss aims to regularize the proximity of the overall features and would be high for closely distributed features. Therefore, it can be conveniently leveraged to measure the semantic consistency between frames. The two losses can be further utilized to perform *contrastive refinement* of the features, which we will show is also more efficient than previous methods.

Nonetheless, background frames with complex contents related to many other frames in the video can also be locally dissimilar and globally consistent. For instance, street scenes will likely appear throughout a car-accident video. Such frames can still be relatively dissimilar because of the moving objects. However, on average, they may be consistent with most frames. Luckily, we can exclude them by leveraging an assumption that these background frames tend to appear in many different videos and thus are not *unique* to their associated videos, *e.g*., street scenes in videos about car accidents, parades, city tours, *etc*. Based on this assumption, we train a *unqueness filter* to filter out ambiguous background frames, which can be readily incorporated into the aforementioned contrastive losses. We illustrate our proposed method together with a comparison with previous work in Fig. 1.

Contributions. Unlike previous work bootstrapping the online-generated summaries, we propose three metrics, local dissimilarity, global consistency, and uniqueness to directly quantify frame-level importance based on theoretically motivated contrastive losses [52]. This is significantly more efficient than previous methods. Specifically, we can obtain competitive F1 scores and better correlation coefficients [39] on SumMe [14] as well as TVSum [44] with the first two metrics calculated using only ImageNet [25] pre-trained features *without any further training*. Moreover, by contrastively refining the features along with training the proposed uniqueness filter, we can further improve the performance given only random videos sampled from the Youtube8M dataset [1]

2. Related Work

Before the dawn of deep learning, video summarization was handled by hand-crafted features and heavy optimization schemes [44, 15, 30, 60, 14]. Zhang *et al*. [56] applied a deep architecture involving a bidirectional recurrent neural network (RNN) to select key frames in a supervised manner, forming a cornerstone for many subsequent work [58, 59, 57, 12, 51]. There are also methods that leverage attention mechanisms [11, 4, 20, 19, 33, 13, 32, 29] or fully convolutional networks [42, 34]. Exploring spatiotemporal information by jointly using RNNs and convolutional neural networks (CNNs) [55, 8, 10] or using graph convolution networks [24, 40] also delivered decent performance. There is also an attempt at leveraging texts to aid video summarization [36].

Unsupervised methods mainly exploit two heuristics: diversity and representativeness. Zhou *et al*. [62] and subsequent work [6, 28] tried to maximize the diversity and representativeness rewards of the produced summaries. Some work [35, 40, 42, 41] applied the repelling regularizer [61] to regularize the similarities between the mid-level frame features in the generated summaries to guarantee their diversity. Mahasseni *et al*. [35] and subsequent work [21, 17, 33] used the reconstruction-based VAE-GAN structure to generate representative summaries, while Rochan *et al*. [41] exploited reconstructing unpaired summaries.

Though also inspired by diversity and representativeness, our methodology differs from all the above unsupervised approaches. Concretely, we directly quantify how much each frame contributes to the final summary by leveraging contrastive loss functions in the representation learning literature [52], which enables us to achieve competitive or better performance compared to these approaches *without any training*. Moreover, we perform contrastive refinement of the features to produce better importance scores, which obviates bootstrapping online generated summaries and is much more efficient than previous work. To the best of our knowledge, we are the first to leverage contrastive learning for video summarization.

3. Preliminaries

As contrastive learning is essential to our approach, we introduce the preliminaries focusing on instance discrimination [54].

3.1. Instance Discrimination via the InfoNCE Loss

As an integral part of unsupervised image representation learning, contrastive learning [7] has been attracting researchers' attention over the years and has been constantly improved to deliver representations with outstanding transferability [54, 17, 63, 52, 18, 38, 47, 5]. Formally, given a set $\mathcal{D} = \{I_n\}_{n=1}^N$ of N images, contrastive representation learning aims to learn an encoder f_θ with learnable θ such that the resulting features $f_\theta(I_n)$ can be readily leveraged by downstream vision tasks. A theoretically founded [38] loss function with favorable empirical behaviors [50] is the so-called InfoNCE loss [38]:

$$\mathcal{L}_{\text{InfoNCE}} = \sum_{I \in \mathcal{D}} -\log \frac{e^{f_\theta(I) \cdot f_\theta(I')/\tau}}{\sum_{J \in \mathcal{D}'(I)} e^{f_\theta(I) \cdot f_\theta(J)/\tau}}, \quad (1)$$

where I' is a positive sample for $I \in X$, usually obtained through data augmentation, and $\mathcal{D}'(I)$ includes I' as well as all negative samples, *e.g.*, any other images. The operator "$\cdot$" is the inner product, and τ is a temperature parameter. Therefore, the loss aims at pulling closer the feature of an instance with that of its augmented views while repelling it from those of other instances, thus performing instance discrimination.

3.2. Contrastive Learning via Alignment and Uniformity

When normalized onto the unit hypersphere, the contrastively learned features that tend to deliver promising downstream performance possess two interesting properties. That is, the semantically related features are usually closely located on the sphere regardless of their respective details, and the overall features' information is retained as much as possible [38, 18, 47]. Wang *et al*. [52] defined these two properties as *alignment* and *uniformity*.

The alignment metric computes the distance between the positive pairs [52]:

$$\mathcal{L}_{\text{align}}(\theta, \alpha) = \mathop{\mathbb{E}}_{(I,I') \sim p_{\text{pos}}} [\|f_\theta(I) - f_\theta(I')\|_2^\alpha], \quad (2)$$

where $\alpha > 0$, and p_{pos} is the distribution of positive pairs (*i.e.*, an original image and its augmentation).

The uniformity is defined as the average pairwise Gaussian potential between the overall features:

$$\mathcal{L}_{\text{uniform}}(\theta, \beta) = \log \left(\mathop{\mathbb{E}}_{I, J \overset{\text{i.i.d}}{\sim} p_{\text{data}}} [e^{-\beta \|f_\theta(I) - f_\theta(J)\|_2^2}] \right), \quad (3)$$

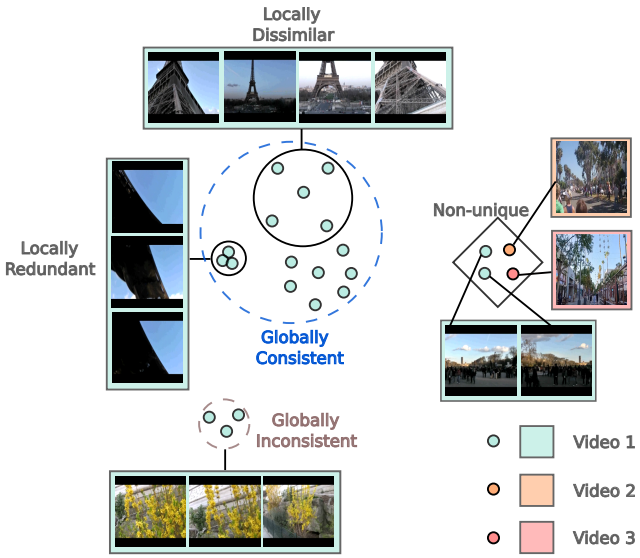

Figure 2: A conceptual illustration for the three metrics: local dissimilarity, global consistency, and uniqueness in the semantic space. The images come from the SumMe [14] and TVSum [44] datasets. The dots with the same color indicate features from the same video. For concise demonstration, we only show one frame for "video 2" and "video 3" to show the idea of uniqueness.

where p_{data} is usually approximated by the empirical data distribution, and β is usually set to 2 as recommended by [52]. This metric encourages the overall feature distribution on the unit hypersphere to approach a uniform distribution and can also be directly used to measure the uniformity of feature distributions [50]. Moreover, Eq. (3) approximates the log of the denominator of Eq. (1) when the number of negative samples goes to infinity [52]. As proved in [52], jointly minimizing Eqs. (2) and (3) can achieve better alignment and uniformity of the features, *i.e.*, they are locally clustered and globally uniform [50].

In this paper, we use Eq. (2) to compute the distance/dissimilarity between the semantically close video frame features to measure frame importance in terms of local dissimilarity. We then use the proposed variant of Eq. (3) to measure the proximity between a specific frame and the overall information of the associated video to estimate their semantic consistency. Moreover, utilizing the two losses, we learn a nonlinear projection of the pre-trained features so that the projected features are more locally aligned and globally uniform.

4. Proposed Method

Unlike previous work that bootstraps the inaccurate candidate summaries generated on the fly, we take a more straightforward perspective to quantify frame importance

directly, as we believe it is highly inefficient to deal with the infinitely many collections of frames. To quantify frame importance, we define three metrics: local dissimilarity, global consistency, and uniqueness. We provide a conceptual illustration in Fig. 2.

4.1. Local Dissimilarity

Inspired by the diversity objective, we consider frames likely to result in a diverse summary as those conveying diverse information even when compared to their semantic nearest neighbors.

Formally, given a video $\mathbf{V}$, we first extract deep features using ImageNet [25] pre-trained backbone, *e.g.*, GoogleNet [45], denoted as F, such that $F(\mathbf{V}) = \{\mathbf{x}_t\}_{t=1}^{T}$, where $\mathbf{x}_t$ represents the deep feature for the t-th frame in $\mathbf{V}$, and T is the total number of frames in $\mathbf{V}$. Each feature is L2-normalized such that $\|\mathbf{x}_t\|_2 = 1$.

To define local dissimilarity for frames in $\mathbf{V}$, we first use cosine similarity to retrieve for each frame $\mathbf{x}_t$ a set $\mathcal{N}_t$ of top $K = aT$ neighbors, where a is a hyperparameter and K is rounded to the nearest integer. The local dissmilarity metric for $\mathbf{x}_t$ is an empirical approximation of Eq. (2), defined as the local alignment loss:

$$\mathcal{L}_{\text{align}}(\mathbf{x}_t) = \frac{1}{|\mathcal{N}_t|} \sum_{\mathbf{x} \in \mathcal{N}_t} \|\mathbf{x}_t - \mathbf{x}\|_2^2, \tag{4}$$

which measures the distance/dissimilarity between $\mathbf{x}_t$ and its semantic neighbors.

The larger $\mathcal{L}_{\text{align}}(\mathbf{x}_t)$ is, the more dissimilar $\mathbf{x}_t$ is with its neighbors. Thus, if a frame has a certain distance from even its nearest neighbors in the semantic space, the frames in their local neighborhood are likely to convey diverse but still semantically cohesive information and hence are desirable key frame candidates. $\mathcal{L}_{\text{align}}(\mathbf{x}_t)$ can be directly used as the importance score of $\mathbf{x}_t$ after proper scaling.

4.2. Global Consistency

$\mathcal{N}_t$ may contain semantically irrelevant frames if $\mathbf{x}_t$ has very few semantic neighbors in the video. Therefore, merely using Eq. (4) as framewise importance scores is insufficient.

Inspired by the reconstruction-based representativeness objective [35], we define another metric called global consistency to quantify how consistent a frame is with the video gist by a modified uniformity loss based on Eq. (3):

$$\mathcal{L}_{\text{uniform}}(\mathbf{x}_t) = \log \left(\frac{1}{T-1} \sum_{\substack{\mathbf{x} \neq \mathbf{x}_t, \\ \mathbf{x} \in F(\mathbf{V})}} e^{-2\|\mathbf{x}_t - \mathbf{x}\|_2^2} \right), \tag{5}$$

$\mathcal{L}_{\text{uniform}}(\mathbf{x}_t)$ measures the proximity between $\mathbf{x}_t$ and the remaining frames, bearing similarity to the reconstruction- and K-medoids-based objectives in [35, 62], but only using a single frame instead of a collection of frames for reconstruction and obviating training an autoencoder [35] or a policy network [62].

4.3. Contrastive Refinement

Eqs. (4) and (5) are calculated using deep features pre-trained on image classification tasks, which may not necessarily well possess the local alignment and global uniformity as discussed in Section 3.2.

Hamilton *et al.* [16] contrastively refines the self-supervised vision transformer features [3] for unsupervised semantic segmentation. They freeze the feature extractor for better efficiency and only train a lightweight projector. Inspired by this work, we also avoid fine-tuning the heavy feature extractor, a deep CNN in our case, but instead only train a lightweight module appended to it.

Formally, given features $F(\mathbf{V})$ from the frozen backbone for a video, we feed them to a learnable module to obtain $\mathbf{z}_t = G_\theta(\mathbf{x}_t)$, where $\mathbf{z}_t$ is L2-normalized[1]. The nearest neighbors in $\mathcal{N}_t$ for each frame are still determined using the pre-trained features $\{\mathbf{x}_t\}_{t=1}^{T}$, which have been shown to be a good proxy for semantic similarities [48, 16]. Similar to [53, 63], we also observed collapsed training when directly using the learnable features for nearest neighbor retrieval, so we stick to using the frozen features.

With the learnable features, the alignment (local dissimilarity) and uniformity (global consistency) losses become [2]

$$\mathcal{L}_{\text{align}}(\mathbf{z}_t; \theta) = \frac{1}{|\mathcal{N}_t|} \sum_{\mathbf{z} \in \mathcal{N}_t} \|\mathbf{z}_t - \mathbf{z}\|_2^2, \tag{6}$$

$$\mathcal{L}_{\text{uniform}}(\mathbf{z}_t; \theta) = \log \left(\frac{1}{T-1} \sum_{\substack{\mathbf{z} \neq \mathbf{z}_t, \\ \mathbf{z} \in G_\theta(F(\mathbf{V}))}} e^{-2\|\mathbf{z}_t - \mathbf{z}\|_2^2} \right), \tag{7}$$

The joint loss function is thus:

$$\mathcal{L}(\mathbf{z}_t; \theta) = \mathcal{L}_{\text{align}}(\mathbf{z}_t; \theta) + \lambda_1 \mathcal{L}_{\text{uniform}}(\mathbf{z}_t; \theta), \tag{8}$$

where λ_1 is a hyperparameter balancing the two loss terms.

During the contrastive refinement, frames that have semantically meaningful nearest neighbors and are consistent with the video gist will have an $\mathcal{L}_{\text{align}}$ and an $\mathcal{L}_{\text{uniform}}$ that mutually resist each other. Specifically, when a nontrivial number of frames beyond $\mathcal{N}_t$ also share similar semantic structures with the anchor $\mathbf{z}_t$, these frames function as "hard negatives" that prevent $\mathcal{L}_{\text{align}}$ to be easily minimized [63, 50]. Therefore, only frames with both moderate local

[1] We leave out the L2-normalization operator for notation simplicity.

[2] We slightly abuse the notation of $\mathcal{L}$ to represent losses both before and after transformation by G_θ.

dissimilarity and global consistency will have balanced values for the two losses. In contrast, the other frames tend to have extreme values compared to those before the refinement.

4.4. The Uniqueness Filter

The two metrics defined above overlook the fact that the locally dissimilar and globally consistent frames can be background frames with complex content that may be related to most of the frames in the video. For instance, dynamic city views can be ubiquitous in videos recorded in a city.

Intriguingly, we can filter out such frames by directly leveraging a common property of theirs: They tend to appear in many different videos that may not necessarily share a common theme and may or may not have a similar context, *e.g.*, city views in videos about car accidents, city tours, and city parades, *etc.*, or scenes with people moving around that can appear in many videos with very different contexts. Therefore, such frames are not *unique* to their associated videos. Similar reasoning is exploited in weakly-supervised action localization literature [37, 31, 26], where a single class is used to capture all the background frames. However, we aim to pinpoint background frames in an unsupervised manner. Moreover, we do not use a single prototype to detect all the backgrounds as it is too restricted [27]. Instead, we treat each frame as a potential background prototype to look for highly-activated frames in random videos, which also determines the backgourndness of the frame itself.

To design a filter for eliminating such frames, we introduce an extra loss to Eq. (8) that taps into cross-video samples. For computational efficiency, we aggregate the frame features in a video $\mathbf{V}_k$ with T_k frames into segments with an equal length of m. The learnable features $\mathbf{z}$ in each segment are average-pooled and L2-normalized to obtain segment features $\mathcal{S}_k = \{\mathbf{s}_l\}_{l=1}^{|\mathcal{S}_k|}$ with $|\mathcal{S}_k| = \lfloor T_k/m \rfloor$. To measure the proximity of a frame with frames from a randomly sampled batch of videos $\mathcal{B}$ (represented now as segment features) including $\mathcal{S}_k$, we again leverage Eq. (3) to define the uniqueness loss for $\mathbf{z}_t \in \mathbf{V}_k$ as:

$$\mathcal{L}_{\text{unique}}(\mathbf{z}_t;\theta) = \log\left(\frac{1}{A}\sum_{\mathcal{S}\in\mathcal{B}/\mathcal{S}_k}\sum_{\mathbf{s}\in\mathcal{S}} e^{-2\|\mathbf{z}_t-\mathbf{s}\|_2^2}\right), \quad (9)$$

where $A = \sum_{\mathcal{S}\in\mathcal{B}/\mathcal{S}_k}|\mathcal{S}|$ is the normalization factor. A large value of $\mathcal{L}_{\text{unique}}$ means that $\mathbf{z}_t$ has nontrivial similarity with segments from randomly gathered videos, indicating that it is likely to be a background frame.

When jointly optimized with Eq. (8), Eq. (9) will be easy to minimize for unique frames, for which most of $\mathbf{s}$ are semantically irrelevant and can be safely repelled. It is not the case for the background frames that have semantically similar $\mathbf{s}$, as the local alignment loss keeps strengthening the closeness of semantically similar features.

As computing Eq. (9) needs random videos, it is not as straightforward to convert Eq. (9) to importance scores after training. To address this, we simply train a model $H_{\hat{\theta}}$ whose last layer is sigmoid to mimic $1 - \bar{\mathcal{L}}_{\text{unique}}(\mathbf{z}_t;\theta)$, where $\bar{\mathcal{L}}_{\text{unique}}(\mathbf{z}_t;\theta)$ is $\mathcal{L}_{\text{unique}}(\mathbf{z}_t;\theta)$ scaled to $[0,1]$ over t. Denoting $y_t = 1 - \text{sg}(\bar{\mathcal{L}}_{\text{unique}}(\mathbf{z}_t;\theta))$ and $r_t = H_{\hat{\theta}}(\text{sg}(\mathbf{z}_t))$, where "sg" stands for stop gradients, we define the loss for training the model as

$$\mathcal{L}_{\text{filter}}(\mathbf{z}_t;\hat{\theta}) = -y_t \log r_t + (1-y_t)\log(1-r_t). \quad (10)$$

4.5. The Full Loss and Importance Scores

With all the components, the loss for each frame in a video is:

$$\begin{aligned}\mathcal{L}(\mathbf{z}_t;\theta,\hat{\theta}) = \mathcal{L}_{\text{align}}(\mathbf{z}_t;\theta) + \lambda_1\mathcal{L}_{\text{uniform}}(\mathbf{z}_t;\theta)\\ +\lambda_2\mathcal{L}_{\text{unique}}(\mathbf{z}_t;\theta) + \lambda_3\mathcal{L}_{\text{filter}}(\mathbf{z}_t;\hat{\theta}),\end{aligned} \quad (11)$$

where we fix both λ_2 and λ_3 as 0.1 and only tune λ_1.

Scaling the local dissimilarity, global consistency, and uniqueness scores to $[0,1]$ over t, the frame-level importance score is simply defined as:

$$p_t = \bar{\mathcal{L}}_{\text{align}}(\mathbf{z}_t;\theta)\bar{\mathcal{L}}_{\text{uniform}}(\mathbf{z}_t;\theta)\bar{H}_{\hat{\theta}}(\mathbf{z}_t) + \epsilon, \quad (12)$$

which means that the importance scores will be high only when all three terms have nontrivial magnitude. ϵ is for avoiding zero values in the importance scores, which helps stabilize the knapsack algorithm for generating final summaries. As the scores are combined from three independent metrics, they tend not to have the temporal smoothness as possessed by the scores given by RNN [56] or attention networks [11]. We thus simply Gaussian-smooth the scores in each video to align with previous work in terms of temporal smoothness of scores.

5. Experiments

5.1. Datasets and Settings

Datasets. Following previous work, we evaluate our method on two benchmarks: TVSum [44] and SumMe [14]. TVSum contains 50 YouTube videos, each annotated by 20 annotators in the form of importance scores for every two-second-long shot. SumMe includes 25 videos, each with 15-18 reference binary summaries. We follow [56] to use OVP (50 videos) and YouTube (39 videos) [9] to augment TVSum and SumMe. Moreover, to test if our unsupervised approach can leverage a larger amount of videos, we randomly selected about 10,000 videos from the Youtube8M dataset [1], which has 3,862 video classes with highly diverse contents.

Evaluation Setting. Again following previous work, we evaluate model performance using five-fold cross-validation, where the dataset (TVSum or SumMe) is randomly divided into five splits. The reported results are averaged over five splits. In the canonical setting (C) [56], the training is only done on the original splits of the two evaluation datasets. In the augmented setting (A) [56], we augment the training set in each fold with three other datasets (*e.g.*, SumMe, YouTube, and OVP when TVSum is used for evaluation). In the transfer setting (T) [56], all the videos in TVSum (or SumMe) are used for testing, and the other three datasets are used for training. Moreover, we introduce an extra transfer setting where the training is solely conducted on the collected Youtube8M videos, and the evaluation is done on TVSum or SumMe. This setting is designed for testing if our model can benefit from a larger amount of data.

5.2. Evaluation Metrics

F1 score. Denoting A as the set of frames in a ground-truth summary and B as the set of frames in the corresponding generated summary, we can calculate the precision and recall as follows:

$$\text{Precision} = \frac{|A \cap B|}{|A|}, \ \text{Recall} = \frac{|A \cap B|}{|B|}, \tag{13}$$

with which we can calculate the F1 score by

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{14}$$

We follow [56] to deal with multiple ground-truth summaries and to convert importance scores into summaries.

Rank correlation coefficients. Recently, Otani *et al*. [39] demonstrated that the F1 scores are unreliable and can be pretty high for even randomly generated summaries. They proposed to use rank correlation coefficients, namely Kendall's τ [23] and Spearman's ρ [2], to measure the correlation between the predicted and the ground-truth importance scores. For each video, we first calculate the coefficient value between the predicted importance scores and each annotator's scores and then take the average over the total number of annotators for that video. The final results are obtained by averaging over all the videos.

5.3. Implementation Details

We follow previous work to use GoogleNet [45] pre-trained features for standard experiments. For experiments with Youtube8M videos, we use the quantized Inception-V3 [46] features provided by the dataset [1]. Both kinds of features are pre-trained on ImageNet [25]. The contrastive refinement module appended to the feature backbone is a lightweight Transformer encoder [49], and so is the uniqueness filter. More architecture and training details can be found in Section 1 in the supplementary.

Following [42], we restricted each video to have an equal length with random sub-sampling for longer videos and nearest-neighbor interpolation for shorter videos. Similar to [42], we did not observe much difference when using different lengths, and we fixed the length to 200 frames, which is quite efficient for the training.

We tune two hyperparameters: The ratio a that determines the size of the nearest neighbor set $\mathcal{N}_t$ and the coefficient λ_1 that controls the balance between the alignment and uniformity losses. Their respective effect will be demonstrated through the ablation study in Section 3 in the supplementary.

5.4. Quantitative Results

In this section, we compare our results with previous work and do the ablation study for different components of our method.

Importance scores calculated with only pre-trained features As shown in Tables 1 and 2, $\bar{\mathcal{L}}^*_{\text{align}}$ and $\bar{\mathcal{L}}^*_{\text{uniform}}$ directly computed with GoogleNet [45] pre-trained features already surpass most of the methods in terms of τ, ρ and F1 score. Especially, the correlation coefficient τ and ρ surpass even supervised methods, *e.g.*, (0.1345, 0.1776) v.s. dpLSTM's (0.0298, 0.0385) and SumGraph's (0.094, 0.138) for TVSum. Though DR-DSN$_{2000}$ has slightly better performance in terms of τ and ρ for TVSum, it has to reach the performance after bootstrapping the online-generated summaries for 2000 epochs while our results are directly obtained with simple computations using the same pre-trained features as those also used by DR-DSN.

More training videos are needed for the contrastive refinement. For the results in Tables 1 and 2, the maximum number of training videos is only 159, coming from the SumMe augmented setting. For the canonical setting, the training set size is 40 videos for TVSum and 20 for SumMe. Without experiencing many videos, the model tends to overfit each specific video and cannot generalize well. This is similar to the observation in contrastive representation learning that a larger amount of data (from a larger dataset or obtained from data augmentation) helps the model generalize [5, 3]. Therefore, the contrastive refinement results in Tables 1 and 2 hardly outperform those computed using pre-trained features.

Contrastive refinement with Youtube8M on TVSum. The model may better generalize to the test videos given sufficient training videos. This can be validated by the results for TVSum in Table 3. After the contrastive refinement, the results with only $\bar{\mathcal{L}}^*_{\text{align}}$ are improved from (0.0595, 0.0779) to (0.0911, 0.1196) for τ and ρ. We can also observe improvement over $\bar{\mathcal{L}}^*_{\text{align}}$&$\bar{\mathcal{L}}^*_{\text{uniform}}$ brought by contrastive refinement.

Contrastive refinement with Youtube8M on SumMe. The reference summaries in SumMe are binary scores, and

Table 1: Ablation results in terms of τ and ρ together with their comparisons with previous work in the canonical setting. Since no previous work provided τ and ρ for the other two settings, we provide our results for them in Section 2 in the supplementary. DR-DSN$_{60}$ means the DR-DSN trained for 60 epochs and similarly for DR-DSN$_{2000}$. Our scores with superscript $*$ are directly computed from pre-trained features. The results were generated with $(\lambda_1, a) = (0.5, 0.1)$. **Boldfaced** scores represent the best among supervised methods and human evaluations, and blue scores are the best among the unsupervised methods. Please refer to the text for analyses of the results.

		TVSum		SumMe	
		τ	ρ	τ	ρ
	Human baseline [43]	0.1755	0.2019	**0.1796**	**0.1863**
Supervised	VASNet [11, 43]	0.1690	0.2221	0.0224	0.0255
	dppLSTM [56, 39]	0.0298	0.0385	-0.0256	-0.0311
	SumGraph [40]	0.094	0.138	-	-
	Multi-ranker [43]	**0.1758**	**0.2301**	0.0108	0.0137
Unsupervised (previous)	DR-DSN$_{60}$[62, 39]	0.0169	0.0227	0.0433	0.0501
	DR-DSN$_{2000}$[62, 43]	0.1516	0.198	-0.0159	-0.0218
	SUM-FCN$_{unsup}$[42, 43]	0.0107	0.0142	0.0080	0.0096
	SUM-GAN [35, 43]	-0.0535	-0.0701	-0.0095	-0.0122
	CSNet+GL+RPE [22]	0.070	0.091	-	-
Unsupervised (ours, w.o. training)	$\bar{\mathcal{L}}^*_{\text{align}}$	0.1055	0.1389	0.0960	0.1173
	$\bar{\mathcal{L}}^*_{\text{align}}$ & $\bar{\mathcal{L}}^*_{\text{uniform}}$	0.1345	0.1776	0.0819	0.1001
Unsupervised (ours, w. training)	$\bar{\mathcal{L}}_{\text{align}}$	0.1002	0.1321	0.0942	0.1151
	$\bar{\mathcal{L}}_{\text{align}}$ & $\bar{\mathcal{L}}_{\text{uniform}}$	0.1231	0.1625	0.0689	0.0842
	$\bar{\mathcal{L}}_{\text{align}}$ & $\bar{H}_{\hat{\theta}}$	0.1388	0.1827	0.0585	0.0715
	$\bar{\mathcal{L}}_{\text{align}}$ & $\bar{\mathcal{L}}_{\text{uniform}}$ & $\bar{H}_{\hat{\theta}}$	0.1609	0.2118	0.0358	0.0437

summary lengths are constrained to be within 15% of the video lengths. Therefore, the majority part of the reference summary has exactly zeros scores. The contrastive refinement may still increase the scores' confidence for these regions, for which the annotators give zero scores thanks to the 15% constraint. This eventually decreases the average correlation with the reference summaries, as per Table 3.

However, suppose the predicted scores are refined to have sufficiently high confidence for regions with nonzero reference scores; in this case, they tend to be captured by the knapsack algorithm for computing the F1 scores. Therefore, we consider scores with both high F1 and high correlations to have high quality, as the former tends to neglect the overall correlations between the predicted and the annotated scores [39], and the latter focuses on their overall ranked correlations but cares less about the prediction confidence. This analysis may explain why the contrastive refinement for $\bar{\mathcal{L}}^*_{\text{align}}$ improves the F1 score but decreases the correlations. We analyze the negative effect of $\bar{\mathcal{L}}_{\text{uniform}}$ for SumMe later.

The effect of $\bar{\mathcal{L}}_{\text{align}}$. As can be observed in Tables 1, 2, and 3, solely using $\bar{\mathcal{L}}_{\text{align}}$ can already well quantify the frame importance. This indicates that $\bar{\mathcal{L}}_{\text{align}}$ successfully selects frames with diverse semantic information, which are indeed essential for a desirable summary. Moreover, we assume that diverse frames are a basis for a decent summary, thus always using $\bar{\mathcal{L}}_{\text{align}}$ for further ablations.

Table 2: Ablation results in terms of F1 together with their comparisons with previous unsupervised methods. The **boldfaced** results are the best ones. Please refer to Table 1's caption for the explanation of the notations and the text for analyses of the results.

	TVSum			SumMe		
	C	A	T	C	A	T
DR-DSN$_{60}$[62]	57.6	58.4	57.8	41.4	42.8	42.4
SUM-FCN$_{unsup}$[42]	52.7	-	-	41.5	-	39.5
SUM-GAN [35]	51.7	59.5	-	39.1	43.4	-
UnpairedVSN [41]	55.6	-	55.7	47.5	-	41.6
CSNet [21]	58.8	59	59.2	**51.3**	**52.1**	45.1
CSNet+GL+RPE [22]	59.1	-	-	50.2	-	-
SumGraph$_{unsup}$[40]	59.3	**61.2**	57.6	49.8	**52.1**	**47**
$\bar{\mathcal{L}}^*_{\text{align}}$	56.4	56.4	54.6	43.5	43.5	39.4
$\bar{\mathcal{L}}^*_{\text{align}}$ & $\bar{\mathcal{L}}^*_{\text{uniform}}$	58.4	58.4	56.8	47.2	46.07	41.7
$\bar{\mathcal{L}}_{\text{align}}$	54.6	55.1	53	46.8	47.1	41.5
$\bar{\mathcal{L}}_{\text{align}}$ & $\bar{\mathcal{L}}_{\text{uniform}}$	58.8	59.9	57.4	46.7	48.4	41.1
$\bar{\mathcal{L}}_{\text{align}}$ & $\bar{H}_{\hat{\theta}}$	53.8	56	54.3	45.2	45	45.3
$\bar{\mathcal{L}}_{\text{align}}$ & $\bar{\mathcal{L}}_{\text{uniform}}$ & $\bar{H}_{\hat{\theta}}$	**59.5**	59.9	**59.7**	46.8	45.5	43.9

The effect of $\bar{\mathcal{L}}_{\text{uniform}}$. $\bar{\mathcal{L}}_{\text{uniform}}$ measures how consistent a frame is with the context of the whole video, thus helping remove frames with diverse contents but hardly related to the video theme. It is clearly demonstrated in Tables 1 and 3 that incorporating $\bar{\mathcal{L}}_{\text{uniform}}$ helps improve the quality of the frame importance for TVSum. We thoroughly discuss why $\bar{\mathcal{L}}_{\text{uniform}}$ hurts SumMe performance in Section 7 of our supplementary.

The effect of the uniqueness filter $\bar{H}_{\hat{\theta}}$. As shown in Tables 1 and 2, though $\bar{H}_{\hat{\theta}}$ works well for TVsum videos, it hardly brings any benefits for the SumMe videos. Thus the good performance of the uniqueness filter for TVSum may simply stem from the fact that the background frames in TVSum are not challenging enough and can be easily detected by the uniqueness filter trained using only a few videos. Therefore, we suppose that $\bar{H}_{\hat{\theta}}$ needs to be trained on more videos in order to filter out more challenging background frames such that it can generalize to a wider range of videos. This is validated by the $\bar{\mathcal{L}}_{\text{align}}$&$\bar{H}_{\hat{\theta}}$ results in Table 3, which indicate both decent F1 scores and correlation coefficients for both TvSum and SumMe. The TVSum performance can be further boosted when $\bar{\mathcal{L}}_{\text{uniform}}$ is incorporated.

Comparison with DR-DSN [62] on Youtube8M. As per Table 1, DR-DSN is the only unsupervised method that has competitive performance with ours in terms of τ and ρ and that has released the official implementation. We thus also trained DR-DSN on our collected Youtube8M videos to compare it against our method. As shown in Table 3, DR-DSN has a hard time generalizing to the evaluation videos. We also compare DR-DSN to our method with varying

Table 3: The transfer evaluation setting with the Youtube8M dataset, where the training is solely conducted on the collected Youtube8M videos and then evaluated on TVSum and SumMe. The results from DR-DSN [62] is also provided for comparison.

	TVSum			SumMe		
	F	τ	ρ	F	τ	ρ
DR-DSN [62]	51.6	0.0594	0.0788	39.8	-0.0142	-0.0176
$\bar{\mathcal{L}}^*_{\text{align}}$	55.9	0.0595	0.0779	45.5	0.1000	0.1237
$\bar{\mathcal{L}}^*_{\text{align}}$ & $\bar{\mathcal{L}}^*_{\text{uniform}}$	56.7	0.0680	0.0899	42.9	0.0531	0.0649
$\bar{\mathcal{L}}_{\text{align}}$	56.2	0.0911	0.1196	46.6	0.0776	0.0960
$\bar{\mathcal{L}}_{\text{align}}$ & $\bar{\mathcal{L}}_{\text{uniform}}$	57.3	0.1130	0.1490	40.9	0.0153	0.0190
$\bar{\mathcal{L}}_{\text{align}}$ & $\bar{H}_{\hat{\theta}}$	58.1	0.1230	0.1612	48.7	0.0780	0.0964
$\bar{\mathcal{L}}_{\text{align}}$ & $\bar{\mathcal{L}}_{\text{uniform}}$ & $\bar{H}_{\hat{\theta}}$	59.4	0.1563	0.2048	43.2	0.0449	0.0553

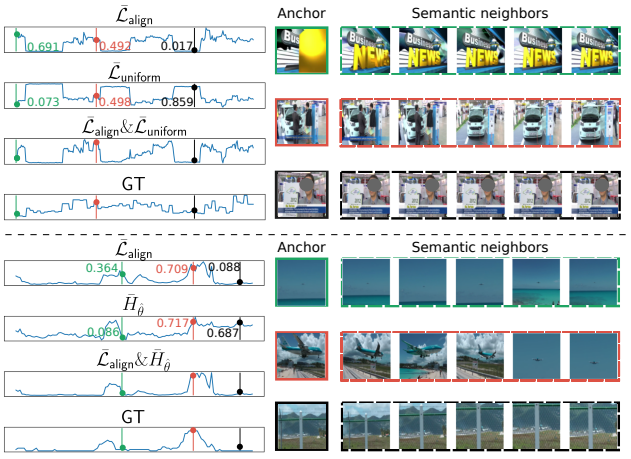

Figure 3: The qualitative analysis for two video examples from TVSum and SumMe. The left column contains importance scores, where "GT" stands for ground truth. The green bar selects an anchor frame with high $\bar{\mathcal{L}}_{\text{align}}$ but low $\bar{\mathcal{L}}_{\text{uniform}}$ or $\bar{H}_{\hat{\theta}}$, the red bar selects one with nontrial magnitude for both metrics, and the **black** bar selects one with low $\bar{\mathcal{L}}_{\text{align}}$ but high $\bar{\mathcal{L}}_{\text{uniform}}$ or $\bar{H}_{\hat{\theta}}$. We show five samples from the top 10 semantic nearest neighbors within the dashed boxes on the right for each selected anchor frame.

model sizes in Section 4 in the supplementary, where we compare the two methods' training efficiency and provide more training details as well.

More experiments. We perform the hyperparameter tuning on the challenging Youtube8M videos. We also evaluated the proposed metrics using different kinds of pre-trained features. Moreover, we observed that the F1 score for TVSum can be volatile to the importance scores' magnitude. The above results are in Sections 3-6 in the supplementary.

5.5. Qualitative Results

We show the effect of the local dissimilarity ($\bar{\mathcal{L}}_{\text{align}}$), the global consistency ($\bar{\mathcal{L}}_{\text{uniform}}$), and the uniqueness scores generated by the uniqueness filter $\bar{H}_{\hat{\theta}}$ in Fig. 3. We visualize and discuss the effects in pairs, *i.e.*, $\bar{\mathcal{L}}_{\text{align}}$&$\bar{\mathcal{L}}_{\text{uniform}}$ and $\bar{\mathcal{L}}_{\text{align}}$&$\bar{H}_{\hat{\theta}}$. We provide an example of how $\bar{H}_{\hat{\theta}}$ improves $\bar{\mathcal{L}}_{\text{align}}$&$\bar{\mathcal{L}}_{\text{uniform}}$ in Section 8 in the supplementary. In the upper half of Fig. 3, the green bar selects a frame with high local similarity but low global consistency, which turns out to be a title frame with a disparate appearance and hardly conveys any valuable information about the video. While the **black** bar selects a frame related to the main content of the video (an interview), it has semantic neighbors with almost the same look and is less likely to contain diverse semantics. The red bar selects a frame with moderate local dissimilarity and global consistency. The frame, together with its semantic neighbors, conveys diverse information *e.g.*, the car with or without people surrounding it. Moreover, it is also highly related to the whole video context: an interview in a car company.

For the lower half of Fig. 3, the green bar selects a frame with information noticeably different from its neighbors, *e.g.*, the sea occupies different proportions of the scene. However, such a frame can appear in any video with water scenes, rendering it not unique to the belonging video. Hence, its uniqueness score is low. The **black** bar selects a frame with an object specifically belonging to this video in the center, but the local semantic neighborhood around it hardly conveys very diverse information. The red bar selects a frame with both high local dissimilarity and high uniqueness, which turns out to be the frame related to the gist of the video: St Maarten landing.

6. Conclusion

We take the first attempt to directly quantify frame-level importance without bootstrapping online-generated summaries for unsupervised video summarization based on the contrastive losses proposed by [52]. With pre-trained deep features, our proposed metrics already yield high-quality importance scores compared to many previous heavily-trained methods. We can further improve the metrics with contrastive refinement by leveraging a sufficient amount of random videos. We also propose a novel uniqueness filter and validate its effectiveness through extensive experiments. It would be interesting to keep exploring the combination of unsupervised video summarization and representation learning.

7. Acknowledgement

This work was partly supported by JST CREST Grant No. JPMJCR20D3 and JST FOREST Grant No. JPMJFR216O.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[2] William H Beyer. *Standard Probability and Statistics: Tables and Formulae*. 1991.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *CVPR*, 2021.

[4] Luis Lebron Casas and Eugenia Koblents. Video summarization with LSTM and deep attention models. In *MMM*, 2019.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[6] Yiyan Chen, Li Tao, Xueting Wang, and Toshihiko Yamasaki. Weakly supervised video summarization by hierarchical reinforcement learning. In *ACM MM Asia*. 2019.

[7] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

[8] Wei-Ta Chu and Yu-Hsin Liu. Spatiotemporal modeling and label distribution learning for video summarization. In *MMSP*, 2019.

[9] Sandra Eliza Fontes De Avila, Ana Paula Brandao Lopes, Antonio da Luz Jr, and Arnaldo de Albuquerque Araújo. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1):56–68, 2011.

[10] Mohamed Elfeki and Ali Borji. Video summarization via actionness ranking. In *WACV*, 2019.

[11] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention. In *ACCV*, 2018.

[12] Litong Feng, Ziyin Li, Zhanghui Kuang, and Wei Zhang. Extractive video summarizer with memory augmented neural networks. In *ACM MM*, 2018.

[13] Tsu-Jui Fu, Shao-Heng Tai, and Hwann-Tzong Chen. Attentive and adversarial learning for video summarization. In *WACV*, 2019.

[14] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *ECCV*, 2014.

[15] Michael Gygli, Helmut Grabner, and Luc Van Gool. Video summarization by learning submodular mixtures of objectives. In *CVPR*, 2015.

[16] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*, 2022.

[17] Xufeng He, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan. Unsupervised video summarization with attentive conditional generative adversarial networks. In *ACM MM*, 2019.

[18] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[19] Zhong Ji, Fang Jiao, Yanwei Pang, and Ling Shao. Deep attentive and semantic preserving video summarization. *Neurocomputing*, 405:200–207, 2020.

[20] Zhong Ji, Kailin Xiong, Yanwei Pang, and Xuelong Li. Video summarization with attention-based encoder–decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[21] Yunjae Jung, Donghyeon Cho, Dahun Kim, Sanghyun Woo, and In So Kweon. Discriminative feature learning for unsupervised video summarization. In *AAAI*, 2019.

[22] Yunjae Jung, Donghyeon Cho, Sanghyun Woo, and In So Kweon. Global-and-local relative position embedding for unsupervised video summarization. In *ECCV*, 2020.

[23] Maurice G Kendall. The treatment of ties in ranking problems. *Biometrika*, 1945.

[24] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[26] Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. Background suppression network for weakly-supervised temporal action localization. In *AAAI*, 2020.

[27] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. Weakly-supervised temporal action localization by uncertainty modeling. In *AAAI*, 2021.

[28] Zutong Li and Lei Yang. Weakly supervised deep reinforcement learning for video summarization with semantically meaningful reward. In *WACV*, 2021.

[29] Jingxu Lin and Sheng-hua Zhong. Bi-directional self-attention with relative positional encoding for video summarization. In *ICTAI*, 2020.

[30] David Liu, Gang Hua, and Tsuhan Chen. A hierarchical visual model for video object summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(12):2178–2190, 2010.

[31] Daochang Liu, Tingting Jiang, and Yizhou Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In *CVPR*, 2019.

[32] Yen-Ting Liu, Yu-Jhe Li, and Yu-Chiang Frank Wang. Transforming multi-concept attention into video summarization. In *ACCV*, 2020.

[33] Yen-Ting Liu, Yu-Jhe Li, Fu-En Yang, Shang-Fu Chen, and Yu-Chiang Frank Wang. Learning hierarchical self-attention for video summarization. In *ICIP*, 2019.

[34] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[35] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial LSTM networks. In *CVPR*, 2017.

[36] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. In *NeurIPS*, 2021.

[37] Phuc Xuan Nguyen, Deva Ramanan, and Charless C Fowlkes. Weakly-supervised action localization with background modeling. In *ICCV*, 2019.

[38] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[39] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkila. Rethinking the evaluation of video summaries. In *CVPR*, 2019.

[40] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. SumGraph: Video summarization via recursive graph modeling. In *ECCV*, 2020.

[41] Mrigank Rochan and Yang Wang. Video summarization by learning from unpaired data. In *CVPR*, 2019.

[42] Mrigank Rochan, Linwei Ye, and Yang Wang. Video summarization using fully convolutional sequence networks. In *ECCV*, 2018.

[43] Yassir Saquil, Da Chen, Yuan He, Chuan Li, and Yong-Liang Yang. Multiple pairwise ranking networks for personalized video summarization. In *ICCV*, 2021.

[44] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. TVSum: Summarizing web videos using titles. In *CVPR*, 2015.

[45] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[46] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[47] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020.

[48] Nikolai Ufer and Bjorn Ommer. Deep semantic feature matching. In *CVPR*, 2017.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[50] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *CVPR*, 2021.

[51] Junbo Wang, Wei Wang, Zhiyong Wang, Liang Wang, Dagan Feng, and Tieniu Tan. Stacked memory network for video summarization. In *ACM MM*, 2019.

[52] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020.

[53] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, 2021.

[54] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

[55] Yuan Yuan, Haopeng Li, and Qi Wang. Spatiotemporal modeling for video summarization using convolutional recurrent neural network. *IEEE Access*, 2019.

[56] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *ECCV, 2016*.

[57] Ke Zhang, Kristen Grauman, and Fei Sha. Retrospective encoders for video summarization. In *ECCV*, 2018.

[58] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. Hierarchical recurrent neural network for video summarization. In *ACM MM*, 2017.

[59] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. HSA-RNN: Hierarchical structure-adaptive RNN for video summarization. In *CVPR*, 2018.

[60] Bin Zhao and Eric P Xing. Quasi real-time summarization for consumer videos. In *CVPR*, 2014.

[61] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

[62] Kaiyang Zhou, Yu Qiao, and Tao Xiang. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *AAAI*, 2018.

[63] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *ICCV*, 2019.