ABSTRACT

Video summarization aims at generating concise video summaries from the lengthy videos, to achieve better user watching experience. Due to the subjectivity, purely supervised methods for video summarization may bring the inherent errors from the annotations. To solve the subjectivity problem, we study the general user summarization process. General users usually watch the whole video, compare interesting clips and select some clips to form a final summary. Inspired by the general user behaviours, we formulate the summarization process as multiple sequential decision-making processes, and propose Comparison-Selection Network (CoSNet) based on multi-agent reinforcement learning. Each agent focuses on a video clip and constantly changes its focus during the iterations, and the final focus clips of all agents form the summary. The comparison network provides the agent with the visual feature from clips and the chronological feature from the past round, while the selection network of the agent makes decisions on the change of its focus clip. The specially designed unsupervised reward and supervised reward together contribute to the policy advancement, each containing local and global parts. Extensive experiments on two benchmark datasets show that CoSNet outperforms state-of-the-art unsupervised methods with the unsupervised reward and surpasses most supervised methods with the complete reward.

CCS CONCEPTS

• Computing methodologies → Video summarization; Multi-agent reinforcement learning.

KEYWORDS

video summarization, multi-agent reinforcement learning

1 INTRODUCTION

Gigantic amounts of videos are produced by mobile phones, wearable devices and surveillance cameras. The lengthy raw videos with sparse information make it difficult for viewing, browsing and retrieving, resulting in the decline of user experience. In the meantime, video summaries can shorten the viewing time, provide dense information and save the storage space. To alleviate the problems of raw videos, we need video summarization to transform the lengthy raw videos into concise video summaries.

Video summarization is a relatively subjective task. In the process of creating datasets, different annotators may produce largely different annotations for the same video. Therefore, the annotation of video summarization datasets requires more annotators than other tasks, to ensure the maximum annotation accuracy. In the analysis of the widely used benchmark datasets, SumMe [10] and TVSum [46], the former dataset has 15 to 18 annotations for each video, while the latter dataset has 20 annotations for each video. However, the annotations may still suffer from subjectivity due to the irreconcilable difference among different annotations.

In order to solve the subjectivity problem, we specially conduct a survey on the methods that can relieve the problem. Unsupervised methods [12, 15, 27, 59, 62] can work without annotations, but may lose some effective information in annotations. Among the various methods proposed in the literature, some methods originate from the inherent characteristics of the videos, and some other methods are from the inspiration of video user behaviours. In [46], Song et al. proposed a method that selects frames which are most relevant to the video titles to form the summaries. Like video titles, Panda et al. [30] used video-level category annotations. The category annotations contain less information than frame-level annotations, but are actually more "accurate", because the opinions about the categories are consistent among almost all annotators. In [38], Rochan and Wang used the idea that unpaired videos and summaries can reduce the subjectivity caused by the interdependence of paired ones. In [54], Xiong et al. applied the thought of "less is more" to the video summarization task, which indicates that shorter videos are more informative than longer ones. The above-mentioned methods...
based on inherent characteristics or user behaviours can largely relieve the subjectivity problem.

We are also inspired by the behaviours of video users. How do general users summarize videos? They usually watch the whole video, compare the interesting clips (p1 percent of the whole video length) and select the final clips to form the summary (p2 percent, p2 < p1 < 1.0). We formulate this user summarization process as multiple sequential decision-making processes, and propose Comparison-Selection Network (CoSNet) that is composed of comparison network and selection network based on multi-agent reinforcement learning (MARL).

The N agents of CoSNet "watch" SN clips at the same time and change their focus clips constantly until all agents select the "stay-still" action or the iteration reaches the maximum, as shown in Fig. 1. Each agent compares five clips (the focus clip, two neighboring clips and two focus clips of the neighboring agents) with the comparison network and selects the next clip to focus on with the selection network (by moving left, staying still or moving right). The comparison network is composed of 3D convolutional layers and long short-term memory (LSTM) layers, while the selection network contains fully connected (FC) layers. We specially design unsupervised reward and supervised reward, each with local and global parts. Unsupervised reward can avoid the subjectivity of the annotations, while supervised reward can bring some promotion from the annotation information. The local rewards act as the feedback for the local comparisons, while the global rewards are the feedback for the value of the global selection. In this way, CoSNet can follow the process that how general users summarize videos and relieve the subjectivity to the utmost extent.

To sum up, the major contributions of this paper are:

- We follow the behaviours that general users usually summarize videos by comparing and selecting video clips, and formulate the user summarization process as multiple sequential decision-making processes.
- CoSNet is proposed based on MARL. It is composed of comparison network and selection network, with specially designed unsupervised reward and supervised reward (each with local and global parts).
- Extensive experiments on benchmark datasets show that CoSNet outperforms state-of-the-art unsupervised methods with the unsupervised reward and surpasses most supervised methods with the complete reward.

2 RELATED WORK

2.1 Video Summarization

Some researchers formulate video summarization as an optimization problem [5, 11, 28, 31, 55, 60]. With the overwhelming trend of deep learning (DL), several kinds of DL based methods have also been applied to video summarization. Due to the temporal attributes of videos, some methods are based on different varieties of recurrent neural network (RNN) [7, 44, 49, 61, 64, 65], including LSTM [61], hierarchical RNN [64, 65] and others. In [39], Rochan et al. formulated video summarization as a sequence labeling problem and used convolutional neural network (CNN) to solve it. Li et al. [25] and Sharghi et al. [41] proposed methods based on determinantal point processes (DPP). Methods based on unsupervised learning [12, 15, 27, 59, 62], like generative adversarial networks (GAN) and variational autoencoders (VAE), try to make the summary features indistinguishable from the raw video features. Weakly supervised methods proposed by Cai et al. [1] and Panda et al. [30] are effective in that some videos have additional web information which can be used for summarization.

There are also some varieties of the video summarization task. The first variety is query-based video summarization [2, 8, 42, 43, 48, 53, 63] that the summaries are generated according to user queries. The second variety is interactive video summarization [4, 14] that the computers interact with users during the summary generation processes. The third variety is 360-degree video summarization [21, 58] that the 360-degree videos are summarized both temporally and spatially. The fourth variety is first-person video summarization [8, 13, 33, 45, 55, 56] that the characteristics of the first-person videos are considered during the summarization process. In this paper, we focus on the general video summarization task.

2.2 Reinforcement Learning

The goal of reinforcement learning (RL) is to learn a good policy for the agent from experimental trials by maximizing expected future rewards. RL has made it to solve tasks in many research areas recently, such as games [19, 29, 57] and robotics [9, 22, 26]. It has succeeded in solving various vision tasks, like visual tracking [37], video face recognition [35], image cropping [23], video captioning [32], activity localization [50] and video classification [6]. MARL also helps to solve some vision tasks. Rosello and Kohenderfer [40] proposed a method based on MARL for multi-object tracking. Wu et al. [52] proposed a frame sampling method based on MARL for video recognition.

As the fast development, RL has also been applied to the video summarization task. In [66], Zhou et al. proposed DSN with specially designed diversity and representativeness rewards. The rewards of DSN can describe how diverse and representative the generated summary is, but may ignore some local information. In [20], Lan et al. proposed FFNet for video fast-forwarding. FFNet is fast in processing speed, but many clips are omitted for "watching" that may bring some feature information loss. In [36], Rathore et al. mainly focused on long egocentric video summarization. With MARL, our proposed CoSNet can simultaneously watch many clips to reduce feature information loss, and use both local and global rewards to reduce local information loss.

3 METHOD

We formulate video summarization as multiple sequential decision-making processes and propose CoSNet based on MARL. CoSNet contains N agents. Each agent is composed of a comparison network and a selection network, with unsupervised reward and supervised reward (both with local and global parts). The agents are identical in network architecture and share the same parameters, for convenient experiments with different numbers of agents. Fig. 2 is a demonstration of CoSNet.

3.1 Problem Formulation

The sequential decision-making process has four elements, i.e., state, action, policy and reward.
Figure 2: Each agent is composed of a comparison network and a selection network. Firstly, the C3D layers transform the raw clips into visual features. Then, the focus and neighboring features are averaged, as the input for the LSTM layers. Finally, the FC layers make the selection decisions based on the hidden states of the LSTM layers. The agents move left, stay still or move right according to the decisions. Unsupervised reward and supervised reward (each with local and global parts) are calculated during each round in the policy-based reinforcement learning processes.

State. The states $s \in S$ include the positions of the $N$ agents’ focus clips, the visual features of clips and the chronological features. During each round, the agents have an temporal order and neighboring relations. We number the agents $\{a_i\}_{i=1}^N$ from left to right temporally by 1 to $N$. The $M$ clips also have an temporal order and neighboring relations. We number the clips $\{c_j\}_{j=1}^M$ from left to right temporally by 1 to $M$.

Action. The actions $u \in U$ are discrete movements of the agents. An agent can move left, stay still or move right. The movement steps are within the scope of $u \in [-l, l]$. To avoid collisions, the agents execute the movement actions one by one. During the movement process, if an agent $a_i$ moves to a position that already has another agent $a_j$, $a_i$ moves further in the same direction until no collision exists. In particular, the video clips form a circle in our setting, which means the first clip $c_1$ is ‘adjacent’ to the last clip $c_M$. Under this setting, an agent may move beyond $c_1$ and appear in one of the tail clips during left movements.

Policy. The policy $\pi$ denotes the possibility of choosing action $u_t$ under state $s_t$, as Eq. (1) shows.

$$\pi(u_t|s_t) = P[ u = u_t | s = s_t ]$$ (1)

Reward. The reward $r$ of each round acts as important feedback for the advancement of the policy. The detailed reward definitions including unsupervised reward, supervised reward, local parts and global parts are presented in Sec. 3.3.

3.2 Network Architecture

Comparison Network. The comparison network is composed of 3D convolutional layers for visual feature representation and LSTM layers for chronological feature reservation. In practice, the 3D convolutional layers are C3D [47] layers pre-trained on Sports-1M dataset [16]. We use the “fc6” layer features $\{x_i\}_{i=1}^M$ of video clips (16 frames per clip).

Each agent $a_i$ focuses on a clip $c_j$ at round $t$. The input for the LSTM layers of $a_i$ is the average of the features from five clips (the focus clip of the left adjacent agent, the left adjacent clip, the focus clip, the right adjacent clip and the focus clip of the right adjacent agent). Then the LSTM layers produce hidden states $\{h_i\}_{i=1}^N$. For the first agent $a_1$, the left adjacent agent is the last agent $a_N$ under the circle setting. If agent $a_1$ focuses on the first clip $c_1$, the left adjacent clip of $a_1$ is the last clip $c_M$.

Selection Network. The selection network contains two FC layers, to make decisions on which action to choose. The FC layers take the hidden states of the LSTM layers as input, output the action values for each possible action $u \in U$ and produce action choices $\{u_i\}_{i=1}^N$ with softmax.

3.3 Reward Definition

The reward definitions include unsupervised reward and supervised reward, each with both local and global parts.
Unsupervised Reward. Unsupervised reward contains local unsupervised reward and global unsupervised reward, and does not need annotations for calculation.

Local unsupervised reward \( r^l_u \) (Eq. (2)) denotes the local centrality of the focus clips. We want to improve the minimum feature similarity between the focus clip and its neighboring clips, to ensure that the focus clip approximates all its neighboring clips. The range of neighboring clips is between the focus clip of the left adjacent agent and the focus clip of the right adjacent agent. In Eq. (2), \( c_{a1} \) denotes the focus clip of agent \( a_1 \), \( c' \) denotes a neighboring clip, \( x(\cdot) \) denotes the C3D feature of the clip, and \( N(a_1) \) denotes the set of neighboring clips of agent \( a_1 \).

\[
r^l_u = \min_{c' \in N(a_1)} \left( \frac{x(c_{a1})^T x(c')}{{\|x(c_{a1})\|}_2 \|x(c')\|_2} \right) \tag{2}
\]

Global unsupervised reward \( r^g_u \) (Eq. (3)) denotes the overall feature difference among all focus clips, to ensure that the generated summaries cover most contents of the videos.

\[
r^g_u = 1 - \frac{1}{N(N-1)} \sum_{i \in [1,N]} \sum_{j \in [1,N]} \frac{x(c_{a1})^T x(c_j)}{{\|x(c_{a1})\|}_2 \|x(c_j)\|_2} \tag{3}
\]

The unsupervised reward \( r^l_u \) (0 \( \leq r^l_u \leq 1 \)) is the combination of local unsupervised reward \( r^l_u \) (0 \( \leq r^l_u \leq 1 \)) and global unsupervised reward \( r^g_u \) (0 \( \leq r^g_u \leq 1 \)), as Eq. (4) shows. In Eq. (4), \( \alpha_1 \) is a scale factor (0 \( < \alpha_1 \leq 1 \)).

\[
r^u = \frac{1}{2} (r^l + \alpha_1 r^g) \tag{4}
\]

Supervised Reward. Supervised reward contains local supervised reward and global supervised reward, and needs annotations for calculation.

Local supervised reward \( r^s_l \) (Eq. (5)) and skip reward (0 \( \leq r^s_l \leq 1 \)) in Eq. (5). Change reward denotes the precision change between the new focus clip and the old focus clip, to ensure that the agents move towards more important clips with more key frames. When agents move, some frames are skipped over. Skip reward denotes the proportion of key frames in the frames skipped over, to ensure that the agents skip over less key frames. In Eq. (5), \( f^k \) and \( f^c \) are the number of frames in a clip at round \( t \), \( f^u \) is the number of frames in a clip. \( f^s \) is the number of frames skipped over between \( t - 1 \) and \( t \), \( f^s \) is the number of frames skipped over between \( t - 1 \) and \( t \), \( \alpha_2 \) is a scale factor (0 \( < \alpha_2 \leq 1 \)).

\[
r^s_l = \frac{1}{4} (1 + \frac{f^k - f^u}{f^c} - \frac{f^k - f^u}{f^c}) + \frac{1}{2} \alpha_2 (1 - \frac{f^s}{f^s}) \tag{5}
\]

Global supervised reward \( r^s_g \) (Eq. (6)) denotes the global precision change between the new round and the old round, which also ensures that the agents move towards more important clips.

\[
r^s_g = \frac{1}{2} (1 + \frac{\sum_{agents} f^k}{\sum_{agents} f^c} - \frac{\sum_{agents} f^k - f^s}{\sum_{agents} f^c}) \tag{6}
\]

The supervised reward \( r^s \) (0 \( \leq r^s \leq 1 \)) is the combination of local supervised reward \( r^s_l \) (0 \( \leq r^s_l \leq 1 \)) and global supervised reward \( r^s_g \) (0 \( \leq r^s_g \leq 1 \)), as Eq. (7) shows. In Eq. (7), \( \alpha_3 \) is a scale factor (0 \( < \alpha_3 \leq 1 \)).

\[
r^s = \frac{1}{2} (r^s_l + \alpha_3 r^s_g) \tag{7}
\]

We can further combine the unsupervised reward and the supervised reward. The reward of agent \( a \) in round \( t \) is shown in Eq. (8), in which \( \alpha_4 \) is a scale factor (0 \( < \alpha_4 \leq 1 \)). \( r^a \) also has the range of 0 \( \leq r^a \leq 1 \).

\[
r^a = \frac{1}{2} (r^u + \alpha_4 r^s) \tag{8}
\]

The accumulated reward of agent \( a \) from round \( t \) on is in the form of Eq. (9), \( \gamma (0 < \gamma \leq 1) \) is the discount factor.

\[
R^a_t = \sum_{t=0}^{T} \gamma^{t-t} r^a_t \tag{9}
\]

3.4 Policy Gradient

The objective function is defined as Eq. (10). Our optimization goal is to learn a policy \( \pi \) with parameters \( \theta^\pi \) by maximizing the objective function \( J(\theta^\pi) \).

\[
J(\theta^\pi) = \sum_{a=1}^{N} \sum_{t=0}^{T} \pi(u^a_t | x^a_t ; \theta^\pi) R^a_t \tag{10}
\]

The parameters \( \theta^\pi = (\theta_c, \theta) \) need to be learned for optimization. \( \theta_c \) represents the parameters of the LSTM layers, and \( \theta \) represents the parameters of the FC layers, while the pre-trained parameters of C3D layers are frozen. The gradient of \( J(\theta^\pi) \) is shown in Eq. (11).

\[
\nabla_{\theta^\pi} J(\theta^\pi) = \sum_{a=1}^{N} \sum_{t=0}^{T} \pi(u^a_t | x^a_t ; \theta^\pi) \nabla_{\theta^\pi} \log \pi(u^a_t | x^a_t ; \theta^\pi) R^a_t \tag{11}
\]

The optimization is non-trivial due to the high-dimensional action sequences. Therefore, We follow REINFORCE [51] algorithm to optimize the parameters. We approximate the policy gradient by K action sequences, as shown in Eq. (12).

\[
\nabla_{\theta^\pi} J(\theta^\pi) \approx \frac{1}{K} \sum_{k=1}^{K} \sum_{a=1}^{N} \sum_{t=0}^{T} \nabla_{\theta^\pi} \log \pi(u^a_t | x^a_t ; \theta^\pi) R^a_t \tag{12}
\]

The loss function has two items, the policy gradient and a \( l_2 \) normalization term. We update the parameters \( \theta^\pi \) by Eq. (13). \( \eta \) is the learning rate, and \( \lambda (0 < \lambda \leq 1) \) is a scale factor.

\[
\theta^\pi = \theta^\pi - \eta \nabla_{\theta^\pi} J(\theta^\pi) + \lambda \sum_{\omega \in \theta^\pi} \omega^2 \tag{13}
\]

4 EXPERIMENTS

4.1 Experimental Setup

Datasets. We conduct experiments on two widely used benchmark datasets, SumMe and TVSum. SumMe contains 25 videos of various topics, like sports and sights. The duration of the videos in
Compare and Select: Video Summarization with Multi-Agent Reinforcement Learning

(a) Sample frames of video-17 (Reuben Sandwich with Corned Beef & Sauerkraut) in TVSum.

(b) CoSNet-US.

(c) CoSNet-U.

(d) CoSNet-S.

(e) CoSNet-LU.

Figure 3: (a) is the sample frames of video-17 in TVSum. The girl is cooking a special beef sandwich. The heights of all the bars in (b) to (e) represent the ground truth importance scores of video-17. The colored bars represent the selected segments with the corresponding methods.

SumMe varies from 1 to 6 minutes. TVSum contains 50 videos of 10 kinds. The duration of the videos in TVSum is between 2 and 10 minutes. The videos in both datasets are annotated by frame-level importance scores. The duration of a ground truth summary is around or below 15 percent of the raw video duration.

We also use the OVP [61] dataset (50 videos) and the YouTube [3] dataset (39 videos) for data augmentation and transfer learning. In order to unify the different annotation forms of different datasets, we follow [61] to make conversions among key-frames, key-shots and frame-level importance scores. During the above conversions processes, We use KTS [34] to temporally divide the raw videos into segments.

**Evaluation Metric.** Similar to other methods, we take the F-score based on key shots used in [61] as the evaluation metric, to ensure fair comparisons. In the form of the harmonic mean of precision $P$ and recall $R$, F-score (Eq. (14)) depicts the similarity between the ground truth summaries and the generated summaries.

$$F = \frac{2PR}{P + R} \times 100\% \quad (14)$$

The experiments are conducted under the three kinds of settings in [61]. Canonical (C) setting represents the standard 5-fold cross validation (5FCV) metric. Augmented (A) setting denotes augmenting the training set with other three datasets under 5FCV metric. Transfer (T) setting means only using other three datasets as the training set.

**Implementation Details.** The 3D convolutional layers are pre-trained C3D layers as mentioned in Sec. 3.2. The initial positions of the $N$ agents are random and different among epochs. The $N$ agents are the same in network architecture and parameters during a round and the parameters update uniformly according to the experiences of all agents.

$M$ is not same for different videos according to the number of total frames $f_\text{total}$. $f_\text{clip}$ is the number of frames in a clip (16 by default). Empirically, the number of frames focused on by agents is around 15 percent of $f_\text{total}$, with the relationship of Eq. (15). Therefore, $N$ is also not same for different videos. In contrast experiments, $f_\text{clip}$ is the integer multiple of 16, and $N$ changes accordingly. For TVSum, if we take $f_\text{clip}$ to be 16, then $f_\text{total}$ ranges from 2500 to 19406, $M$ ranges from 157 to 1213, and $N$ ranges from 25 to 182. For SumMe, if we take $f_\text{clip}$ to be 16, then $f_\text{total}$ ranges from 950 to 9721, $M$ ranges from 60 to 608, and $N$ ranges from 9 to 92.

$$N f_\text{clip} \leq 15\% f_\text{total} \quad (15)$$

Unlike other MARL-based methods, our $N$ is dynamic due to the characteristics of the video summarization task which requires

| Method    | SumMe | TVSum |
|-----------|-------|-------|
| CoSNet-GU | 43.5  | 56.9  |
| CoSNet-GS | 44.6  | 57.3  |
| CoSNet-LU | 45.9  | 58.4  |
| CoSNet-LS | 46.3  | 58.7  |
| CoSNet-U  | 46.4  | 59.3  |
| CoSNet-S  | 47.1  | 59.4  |
| CoSNet-US | 47.8  | 59.7  |
Figure 4: The visualization of the generated summaries for four videos in TVSum. (a) video-18: "Poor Man’s Meals: Spicy Sausage Sandwich". (b) video-22: "Charlotte Parkour". (c) video-28: "LA Kings Stanley Cup South Bay Parade". (d) video-46: "German Shepherd Dog Show".

Table 2: The evaluation results of CoSNet-US (%) with different clip sizes on SumMe and TVSum.

| Clip Size (frames) | SumMe | TVSum |
|-------------------|-------|-------|
| 16                | 47.8  | 59.7  |
| 32 (16×2)         | 46.1  | 58.5  |
| 48 (16×3)         | 45.3  | 57.4  |

Table 3: The evaluation results (%) of CoSNet-U and other unsupervised methods on SumMe and TVSum.

| Method                   | SumMe | TVSum |
|--------------------------|-------|-------|
| Video-MMR [24]           | 26.6  | -     |
| Vsumm [3]                | 33.7  | -     |
| Web image [17]           | -     | 36.0  |
| Co-archetypal [46]       | -     | 50.0  |
| SUM-GANdpp [27]          | 39.1  | 51.7  |
| DR-DSN [66]              | 41.4  | 57.6  |
| SUM-FCNunsup [39]        | 41.5  | 52.7  |
| Xiong et al. [54]        | -     | 56.3  |
| Rathore et al. [36]      | 45.6  | 59.1  |
| ACGAN [12]               | 46.0  | 58.5  |
| CoSNet-U                 | 46.4  | 59.3  |

We experiment on seven reward combinations, complete reward (CoSNet-US), only supervised reward (CoSNet-S), only unsupervised reward (CoSNet-LU), only global unsupervised reward (CoSNet-GU), only local supervised reward (CoSNet-LS) and only global supervised reward (CoSNet-GS). By setting up the seven different reward combinations, we can draw the difference between unsupervised reward and supervised reward, and the difference between local parts and global parts. Fig. 3 visualizes the generated summaries of video-17 in TVSum with four different reward combinations.

Comparing CoSNet-S with CoSNet-U, we can find that supervised reward performs slightly better than unsupervised reward.
Table 4: The evaluation results (%) of CoSNet-US and other supervised methods on SumMe and TVSum.

| Method            | SumMe | TVSum |
|-------------------|-------|-------|
| Interestingness [10] | 39.4  | -     |
| Submodularity [11]   | 39.7  | -     |
| Summary transfer [60] | 40.9  | -     |
| Bi-LSTM [61]        | 37.6  | 54.2  |
| DPP-LSTM [61]       | 38.6  | 54.7  |
| SUM-GAN$_{sup}$ [27] | 41.7  | 56.3  |
| DR-DSN$_{sup}$ [66] | 42.1  | 58.1  |
| MAVS [7]           | 43.1  | 67.5  |
| DySeqDPP [25]       | 44.3  | 58.4  |
| ACGAN$_{sup}$ [12]  | 47.2  | 59.4  |
| SUM-FCN [39]        | 47.5  | 56.8  |
| SUM-DeepLab [39]    | 48.8  | 58.4  |
| SMN [49]           | 58.3  | 64.5  |
| CoSNet-US          | 47.8  | 59.7  |

Table 5: The evaluation results (%) of CoSNet-U and other unsupervised methods on SumMe and TVSum under Augmented and Transfer settings.

| Method            | SumMe | TVSum |
|-------------------|-------|-------|
| SUM-GAN$_{aug}$ [27] | 43.4  | 59.5  |
| DR-DSN [66]        | 42.8  | 58.4  | 57.8 |
| ACGAN [12]         | 47.0  | 44.5  | 58.9 |
| CoSNet-U           | 47.1  | 45.2  | 59.8 |

Table 6: The evaluation results (%) of CoSNet-U on YouTube, OVP, SumMe and TVSum under one-to-one Transfer setting.

| Test | Training | YouTube | OVP | SumMe | TVSum |
|------|----------|---------|-----|-------|-------|
|      |          |         |     |       |       |
| SumMe| 42.8     | 43.5    | -   | 44.1  |       |
| TVSum| 56.5     | 57.4    | 56.6| -     | -     |

(47.1% vs. 46.4% on SumMe and 59.4% vs. 59.3% on TVSum), which is the usual case in the contrast between unsupervised methods and supervised methods. Meanwhile, the difference in our methods is smaller than that of other methods (only 0.7% on SumMe and 0.1% on TVSum). The small difference justifies that the user summarization process modeling has less relationship with the subjective annotations than other methods.

We can also find that CoSNet-GS achieves higher F-score than CoSNet-GU (46.3% vs. 44.6% on SumMe and 58.7% vs. 57.3% on TVSum) and CoSNet-LU achieves higher F-score than CoSNet-GU (45.9% vs. 43.5% on SumMe and 58.4% vs. 56.9% on TVSum). The reason why methods with only local parts perform better than methods with only global parts is that the local parts in the rewards reflect local agent-level gains, which can benefit more from the multi-agent setting of CoSNet than the global parts.

Then, we can analyze this result further by comparing what the different reward parts have done respectively during the learning process. Specifically, for unsupervised reward, the local clip centrality ensures that each focus clip approximates and covers its neighboring clips, while the global feature difference only makes the summary diverse. Similarly, for supervised reward, the local change and local clip skip help to save more key frames at clip level, while the global change only helps to save key frames at video level. Both global parts of the rewards can not benefit from the many agents which is the key characteristic of CoSNet.

**Clip Sizes.** We also conduct comparative experiments on different clip sizes. In the Implementation Details part of Sec. 4.1, we introduce the relationship of \( N, f^{clip} \) and \( f^{total} \) with Eq. (15). The default clip size \( f^{clip} = 16 \) frames. For comparison, we further expand \( f^{clip} \) to 32 frames (\( 2 \times 16 \)) and 48 frames (\( 3 \times 16 \)), and \( N \) changes accordingly. The \( C3D \) features are extracted every 16 frames. Therefore, the input feature vector for larger \( f^{clip} \) is the average of two or three \( C3D \) feature vectors.

Tab. 2 shows the evaluation results of CoSNet-US with different clip sizes on SumMe and TVSum. The F-score decreases with the increase of clip size, which indicates that the clip size of 16 is moderate for the video summarization task. Smaller clips can be more accurate in granularity than larger clips when making comparisons and selections.

### 4.3 Method Comparison

We compare CoSNet with other methods. For comparison with unsupervised methods, we use CoSNet-U with only unsupervised reward. For comparison with supervised methods, we use CoSNet-US with the complete reward. We also make comparisons under Augmented and Transfer settings. To ensure fair comparison, the numerical results are mostly from [12] and partly from other state-of-the-art methods not included in [12].

**Comparison with Unsupervised Methods.** Tab. 3 shows the evaluation results of CoSNet-U and other unsupervised methods on SumMe and TVSum. CoSNet-U outperforms state-of-the-art unsupervised methods on both datasets. Specifically, CoSNet-U performs 0.4% better than ACGAN on SumMe and 0.2% better than the method by Rathore et al. [36] on TVSum. The performance with unsupervised reward demonstrates that CoSNet-U solves the subjectivity problem to some extent.

**Comparison with Supervised Methods.** Tab. 4 shows the evaluation results of CoSNet-US and other supervised methods on SumMe and TVSum. CoSNet-US outperforms most supervised methods on both datasets. Meanwhile, the F-score of CoSNet-US is lower than SUM-DeepLab (SumMe), MAVS (TVSum) and SMN (both). Besides, the performance of MAVS on SumMe is inferior to CoSNet-US, and the performance of SUM-DeepLab on TVSum is also inferior to CoSNet-US. The reason is that the supervised reward brings some promotion with the annotation information, but also brings the inherent subjectivity in the annotations.
Figure 5: The contrast between the ground truth importance score distributions of the given summaries and the binary score distributions of the generated summaries for video-2, video-36, video-39 and video-41 in TVSum.

Comparison under Augmented and Transfer Settings. Tab. 5 shows the evaluation results of CoSNet-U and other unsupervised methods on SumMe and TVSum under Augmented and Transfer settings. CoSNet-U outperforms other unsupervised methods on both datasets under both settings.

Tab. 6 shows the evaluation results of CoSNet-U on YouTube, OVP, SumMe and TVSum under one-to-one Transfer setting. The training sets include all the four datasets, while the test sets only include SumMe and TVSum. With only one dataset for training, CoSNet-U performs worse than with three datasets for training. The reason is that the number of videos in the training set with three datasets is larger than with only one dataset, which also applies to the difference among training sets with only one dataset (TVSum-50 = OVP-50 > YouTube-39 > SumMe-25).

4.4 Qualitative Analysis
We analyze CoSNet qualitatively with visual illustrations in this subsection. Fig. 3 and Fig. 4 are the visualization for the generated summaries of five videos in TVSum. The heights of all the bars (grey or colored) represent the ground truth importance scores of the videos, while the colored bars represent the selected segments. We can find that most bars with relatively higher ground truth importance scores are selected by different varieties of CoSNet.

In Fig. 3, there is a comparison among four different methods with different reward combinations on video-17 in TVSum. The four methods (CoSNet-US, CoSNet-U, CoSNet-S and CoSNet-LU) perform likely with slight difference in the selection of segments.

Fig. 5 shows the contrast between the ground truth importance score distributions of the given summaries and the binary score distributions of the generated summaries. With CoSNet, the summarization results are in the form of key clips, which can not be directly transformed into continuous scores ([0, 1]). Therefore, we make comparisons between two different forms of scores, which can still demonstrate the quality of the generated summaries intuitively. We can find that the high scores of the given ground truth summaries and the “one” scores of the generated summaries appear in similar segments.

5 CONCLUSION
In this paper, we propose Comparison-Selection Network (CoSNet) based on multi-agent reinforcement learning for video summarization. In order to solve the subjectivity problem and inspired by the general user behaviours, we formulate the user summarization process as multiple sequential decision-making processes. Each process is modeled by a comparison network for feature extraction and a selection network for decision making. The unsupervised reward and supervised reward, each with local and global parts, together contribute to the advancement of the policy. Experiments on two benchmark datasets show that CoSNet outperforms state-of-the-art unsupervised methods and most supervised methods.

REFERENCES
[1] Sijia Cai, Wangmeng Zuo, Larry S Davis, and Lei Zhang. 2018. Weakly-supervised video summarization using variational encoder-decoder and web prior. In Proceedings of the European Conference on Computer Vision (ECCV). 184–200.
[2] Jinseok Choi, Tae-Hyun Oh, and In So Kweon. 2018. Contextually customized video summaries via natural language. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 1718–1726.
[3] Sandra Eliza Fontes De Avila, Ana Paula Brandão Lopes, Antonio da Luz Jr, and Arnaldo de Albuquerque Araújo. 2011. VSUMM: A mechanism designed
to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters* 32, 1 (2011), 56–68.

[4] Ana Garcia del Molino, Xavier Boix, Joo-Hwee Lim, and Ah-Hwee Tan. 2017. Active video summarization: Customized summaries via on-line interaction with the user. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[5] Ehsan Elhamifar and M Clara De Paolis Kaluza. 2017. Online summarization via submodular and convex optimization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 1783–1791.

[6] Hehe Fan, Zhongwen Xu, Linchao Zhu, Chenggang Yang, Jianjun Ge, and Yi Yang. 2018. Watching a small portion could be as good as watching all: towards efficient video classification. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 795–811.

[7] Litong Feng, Ziyin Li, Zhanghui Kuang, and Wei Zhang. 2018. Extractive video summarizer with memory augmented neural networks. In *Proceedings of the 26th ACM international conference on Multimedia*. 976–983.

[8] Ana Garcia del Molino. 2016. First person view video summarization subject to the user needs. In *Proceedings of the 24th ACM international conference on Multimedia*. 1440–1444.

[9] Shuxiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. 2017. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 3389–3396.

[10] Michael Gylfi, Helmut Grabner, Hayriko Riemensdemeier, and Luc Van Gool. 2014. Creating summaries from user videos. In *European conference on computer vision*. Springer, 505–520.

[11] Michael Gylfi, Helmut Grabner, and Luc Van Gool. 2015. Video summarization by learning submodular objectives of subjective descriptions. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2296–2304.

[12] Xufeng He, Yang Hua, Tao Song, Zongguo Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan. 2019. Unsupervised Video Summarization with Attentive Conditional Generative Adversarial Networks. In *Proceedings of the 27th ACM International Conference on Multimedia*. 3090–3098.

[13] Xufeng He, Yang Hua, Tao Song, Zongguo Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan. 2019. Unsupervised Video Summarization with Attentive Conditional Generative Adversarial Networks. In *Proceedings of the 27th ACM International Conference on Multimedia*. 3090–3098.

[14] Huan-I Ho, Wei-Chen Chiu, and Yu-Chang Frank Wang. 2018. Summarizing First-Person Videos from Third Persons’ Points of View. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 70–85.

[15] Haojian Jin, Yale Song, and Koji Yatani. 2017. Elasticplay: Interactive video summarization with dynamic time budgets. In *Proceedings of the 25th ACM international conference on Multimedia*. 1164–1172.

[16] Junjie Jiang, Donghyun Cho, Dahlin Kim, Sanghyun Woo, and In So Kweon. 2019. Discriminative feature learning for unsupervised video summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33*. 8537–8544.

[17] Aidean Sharghi, Ali Borji, Chengtao Li, Tianbao Yang, and Boqing Gong. 2018. Improving sequential determinantal point processes for supervised video summarization. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 792–808.

[18] Anuj Rathore, Pravin Nagar, Chetan Arora, and CV Jawahar. 2019. Generating storylines with skipping recurrent neural networks. In *European Conference on Computer Vision (ECCV)*. 7902–7911.

[19] Robert Robertson, and Haibing Guan. 2019. Unsupervised Video Summarization with adversarial lstm networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3090–3098.

[20] Yair Pogel, Tavi Halperin, Chetan Arora, and Shmuel Peleg. 2015. Egosampling: Fast-forward and stereo for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4766–4776.

[21] Shiyue Lan, Rameswar Panda, Qi Zhu, and Amit K Roy-Chowdhury. 2018. Category-specific video summarization. In *European conference on computer vision*. Springer, 540–555.

[22] Donghyeon Cho, Dahun Kim, Sanghyun Woo, and In So Kweon. 2013. Large-scale video summarization using web image priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2698–2705.

[23] Bhesebro Mahaussen, Michael Lam, and Sinisa Todorovic. 2017. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 202–211.

[24] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Reddick, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533.

[25] Rameswar Panda, Abir Das, Ziyao Wu, Jan Ernst, and Amit K Roy-Chowdhury. 2017. Weakly supervised summarization of web videos. In *Proceedings of the IEEE International Conference on Computer Vision*. 3657–3666.

[26] Rameswar Panda and Amit K Roy-Chowdhury. 2017. Collaborative summarization of topic-related videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7083–7092.

[27] Robert Robertson, and Haibing Guan. 2019. Unsupervised Video Summarization with adversarial lstm networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8537–8544.

[28] Baihao Mirzae, Stefanie Jegelka, and Andreas Krause. 2018. Streaming learning submodular mixtures of objectives. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and MultiAgent Systems, 1397–1404.

[29] Baharan Mirzasoleiman, Stefanie Jegelka, and Andreas Krause. 2018. Streaming video summarization by learning from unlabeled data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4768–4779.

[30] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning visualabo, and Abhinav Gupta. 2016. Learning visual storylines with skipping recurrent neural networks. In *European conference on computer vision*. Springer, 78–92.

[31] Michel Silva, Washington Ramos, Joao Ferreira, Felipe Chamone, Mario Campos, and Erickson R Nascimento. 2018. A weighted sparse sampling and smoothing frame transition approach for semantic fast-forward first-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2383–2392.

[32] Yale Song, Jordi Villalatina, Amanda Stent, and Alejandro Jaime. 2015. Tsvm: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5179–5187.

[33] Dui Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[34] Arun Baljee Yasavdean, Michael Gylfi, Anna Volokitin, and Luc Van Gool. 2017. Query-adaptive video summarization via quality-aware relevance estimation. In *Proceedings of the 25th ACM international conference on Multimedia*. 582–590.

[35] Yingbo Li, Lique Wang, Weixiang Wang, and Boqing Gong. 2018. How local is the local diversity? Reinforcing sequential determinantal point processes with dynamic ground sets for supervised video summarization. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 151–167.

[36] Timothy P Lillicrap, Jonathan H Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971 (2015).

[37] Vladimir Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Reddick, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533.

[38] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8, 3–4 (1992), 229–256.
[52] Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, and Shilei Wen. 2019. Multi-Agent Reinforcement Learning Based Frame Sampling for Effective Untrimmed Video Recognition. In Proceedings of the IEEE International Conference on Computer Vision. 6222–6231.

[53] Shuwen Xiao, Zhou Zhao, Zijian Zhang, Xiaohui Yan, and Min Yang. 2020. Convolutional Hierarchical Attention Network for Query-Focused Video Summarization. arXiv preprint arXiv:2002.03740 (2020).

[54] Bo Xiong, Yannis Kalantidis, Deepi Ghaidyaram, and Kristen Grauman. 2019. Less is more: Learning highlight detection from video duration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 1258–1267.

[55] Jia Xu, Lopanoudra Mukherjee, Yin Li, Jameson Warner, James M Rehg, and Vikas Singh. 2015. Gaze-enabled egocentric video summarization via constrained submodular maximization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2235–2244.

[56] Ting Yao, Tao Mei, and Yong Rui. 2016. Highlight detection with pairwise deep ranking for first-person video summarization. In Proceedings of the IEEE conference on computer vision and pattern recognition. 982–990.

[57] Ting Yao, Tao Mei, and Yong Rui. 2016. Highlight detection with pairwise deep ranking for first-person video summarization. In Proceedings of the IEEE conference on computer vision and pattern recognition. 982–990.

[58] Ting Yao, Tao Mei, and Yong Rui. 2016. Highlight detection with pairwise deep ranking for first-person video summarization. In Proceedings of the IEEE conference on computer vision and pattern recognition. 982–990.

[59] Yujia Zhang, Michael Kampffmeyer, Xiaodan Liang, Min Tan, and Eric P Xing. 2018. Query-conditioned three-player adversarial network for video summarization. arXiv preprint arXiv:1807.06677 (2018).

[60] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video summarization with long short-term memory. In European conference on computer vision. Springer. 766–782.

[61] Ke Zhang, Kristen Grauman, and Fei Sha. 2018. Retrospective encoders for video summarization. In Proceedings of the European Conference on Computer Vision (ECCV). 383–399.

[62] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video summarization with long short-term memory. In European conference on computer vision. Springer. 766–782.

[63] Ke Zhang, Kristen Grauman, and Fei Sha. 2018. Retrospective encoders for video summarization. In Proceedings of the European Conference on Computer Vision (ECCV). 383–399.

[64] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. 2017. Hierarchical recurrent neural network for video summarization. In Proceedings of the 25th ACM international conference on Multimedia. 863–871.

[65] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. 2018. Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 7405–7414.

[66] Kaiyang Zhou, Yu Qiao, and Tao Xiang. 2018. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In Thirty-Second AAAI Conference on Artificial Intelligence.