Joint Video Summarization and Moment Localization by Cross-Task Sample Transfer

Hao Jiang
Peking University
jianghao@stu.pku.edu.cn

Yadong Mu*
Peking University
myd@pku.edu.cn

Abstract

Video summarization has recently engaged increasing attention in computer vision communities. However, the scarcity of annotated data has been a key obstacle in this task. To address it, this work explores a new solution for video summarization by transferring samples from a correlated task (i.e., video moment localization) equipped with abundant training data. Our main insight is that the annotated video moments also indicate the semantic highlights of a video, essentially similar to video summary. Approximately, the video summary can be treated as a sparse, redundancy-free version of the video moments. Inspired by this observation, we propose an importance Propagation based collaborative Teaching Network (iPTNet). It consists of two separate modules that conduct video summarization and moment localization, respectively. Each module estimates a frame-wise importance map for indicating keyframes or moments. To perform cross-task sample transfer, we devise an importance propagation module that realizes the conversion between summarization-guided and localization-guided importance maps. This way critically enables optimizing one of the tasks using the data from the other task. Additionally, in order to avoid error amplification caused by batch-wise joint training, we devise a collaborative teaching scheme, which adopts a cross-task mean teaching strategy to realize the joint optimization of the two tasks and provide robust frame-level teaching signals. Extensive experiments on video summarization benchmarks demonstrate that iPTNet significantly outperforms previous state-of-the-art video summarization methods, serving as an effective solution that overcomes the data scarcity issue in video summarization.

1. Introduction

In recent years, with the popularization of video-sharing platforms, the number of videos that record activities of daily living has witnessed an explosive growth. Techniques that help people quickly browse videos and search the key information become a valuable research topic. Video summarization [5, 20, 46], a technology that automatically extracts representative segments from untrimmed videos to concisely depict the original video content, has attracted increasing attention from both academia and industries.

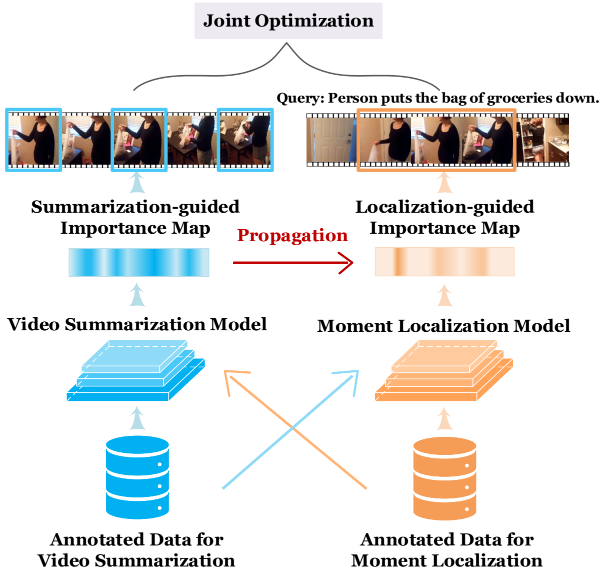

Joint Optimization
Query: Person puts the bag of groceries down.
Summarization-guided Importance Map
Propagation
Localization-guided Importance Map
Video Summarization Model
Moment Localization Model
Annotated Data for Video Summarization
Annotated Data for Moment Localization

Figure 1. The video summarization and video moment localization models can be jointly optimized using the proposed cross-task sample transfer.

The earlier work in video summarization mostly adopts hand-crafted heuristics to attain certain properties of frames (*e.g.*, diversity, representativeness) [11, 31, 34, 42, 49, 52]. These modeling paradigms are recently less used since the revival of deep learning techniques [2, 30, 50, 53, 56, 75, 85]. Specifically, the latest video summarization models have been empowered by recurrent neural networks (RNN) [24, 86, 87, 93] and the attention mechanism [29, 46, 70], which drastically advance the state-of-the-art.

Most of these methods employ a data-hungry supervised learning setting for training [24, 46, 93]. Despite the performance gains achieved by these efforts, current research has been still suffering from the scarcity of large-

[*] Corresponding author.

scale annotated video summaries, which require extensive time and effort to construct. Some weakly-supervised approaches [5, 11, 31] have been proposed to alleviate this problem. However, they not only require additional video auxiliary information, but are still hard to achieve competitive results. Heretofore, learning video summarization under limited data remains an untapped problem.

To address this problem, this work explores the idea of cross-task sample transfer from related tasks, particularly the video moment localization task that aims to temporally spot the video segments corresponding to an arbitrary sentence query. The idea is illustrated in Figure 1. Note that video moment localization is query-driven. We observe that the user-provided queries usually describe the key events in the video, thus the task is essentially correlated with video summarization. Considering the compactness of video summaries, they can be approximately regarded as sparse, redundancy-free version of the video moments, which shed light on transferring the abundant annotations in moment localization for helping the video summarization model. In light of this, we explore a collaborative optimization scheme for these two tasks. Nevertheless, it is non-trivial to achieve cross-task sample transfer satisfactorily. Overlooking the domain gap between the two tasks will inevitably cause the learned model to suffer from collaborative signals with domain bias and lead to performance degradation. Moreover, batch-wise joint training of the two models makes the optimization susceptible to current batch noise. This easily causes the error of one task to spread to another task, resulting in error amplification and failure to provide stable and robust collaborative signals.

In this paper, we propose an importance Propagation based collaborative Teaching Network (iPTNet), as shown in Figure 2. It consists of four parts: the *video summarization* module (SM), the *moment localization* module (LM), the *importance propagation* module (PM), and the *collaborative teaching* module (TM). To be more specific, SM and LM do the job of video summarization / moment localizaiton respectively, under the supervision of corresponding task-related data and annotations. The main functionality of PM is to connect the two frame-wise importance maps generated by SM and LM. The collaborative teaching module implements a cross-task mean teaching strategy and enforces the soundness of our main assumption (*i.e.*, the ground-truth video summaries can be approximately expanded into video moments via some inter-frame propagations). The main contributions of this work are summarized as below:

- To the best of our knowledge, this is the first work that utilizes a second correlated task (*i.e.*, video moment localization) with sufficient training data to help the training of video summarization. Through jointly optimizing two models, it surmounts the obstacle of insufficient annotated video summaries without the requirement of additional annotations or any auxiliary video information.
- To fully harness the ensembles of training data from two tasks, we devise an importance propagation algorithm, which realizes the conversion between the summarization-guided and localization-guided importance maps and thereby accomplishes cross-task sample transfer during model optimization.
- To avoid the error amplification caused by batch-wise joint training, we propose a collaborative teaching scheme based on a cross-task mean-teaching strategy for the modules SM and LM.

Extensive experiments conducted on video summarization benchmark datasets demonstrate that iPTNet significantly outperforms the state-of-the-art methods. The code and data of this work have been released to facilitate further research[1].

2. Related Work

Video summarization. Existing methods could be cast into three categories: unsupervised approaches [14–16, 28, 41, 77, 94], weakly supervised approaches [5, 48, 52, 60], and supervised approaches [20, 56, 57]. Recent unsupervised methods can be divided into dictionary based [16, 41, 94], subset selection based [14, 15], reinforcement learning based [96], and adversarial learning based [28, 77] methods. Weakly supervised approaches often harness auxiliary information (*e.g.*, web priors [11, 31], video titles [60]). For example, Song *et al.* [60] proposed to use title-based image search results to find visually important shots. Supervised approaches learn to generate video summaries based on manual annotations. Zhang *et al.* [86] firstly proposed a supervised learning technique for summarizing videos based on the annotated video summarization datasets such as SumMe [21], TVSum [60], and OVP [12]. Some methods explored the use of RNN [87, 93] and attention mechanisms [46, 70] to capture long-range representations in video sequences. Zhao *et al.* [93] proposed a structure-adaptive approach and integrated shot segmentation and video summarization into a hierarchical RNN. However, these methods rarely addressed the issue of insufficient summary annotations. Although weakly supervised methods can alleviate this problem to some extent, they still require auxiliary video information and are hard to achieve competitive results.

Video moment localization. The moment localization task aims to locate key events in the video specified by some natural language queries [1, 18, 33, 39, 40,

[1] https://code-website.wixsite.com/iptnet.

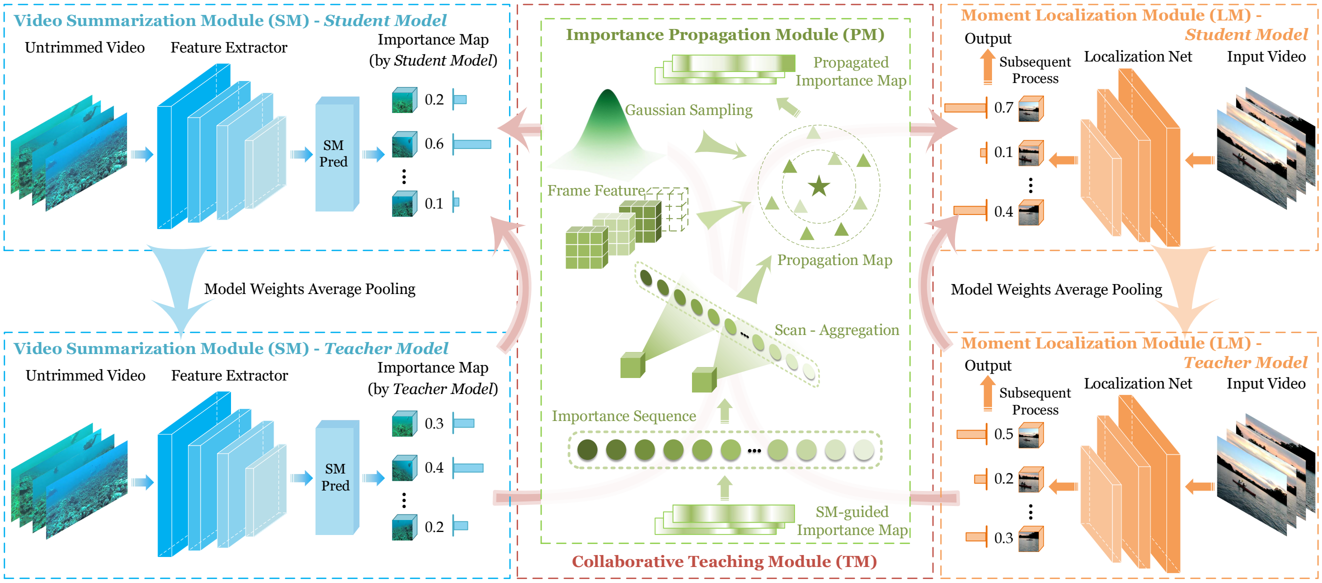


Figure 2. Illustration of the proposed importance Propagation based collaborative Teaching Network (iPTNet) for video summarization. It consists of four parts: the video summarization module (SM), the moment localization module (LM), the collaborative teaching module (TM), and the importance propagation module (PM). See the main text for more details.

47, 65, 88, 95]. Previous methods could be roughly divided into four categories: ranking-based [38, 73], anchor-based [78, 82], reinforcement learning-based [67, 69], and regression-based [45, 80]. Ranking-based methods adopt the matching and ranking architecture to localize temporal moments [8]. Anchor-based methods generate multiple anchors of different scales and select the anchor with the highest confidence [66, 91]. Reinforcement learning-based methods regard moment localization as a decision-making problem and the agent with certain policies is utilized to help moment localization [23]. Regression-based methods can flexibly regress temporal boundaries of localized moments without proposal generation [7, 79]. Zeng *et al.* [80] designed a regression network to regress the distance from each frame to the start / end frame of the video segment described by the query. Zhang *et al.* [89] proposed to model the temporal moment relations by a 2D map. Different from the previous work, we explore a new direction of moment localization, *i.e.*, joint training with the video summarization model to obtain informative cross-task assistance.

Teacher-student models. They have been used for knowledge distillation [3,9,10,35,59,64,71,97] and applied in various vision tasks, such as object detection [4, 13, 62], image segmentation [25, 26], video captioning [90], *etc.* Mean teacher model [63] was proposed to average model weights at different training iterations to form a target-generating teacher model. Yang *et al.* [74] developed an interactive form of self-training using mean teacher models for object detection. However, existing mean teacher models are designed for the same task and could not be directly used for cross-task joint optimization. To solve this problem, we propose a collaborative teaching mechanism that promotes the performance of multiple tasks simultaneously.

3. The Proposed Approach

3.1. Task Definition

Let $\mathcal{V} = \{v_i\}_{i=1}^{N}$ be the frame set of an untrimmed video, where N is the number of frames in $\mathcal{V}$. The video summarization model primarily seeks for a representative set of key frames as video summaries. Let $\mathcal{S} = \{s_j\}_{j=1}^{M}$ denotes the frame set of video summaries, where M is the number of frames in $\mathcal{S}$. Typically, M is below some pre-specified length proportion (say, 15%) of N. The subset $\mathcal{S}$ can be obtained by optimizing some objective functions that encode the belief about a good video summary. For the moment localization model, its goal is to precisely find the starting / ending timestamps for a semantic moment specified by a natural language query.

3.2. Overview of Network Design

In this section, we introduce the architecture of the proposed importance Propagation based collaborative Teaching Network (iPTNet). As shown in Figure 2, the proposed model is comprised of four modules, including the *video summarization* module (SM), the *moment localization* module (LM), the *importance propagation* module (PM), and the *collaborative teaching* module (TM). SM is designed for selecting keyshots based on the input video. LM takes the video and the text query as input, and aims to localize the corresponding moments according to the query. Importantly, both SM and LM will read all samples that are annotated for either task. Yet only the samples for video summarization activate the calculation of loss functions in SM in the forward pass. Similar treatment for LM. In the following sections, we describe each module in details.

3.3. Video Summarization Module (SM)

Given an input untrimmed video $\mathcal{V} = \{v_i\}_{i=1}^{N}$, we first extract the visual features of each frame v_i. In our implementation, GoogLeNet [61] is adopted for frame feature extraction that generates a 1,024-dimensional feature vector $\boldsymbol{h}_i$ for frame v_i. Other neural backbones (*e.g.*, ResNet) can be alternatively used. In addition, to further capture long-range correlations, we use the multi-head attention operation to enhance each $\boldsymbol{h}_i$. Let the final feature set be $\boldsymbol{H}' = \{\boldsymbol{h}'_1, \boldsymbol{h}'_2, ..., \boldsymbol{h}'_N\}$.

SM calculates an importance map $\boldsymbol{I}_{(Sum)}(\boldsymbol{H}')$ from $\boldsymbol{H}'$ in order to generate the final video summary. Suppose the importance map is calculated as follows:

$$\boldsymbol{I}_{(Sum)}(\boldsymbol{H}') = \mathcal{F}^{(I)}(\boldsymbol{H}'), \tag{1}$$

where $\mathcal{F}^{(I)}$ denotes a feed forward network. Unless otherwise clarified, configurations of all neural designs are postponed to the supplemental materials for space limitation.

We also predict the boundary offset $\boldsymbol{B}$ and the centrality score $\boldsymbol{C}$ of video frames [98] to help SM learn to locate key segments, where $\boldsymbol{B}$ is a 2D matrix that represents the offsets of each video frame from the left and right boundaries of the associated segment, and $\boldsymbol{C}$ is a 1D matrix that represents whether each frame is at the center of the segment. Suppose the calculations of $\boldsymbol{B}$ and $\boldsymbol{C}$ are conducted via:

$$\begin{cases} \boldsymbol{B} = \mathcal{F}^{(B)}(\boldsymbol{H}'), \\ \boldsymbol{C} = \mathcal{F}^{(C)}(\boldsymbol{H}'), \end{cases} \tag{2}$$

where $\mathcal{F}^{(B)}$ and $\mathcal{F}^{(C)}$ denote some feed forward networks for $\boldsymbol{B}, \boldsymbol{C}$, respectively. The loss function $\mathcal{L}_{SM}$ for optimizing SM is consistent with [98]. Non-maximum suppression is adopted to remove the predicted segments with large overlaps or low confidences, and the $0/1$ knapsack algorithm is employed to generate video summaries.

3.4. Moment Localization Module (LM)

Given an input untrimmed video $\mathcal{V} = \{v_i\}_{i=1}^{N}$ and the text query $\mathcal{W} = \{w_q\}_{q=1}^{L}$, where w_q represents a word in the query and L is the total number of words. LM aims to locate the video segment corresponding to the query. Similar to the video summarization module, we first extract the I3D feature [6] (which is the most widely-used feature in the moment localization literature) around each frame v_i and GloVe [51] for each word w_q. Both are fed into an additional feature encoding sub-network with convolutional layers and multi-head attention operations, which contextualizes both features. Denote the encoded visual features as $\widetilde{\boldsymbol{H}} = \{\widetilde{\boldsymbol{h}}_i\}_{i=1}^{N}$ and textual features as $\widetilde{\boldsymbol{W}} = \{\widetilde{\boldsymbol{w}}_q\}_{q=1}^{L}$.

Next, we employ the cross-modal attention module [55, 72, 76] that attends to visual and textual features simultaneously and captures the interactions between different modalities. The QGH module [83, 84] is adopted to calculate the importance map $\boldsymbol{I}_{(Loc)}(\widetilde{\boldsymbol{H}})$. The prediction and optimization of LM essentially follow the practice in [83]. We omit more details regarding the network design and optimization since designing novel moment localization module is not the main scope of this work.

Algorithm 1 Importance Propagation Algorithm

```
Input:
    The input video features, $\boldsymbol{H}'$;
    The importance map before propagation, $\boldsymbol{I}_{(Sum)}(\boldsymbol{H}')$;
    Half of the scan range for propagation, $s$;
Output:
    The importance map after propagation, $\boldsymbol{J}_{(Sum)}(\boldsymbol{H}')$;
for each $i \in \{1, 2, ..., N\}$ do
    $\chi_i^{(l)} = max(0, i - s)$;
    $\chi_i^{(r)} = min(i + s + 1, N + 1)$;
    $\Upsilon_i^{(l \leftarrow r)} = \{\boldsymbol{h}'_j | \chi_i^{(l)} \leq j < \chi_i^{(r)} \wedge \boldsymbol{h}'_j \in \boldsymbol{H}'\}$;
    $\mathcal{O}_i^{(l \leftarrow r)} = \text{SimFunction}(\Upsilon_i^{(l \leftarrow r)}, \boldsymbol{h}'_i)$;
    $\boldsymbol{I}_i \leftarrow \Big(\sum_{j:j \in \{\chi_i^{(l)}, ..., \chi_i^{(r)}\}} \boldsymbol{I}_j \cdot \mathcal{O}_i^j\Big) / \Big(\chi_i^{(r)} - \chi_i^{(l)}\Big)$;
end for
for each $i \in \{1, 2, ..., N\}$ do
    $\varrho_i^{(s)} = \{\boldsymbol{I}_j | max(0, i - s) \leq j < min(i + s + 1, N + 1), \boldsymbol{I}_j \in \boldsymbol{I}\}$;
    $\tilde{\varrho}_i^{(s)} = \arg\max_i(\varrho_i^{(s)})$;
    $\Delta \sim Gaussian(min(s, i) - \tilde{\varrho}_i^{(s)})$;
    $\Gamma \sim FeedForwardNetwork(\boldsymbol{H}')$;
    $\boldsymbol{J}_i \leftarrow \boldsymbol{I}_i \cdot \Delta + \Gamma$;
end for
$\boldsymbol{J}_{(Sum)}(\boldsymbol{H}') = \{\boldsymbol{J}_i | i \in \{1, 2, ..., N\}\}$;
return $\boldsymbol{J}_{(Sum)}(\boldsymbol{H}')$
```

3.5. Importance Propagation Module (PM)

As stated earlier, the main belief underlying our work is that the importance maps generated by SM and LM approximately correspond to the same set of semantic events, yet at different levels of compactness and redundancy. Guided by such belief, it is extremely important to establish connections between these two kinds of importance maps, such that collaborative teaching can be enabled. To this end, we craft an importance propagation algorithm for realizing the conversion between the summarization-guided and localization-guided importance maps, implemented by the

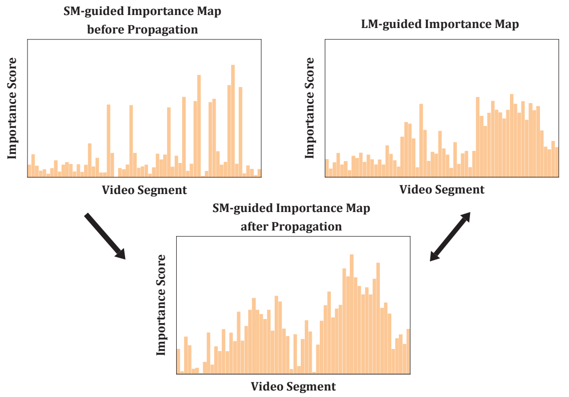


Figure 3. Illustration of the importance propagation process. Note that the process densifies the SM-guided importance map, returning a new map more amenable for video moment localization.

importance propagation module (PM). It takes the importance map $\boldsymbol{I}_{(Sum)}(\boldsymbol{H}')$ as input, and outputs $\boldsymbol{J}_{(Sum)}(\boldsymbol{H}')$ after propagation. The process is detailed in Algorithm 1.

For each frame i in the video, we first determine its scan range during importance propagation, controlled by a hyper-parameter s. As shown in lines 2-3, according to s we calculate the left endpoint $\chi_i^{(l)}$ and right endpoint $\chi_i^{(r)}$ of the scan range respectively. Next, we select the frame feature set $\Upsilon_i^{(l\leftarrow r)}$ required for importance propagation based on the obtained $\chi_i^{(l)}$ and $\chi_i^{(r)}$ (line 4). In order to calculate the weighting factor in the importance propagation process, we adopt some similarity functions (*e.g.*, cosine) to calculate the similarities $\mathcal{O}_i^{(l\leftarrow r)}$ between the visual feature $\boldsymbol{h}_i'$ of the i-th frame and the feature $\boldsymbol{h}_j'$ ($\chi_i^{(l)} \leq j < \chi_i^{(r)}$) in the set $\Upsilon_i^{(l\leftarrow r)}$ (line 5). Afterwards, we perform importance propagation within the scan range of frame i according to the obtained weighting factors $\mathcal{O}_i^{(l\leftarrow r)}$, as shown in line 6, and then update $\boldsymbol{I}_i$ with the propagation results of the first stage.

Considering that the process of importance propagation is not only affected by the frame visual features, but also by the relative distance between different frames. In light of this, we find the position $\tilde{\varrho}_i^{(s)}$ of the maximum frame importance $\varrho_i^{(s)}$ in the scan range $\boldsymbol{I}_j, j \in [max(0, i-s), min(i+s+1, N+1))$ centered on video frame i, and calculate the relative distance offset between $\tilde{\varrho}_i^{(s)}$ and frame i in the scan range (lines 9-10). Since the Gaussian distribution can approximate the attenuation of the influence of adjacent video frames with increasing distance, we perform Gaussian sampling based on the relative distance offset and obtain the sampling result Δ (line 11). In addition, we also include a learnable feedforward network to calculate the shift term Γ based on the video feature $\boldsymbol{H}'$. Γ is added to the scaled frame importance map $\boldsymbol{I}_i \cdot \Delta$ to generate the propagated importance map, as shown in lines 12-13. In this way, we return $J_{(Sum)}(\boldsymbol{H}')$ as the final result. Figure 3 presents an example of importance propagation.

3.6. Collaborative Teaching Module (TM)

Once we have obtained $\boldsymbol{I}_{(Sum)}(\boldsymbol{H}')$ and $\boldsymbol{I}_{(Loc)}(\widetilde{\boldsymbol{H}})$ from SM and LM respectively, the next challenge is how to effectively combine these two modules to achieve cross-task joint optimization. Direct batch-wise joint training is susceptible to the noises in current batch and numerically unstable. This easily causes the error of one task to spread to the other task, resulting in error amplification and failure of convergence.

We mitigate this issue using the mean teacher mechanism. Existing mean teacher methods [19, 63] propose to average the student model weights at different training iterations to form the teacher model, which continuously aggregates and updates the distilled knowledge from students. However, these methods are limited to the same task and not directly applicable for cross-task joint optimization. As shown in Figure 2, we propose a collaborative teaching module (TM) over the video summarization and moment localization modules, which enables the teacher model of one task to transfer the distilled knowledge to the student model of the other task through importance maps, thereby achieving collaborative training. For video summarization, let its teacher / student network be $\boldsymbol{\Pi}_{Sum}^{(T)}$, $\boldsymbol{\Pi}_{Sum}^{(S)}$ respectively. Likewise we introduce the notations $\boldsymbol{\Pi}_{Loc}^{(T)}$, $\boldsymbol{\Pi}_{Loc}^{(S)}$ for moment localization. The parameters of $\boldsymbol{\Pi}_{Sum}^{(T)}$ are determined by:

$$\mathbb{E}^{(t+1)}\left[\boldsymbol{\theta}\left(\boldsymbol{\Pi}_{Sum}^{(S)}\right)\right] = \gamma\mathbb{E}^{(t)}\left[\boldsymbol{\theta}\left(\boldsymbol{\Pi}_{Sum}^{(S)}\right)\right] + (1-\gamma)\boldsymbol{\theta}\left(\boldsymbol{\Pi}_{Sum}^{(S)}\right), \tag{3}$$

where $\mathbb{E}^{(t+1)}\left[\boldsymbol{\theta}\left(\boldsymbol{\Pi}_{Sum}^{(S)}\right)\right]$ and $\mathbb{E}^{(t)}\left[\boldsymbol{\theta}\left(\boldsymbol{\Pi}_{Sum}^{(S)}\right)\right]$ denote the parameters of model $\boldsymbol{\Pi}_{Sum}^{(T)}$ in the iteration $t+1$ and t, respectively. $\boldsymbol{\theta}\left(\boldsymbol{\Pi}_{Sum}^{(S)}\right)$ indicates the parameters of model $\boldsymbol{\Pi}_{Sum}^{(S)}$ in the iteration $t+1$ and γ is the ensembling momentum. Likewise, we can also define the updating formula for $\boldsymbol{\Pi}_{Loc}^{(T)}$, which is omitted for saving space.

Following the previous description, we denote the importance map generated by $\boldsymbol{\Pi}_{Sum}^{(S)}$ and $\boldsymbol{\Pi}_{Sum}^{(T)}$ as $\boldsymbol{I}_{Sum}^{(S)}(\boldsymbol{H}')$ and $\boldsymbol{I}_{Sum}^{(T)}(\boldsymbol{H}')$, respectively. The importance map generated by the student network should be as close as possible to the map generated by the teacher network, so we measure the difference between $\boldsymbol{I}_{Sum}^{(S)}(\boldsymbol{H}')$ and $\boldsymbol{I}_{Sum}^{(T)}(\boldsymbol{H}')$ by:

$$\begin{aligned} D(\boldsymbol{I}_{Sum}^{(S)}(\boldsymbol{H}' | \boldsymbol{\theta}(\boldsymbol{\Pi}_{Sum}^{(S)})), \boldsymbol{I}_{Sum}^{(T)}(\boldsymbol{H}' | \mathbb{E}[\boldsymbol{\theta}(\boldsymbol{\Pi}_{Sum}^{(S)})])) \\ = \mathbb{E}_{\boldsymbol{h}' \sim \boldsymbol{I}_{Sum}^{(S)}(\boldsymbol{h}')}\left[\log \boldsymbol{I}_{Sum}^{(S)}(\boldsymbol{h}') - \log \boldsymbol{I}_{Sum}^{(T)}(\boldsymbol{h}')\right]. \end{aligned} \tag{4}$$

The calculation of the difference between $\boldsymbol{I}_{Loc}^{(S)}(\widetilde{\boldsymbol{H}})$ and

$\boldsymbol{I}^{(T)}_{Loc}(\widetilde{\boldsymbol{H}})$ is similarly defined as follows:

$$\begin{aligned} &D(\boldsymbol{I}^{(S)}_{Loc}(\widetilde{\boldsymbol{H}}|\boldsymbol{\theta}(\boldsymbol{\Pi}^{(S)}_{Loc})), \boldsymbol{I}^{(T)}_{Loc}(\widetilde{\boldsymbol{H}}|\mathbb{E}[\boldsymbol{\theta}(\boldsymbol{\Pi}^{(S)}_{Loc})])) \\ &\quad = \mathbb{E}_{\widetilde{\boldsymbol{h}}\sim \boldsymbol{I}^{(S)}_{Loc}(\widetilde{\boldsymbol{h}})}\left[\log \boldsymbol{I}^{(S)}_{Loc}(\widetilde{\boldsymbol{h}}) - \log \boldsymbol{I}^{(T)}_{Loc}(\widetilde{\boldsymbol{h}})\right]. \end{aligned} \tag{5}$$

When we use the LM data (*i.e.*, samples from the moment localization task, more details in Sec. 3.7) for collaborative teaching, LM could generate more trustworthy importance maps than SM. Therefore, we make $\boldsymbol{\Pi}^{(S)}_{Sum}$ learn from $\boldsymbol{\Pi}^{(T)}_{Loc}$ and measure the difference between $\boldsymbol{J}^{(S)}_{Sum}(\widetilde{\boldsymbol{H}})$ (obtained by $\boldsymbol{I}^{(S)}_{Sum}(\widetilde{\boldsymbol{H}})$ from PM) and $\boldsymbol{I}^{(T)}_{Loc}(\widetilde{\boldsymbol{H}})$ as:

$$\begin{aligned} &D(\boldsymbol{J}^{(S)}_{Sum}(\widetilde{\boldsymbol{H}}|\boldsymbol{\theta}(\boldsymbol{\Pi}^{(S)}_{Sum})), \boldsymbol{I}^{(T)}_{Loc}(\widetilde{\boldsymbol{H}}|\mathbb{E}[\boldsymbol{\theta}(\boldsymbol{\Pi}^{(S)}_{Loc})])) \\ &\quad = \mathbb{E}_{\widetilde{\boldsymbol{h}}\sim \boldsymbol{J}^{(S)}_{Sum}(\widetilde{\boldsymbol{h}})}\left[\log \boldsymbol{J}^{(S)}_{Sum}(\widetilde{\boldsymbol{h}}) - \log \boldsymbol{I}^{(T)}_{Loc}(\widetilde{\boldsymbol{h}})\right]. \end{aligned} \tag{6}$$

Similarly, when the SM data is used for collaborative teaching, SM could generate more trustworthy importance maps than LM[2], so we make $\boldsymbol{\Pi}^{(S)}_{Loc}$ learn from $\boldsymbol{\Pi}^{(T)}_{Sum}$ and calculate the difference between $\boldsymbol{J}^{(T)}_{Sum}(\boldsymbol{H}')$ and $\boldsymbol{I}^{(S)}_{Loc}(\boldsymbol{H}')$:

$$\begin{aligned} &D(\boldsymbol{J}^{(T)}_{Sum}(\boldsymbol{H}'|\mathbb{E}[\boldsymbol{\theta}(\boldsymbol{\Pi}^{(S)}_{Sum})]), \boldsymbol{I}^{(S)}_{Loc}(\boldsymbol{H}'|\boldsymbol{\theta}(\boldsymbol{\Pi}^{(S)}_{Loc}))) \\ &\quad = \mathbb{E}_{\boldsymbol{h}'\sim \boldsymbol{J}^{(T)}_{Sum}(\boldsymbol{h}')}\left[\log \boldsymbol{J}^{(T)}_{Sum}(\boldsymbol{h}') - \log \boldsymbol{I}^{(S)}_{Loc}(\boldsymbol{h}')\right]. \end{aligned} \tag{7}$$

Regardless of which task data is used, we can always make $\boldsymbol{\Pi}^{(S)}_{Loc}$ and $\boldsymbol{\Pi}^{(S)}_{Sum}$ learn from each other. The reason lies in when the SM data is used, $\boldsymbol{\Pi}^{(S)}_{Sum}$ can help $\boldsymbol{\Pi}^{(S)}_{Loc}$ learning; otherwise, $\boldsymbol{\Pi}^{(S)}_{Loc}$ can help $\boldsymbol{\Pi}^{(S)}_{Sum}$ learning. Take the use of the LM data as an example, we calculate the difference between $\boldsymbol{J}^{(S)}_{Sum}(\widetilde{\boldsymbol{H}})$ and $\boldsymbol{I}^{(S)}_{Loc}(\widetilde{\boldsymbol{H}})$ as follows:

$$\begin{aligned} &D(\boldsymbol{J}^{(S)}_{Sum}(\widetilde{\boldsymbol{H}}|\boldsymbol{\theta}(\boldsymbol{\Pi}^{(S)}_{Sum})), \boldsymbol{I}^{(S)}_{Loc}(\widetilde{\boldsymbol{H}}|\boldsymbol{\theta}(\boldsymbol{\Pi}^{(S)}_{Loc}))) \\ &\quad = \mathbb{E}_{\widetilde{\boldsymbol{h}}\sim \boldsymbol{J}^{(S)}_{Sum}(\widetilde{\boldsymbol{h}})}\left[\log \boldsymbol{J}^{(S)}_{Sum}(\widetilde{\boldsymbol{h}}) - \log \boldsymbol{I}^{(S)}_{Loc}(\widetilde{\boldsymbol{h}})\right]. \end{aligned} \tag{8}$$

The optimization function of TM needs to be discussed in two situations, namely when the LM data is used for collaborative teaching and when the SM data is used. When the LM data is used, the optimization function is drawn by:

$$\mathcal{L}_{TM\_LL} = D(\boldsymbol{I}^{(S)}_{Loc}, \boldsymbol{I}^{(T)}_{Loc}) + \mathcal{L}_{LM} \tag{9}$$

$$\mathcal{L}_{TM\_LS} = D(\boldsymbol{J}^{(S)}_{Sum}, \boldsymbol{I}^{(T)}_{Loc}) + D(\boldsymbol{J}^{(S)}_{Sum}, \boldsymbol{I}^{(S)}_{Loc}) \tag{10}$$

where $\mathcal{L}_{LM}$ represents the loss of LM itself, $\mathcal{L}_{TM\_LL}$ is the total optimization function of LM in the presence of the LM data, and $\mathcal{L}_{TM\_LS}$ is the optimization function of SM in the presence of the LM data. When the SM data is used, the optimization function is drawn by:

$$\mathcal{L}_{TM\_SS} = D(\boldsymbol{I}^{(S)}_{Sum}, \boldsymbol{I}^{(T)}_{Sum}) + \mathcal{L}_{SM} \tag{11}$$

$$\mathcal{L}_{TM\_SL} = D(\boldsymbol{J}^{(T)}_{Sum}, \boldsymbol{I}^{(S)}_{Loc}) + D(\boldsymbol{J}^{(S)}_{Sum}, \boldsymbol{I}^{(S)}_{Loc}) \tag{12}$$

where $\mathcal{L}_{SM}$ represents the loss of SM itself, $\mathcal{L}_{TM\_SS}$ is the total optimization function of SM in the presence of the SM data, and $\mathcal{L}_{TM\_SL}$ is the optimization function of LM in the presence of the SM data.

Note that the teacher network does not perform gradient back-propagation, calculate or store gradients. Thus it does not significantly increase the computational complexity and memory usage.

3.7. Data Flows during Model Learning

The data used in training is an ensemble of annotated samples from two tasks. It is thus critical to understand the data flows in our model. The data flows of our framework can be divided into two parts: the data flow of the LM data and the data flow of the SM data. When the LM data is used, it only activates the loss functions $\mathcal{L}_{TM\_LL}$ and $\mathcal{L}_{TM\_LS}$ to update the parameters. Similarly, when the SM data is fed, it activates $\mathcal{L}_{TM\_SS}$ and $\mathcal{L}_{TM\_SL}$. Importantly, all samples will be fed into both SM and LM, whatever sources they are from. This way, a video originally annotated for moment localization can also help refine the parameters of the video summarization module, through optimizing the loss in TM, and vice versa. This treatment is key for the realization of cross-task sample transfer between the two tasks and overcoming the problem of insufficient video summarizes.

Considering the severe imbalance of the SM and LM data, we adopt a non-mixing sampling strategy in practice. The optimization starts from training SM for several epochs (*e.g.*, 20) purely with SM data, and then starts the collaborative teaching that first reads only the LM data to update both modules and then samples only the SM data. Once a fixed number of epochs are reached, it will exit the collaborative teaching and repeat the above process until convergence.

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate our iPTNet framework on two benchmarks: SumMe [21] and TVSum [60]. The former is a collection of 25 user-generated videos and covers multiple types of scenes (*e.g.*, cooking), with each video having frame-level importance scores annotated by $15-18$ users. TVSum consists of 50 YouTube videos annotated by 20 users, belonging to 10 categories (*e.g.*, parades). Following the previous approaches [54, 92], we use the additional Open Video Project (OVP) [12] and YouTube datasets [12] to perform experiments under the augmented and transfer settings. More details of the datasets are shown in Table 1. If only frame-level importance scores are provided, we follow previous methods to convert them into keyshot-based summaries for evaluation [86]. For LM, we adopt

[2] In this situation, LM only takes video features as input to calculate the importance map and does not localize moments, since the SM data does not have query-moment annotations.

Table 1. Statistics of the SumMe [21], TVSum [60], OVP [12], and YouTube [12] datasets, including the number of videos, the number of annotations, and the durations of videos in the datasets.

Dataset	Num of Videos	Num of annotations	Avg Duration	Min Duration	Max Duration
SumMe [21]	25	15-18	2min 26s	32s	5min 24s
TVSum [60]	50	20	3min 55s	1min 23s	10min 47s
OVP [12]	50	5	1min 38s	46s	3min 29s
YouTube [12]	39	5	3min 16s	9s	9min 32s

Table 2. Overall performance comparison (in terms of F-score %) on SumMe and TVSum datasets under the canonical setting.

Method	SumMe Dataset	TVSum Dataset
Video MMR [37]	26.6	–
LiveLight [94]	–	46.0
ERSUM [36]	43.1	59.4
MSDS-CC [44]	40.6	52.3
vsLSTM [86]	37.6	54.2
dppLSTM [86]	38.6	54.7
SUM-GAN [43]	41.7	56.3
A-AVS [27]	43.9	59.4
M-AVS [27]	44.4	61.0
SASUM [68]	45.3	58.2
DR-DSN [96]	42.1	58.1
TS-STN [24]	46.1	60.0
SUM-FCN [54]	48.8	58.4
VASNet [17]	49.7	61.4
DSNet$_{\text{anc\_based}}$ [98]	50.2	62.1
DSNet$_{\text{anc\_free}}$ [98]	51.2	61.9
RSGN [92]	45.0	60.1
iPTNet (Ours)	**54.5**	**63.4**

Charades-STA [18, 58] with $12,408$ and $3,720$ sentence-moment pairs for training and testing, respectively.

Evaluation protocol. Following previous work [54, 85], we evaluate the proposed method under the canonical, augmented, and transfer settings. In the canonical setting, we randomly divide the dataset into 5 splits, using 80% of the dataset for training, and the remaining for evaluation. In the augmented setting, for a given dataset, we randomly select 20% data for evaluation, and the rest 80% of the dataset augmented with the other three datasets are used for training. In the transfer setting, for the given dataset, the model is trained on other three datasets and evaluated on the remaining dataset. In all settings, we run the models five times and report the averaged results [98].

We adopt F-score to measure the matching degree of the generated summaries $\mathcal{S}_i$ and the ground-truth summaries $\hat{\mathcal{S}}_i$ for video i. The precision and recall according to the temporal overlap between $\mathcal{S}_i$ and $\hat{\mathcal{S}}_i$ are calculated by $Precision = \frac{|\mathcal{S}_i \cap \hat{\mathcal{S}}_i|}{|\mathcal{S}_i|}, Recall = \frac{|\mathcal{S}_i \cap \hat{\mathcal{S}}_i|}{|\hat{\mathcal{S}}_i|}$, and the F-score is calculated by $\frac{2 \times Precision \times Recall}{Precision + Recall}$. For videos with multiple user-annotated summaries, we follow [22, 86] to calculate the metrics.

Implementation details. In SM, we follow the previous work [22, 86, 92] to extract 1024-dimensional visual features from the `poo5` layer of GoogLeNet [61] pre-trained on ImageNet. In LM, we use 3D ConvNet to extract the visual features [81, 83]. The learning rates of SM and LM are set to 0.00001 and 0.0005, respectively. γ is set to 0.3 in TM. s is set to 10 in PM. The number of heads in the multi-head attention layer is 8. The hidden size of the video summarization model is set to 128. Adam [32] optimizer is adopted to update our model.

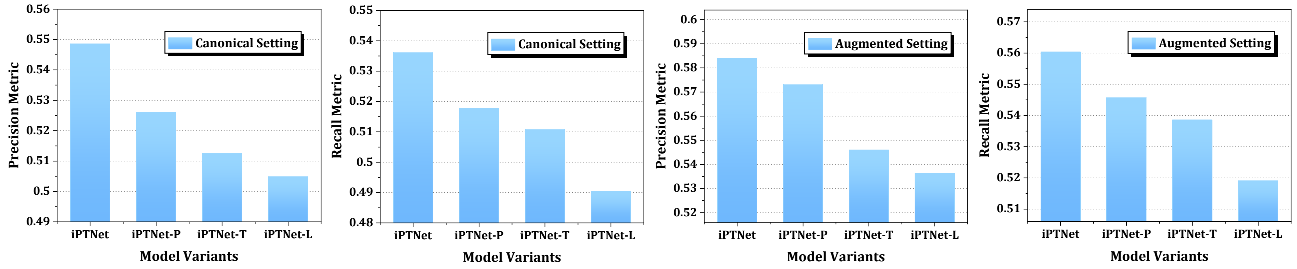
4.2. Performance Comparison

We compare the proposed method iPTNet with previous methods on SumMe and TVSum. Table 2 shows the performance comparison under the canonical setting. We observe that iPTNet achieves the best performance on both datasets. Compared with previous methods, such as RSGN, DSNet, VASNet, and SUM-FCN, our method considers employing the moment localization model to help the training of video summarization, which demonstrates that the idea of joint training facilitates the model performance. In addition, the distilled knowledge obtained by the proposed collaborative teaching mechanism provides helpful teaching signals for SM in the joint learning process, thereby further improving the performance.

We further conduct experiments under the augmented and transfer settings on SumMe and TVSum along with additional OVP and YouTube datasets, as illustrated in Table 3. We observe that iPTNet still outperforms the previous work under the two settings. This shows that the proposed cross-task sample transfer endows our model the competitive ability and helps the model maintain superior performance under different experimental settings.

4.3. Ablation Investigation

In this section, we study the effectiveness of each component of the proposed method by comparing iPTNet with several variants (*i.e.*, iPTNet-P, iPTNet-T, iPTNet-L). Specifically, in iPTNet-P, we remove the proposed importance propagation algorithm in PM. In iPTNet-T, we remove the proposed collaborative teaching mechanism and do not employ the teacher network for knowledge distillation and model teaching. As for iPTNet-L, we do not use the moment localization model to help the training of video summarization. Figure 4 reports the performance comparison of iPTNet, iPTNet-P, iPTNet-T, and iPTNet-



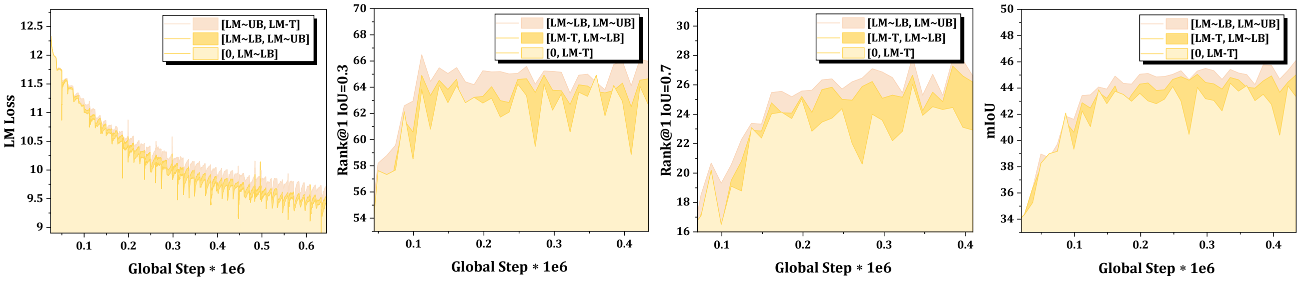
(a) precision (canonical setting)

(b) recall (canonical setting)

(c) precision (augmented setting)

(d) recall (augmented setting)

Figure 4. Performance comparison of different model variants on SumMe dataset under the canonical and augmented settings.



(a) loss curve

(b) Rank@1, IoU=0.3

(c) Rank@1, IoU=0.7

(d) mIoU

Figure 5. Effectiveness analysis of the proposed method to the video moment localization model.

Table 3. Performance comparison (in terms of F-score %) on SumMe and TVSum under the augmented and transfer settings.

Method	SumMe			TVSum		
	C	A	T	C	A	T
vsLSTM [86]	37.6	41.6	40.7	54.2	57.9	56.9
dppLSTM [86]	38.6	42.9	41.8	54.7	59.6	58.7
A-AVS [27]	43.9	44.6	–	59.4	60.8	–
M-AVS [27]	44.4	46.1	–	61.0	61.8	–
DR-DSN [96]	42.1	43.9	42.6	58.1	59.8	58.9
SUM-FCN [54]	48.8	50.2	45.0	58.4	59.1	57.4
DSNet$_{\text{a\_base}}$ [98]	50.2	50.7	46.5	62.1	63.9	59.4
DSNet$_{\text{a\_free}}$ [98]	51.2	53.3	47.6	61.9	62.2	58.0
RSGN [92]	45.0	45.7	44.0	60.1	61.1	**60.0**
iPTNet (Ours)	**54.5**	**56.9**	**49.2**	**63.4**	**64.2**	59.8

L on the SumMe dataset in terms of Precision and Recall. We observe that iPTNet-L performs the worst, indicating the effectiveness of our idea of employing LM to help SM training. iPTNet-T performs better than iPTNet-L, but worse than iPTNet-P. It shows that our collaborative teaching mechanism is significant and the distilled knowledge provides helpful teaching signals for SM in the joint training process. In addition, compared with iPTNet, the performance degradation of iPTNet-P verifies the effectiveness of the proposed importance propagation algorithm.

4.4. Study the Effect of iPTNet to LM

In this part, we study the effect of iPTNet on LM learning. We train the moment localization model separately under the setting of using and removing the proposed joint training method, and observe the performance. Figure 5 summarizes the experimental results, wherein LM-T indicates independent training of LM. We repeat multiple experiments and calculate the mean and standard deviation of the experimental results. LM∼UB and LM∼LB represent the upper and lower bounds of the experimental results of LM. We observe that LM with the proposed method achieves lower loss and higher performance (*i.e.*, in terms of mIoU, Rank@1, IoU=0.3, and Rank@1, IoU=0.7), which demonstrates that iPTNet also provides beneficial information for LM training and helps LM perform better.

5. Concluding Remarks

We address the data scarcity issue in video summarization, *i.e.*, employing the moment localization model to help the training of video summarization. We devise a new framework iPTNet, which adopts a collaborative teaching scheme to perform cross-task mean teaching and provide robust teaching signals. An importance propagation algorithm is designed to deal with domain gaps and realize cross-task sample transfer. Extensive experiments on video summarization datasets demonstrate the effectiveness of iPTNet.

Acknowledgement: This work is supported by National Key R&D Program of China (2020AAA0104401) and Beijing Natural Science Foundation (Z190001).

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 2

[2] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. Video summarization using deep neural networks: A survey. *IEEE*, 2021. 1

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *CVPR*, 2020. 3

[4] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *CVPR*, 2019. 3

[5] Sijia Cai, Wangmeng Zuo, Larry S Davis, and Lei Zhang. Weakly-supervised video summarization using variational encoder-decoder and web prior. In *ECCV*, 2018. 1, 2

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 4

[7] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. Rethinking the bottom-up framework for query-based video localization. In *AAAI*, 2020. 3

[8] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *AAAI*, 2019. 3

[9] Zhihao Chen, Lei Zhu, Liang Wan, Song Wang, Wei Feng, and Pheng-Ann Heng. A multi-task mean teacher for semi-supervised shadow detection. In *CVPR*, 2020. 3

[10] Ali Cheraghian, Shafin Rahman, Pengfei Fang, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Semantic-aware knowledge distillation for few-shot class-incremental learning. In *CVPR*, 2021. 3

[11] Wen-Sheng Chu, Yale Song, and Alejandro Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *CVPR*, 2015. 1, 2

[12] 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. *PRL*, 2011. 2, 6, 7

[13] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, 2021. 3

[14] Ehsan Elhamifar and M Clara De Paolis Kaluza. Online summarization via submodular and convex optimization. In *CVPR*, 2017. 2

[15] Ehsan Elhamifar, Guillermo Sapiro, and S Shankar Sastry. Dissimilarity-based sparse subset selection. *TPAMI*, 2015. 2

[16] Ehsan Elhamifar, Guillermo Sapiro, and Rene Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *CVPR*, 2012. 2

[17] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention. In *ACCV*, 2018. 7

[18] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017. 2, 7

[19] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020. 5

[20] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. In *NeurIPS*, 2014. 1, 2

[21] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *ECCV*, 2014. 2, 6, 7

[22] Michael Gygli, Helmut Grabner, and Luc Van Gool. Video summarization by learning submodular mixtures of objectives. In *CVPR*, 2015. 7

[23] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *AAAI*, 2019. 3

[24] Siyu Huang, Xi Li, Zhongfei Zhang, Fei Wu, and Junwei Han. User-ranking video summarization with multi-stage spatio–temporal representation. *TIP*, 2018. 1, 7

[25] Xinyue Huo, Lingxi Xie, Jianzhong He, Zijie Yang, Wengang Zhou, Houqiang Li, and Qi Tian. Atso: Asynchronous teacher-student optimization for semi-supervised image segmentation. In *CVPR*, 2021. 3

[26] Mingi Ji, Seungjae Shin, Seunghyun Hwang, Gibeom Park, and Il-Chul Moon. Refine myself by teaching myself: Feature refinement via self-knowledge distillation. In *CVPR*, 2021. 3

[27] Zhong Ji, Kailin Xiong, Yanwei Pang, and Xuelong Li. Video summarization with attention-based encoder–decoder networks. *TCSVT*, 2019. 7, 8

[28] Yunjae Jung, Donghyeon Cho, Dahun Kim, Sanghyun Woo, and In So Kweon. Discriminative feature learning for unsupervised video summarization. In *AAAI*, 2019. 2

[29] Yunjae Jung, Donghyeon Cho, Sanghyun Woo, and In So Kweon. Global-and-local relative position embedding for unsupervised video summarization. In *ECCV*, 2020. 1

[30] Atsushi Kanehira, Luc Van Gool, Yoshitaka Ushiku, and Tatsuya Harada. Viewpoint-aware video summarization. In *CVPR*, 2018. 1

[31] Aditya Khosla, Raffay Hamid, Chih-Jen Lin, and Neel Sundaresan. Large-scale video summarization using web-image priors. In *CVPR*, 2013. 1, 2

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 7

[33] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 2

[34] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 1

[35] Xin Li, Deng-Ping Fan, Fan Yang, Ao Luo, Hong Cheng, and Zicheng Liu. Probabilistic model distillation for semantic correspondence. In *CVPR*, 2021. 3

[36] Xuelong Li, Bin Zhao, and Xiaoqiang Lu. A general framework for edited video and raw video summarization. *TIP*, 2017. 7

[37] Yingbo Li and Bernard Merialdo. Multi-video summarization based on video-mmr. In *WIAMIS*, 2010. 7

[38] Bingbin Liu, Serena Yeung, Edward Chou, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Temporal modular networks for retrieving complex compositional activities in videos. In *ECCV*, 2018. 3

[39] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *CVPR*, 2021. 2

[40] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *ICCV*, 2019. 2

[41] Shiyang Lu, Zhiyong Wang, Tao Mei, Genliang Guan, and David Dagan Feng. A bag-of-importance model with locality-constrained coding based feature learning for video summarization. *TMM*, 2014. 2

[42] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *CVPR*, 2013. 1

[43] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial lstm networks. In *CVPR*, 2017. 7

[44] Jingjing Meng, Suchen Wang, Hongxing Wang, Junsong Yuan, and Yap-Peng Tan. Video summarization via multiview representative selection. In *ICCV*, 2017. 7

[45] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *CVPR*, 2020. 3

[46] Saiteja Nalla, Mohit Agrawal, Vishal Kaushal, Ganesh Ramakrishnan, and Rishabh Iyer. Watch hours in minutes: Summarizing videos with user intent. In *ECCV*, 2020. 1, 2

[47] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *CVPR*, 2021. 2

[48] Rameswar Panda, Abir Das, Ziyan Wu, Jan Ernst, and Amit K Roy-Chowdhury. Weakly supervised summarization of web videos. In *ICCV*, 2017. 2

[49] Rameswar Panda and Amit K Roy-Chowdhury. Collaborative summarization of topic-related videos. In *CVPR*, 2017. 1

[50] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Sumgraph: Video summarization via recursive graph modeling. In *ECCV*, 2020. 1

[51] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 4

[52] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. Category-specific video summarization. In *ECCV*, 2014. 1, 2

[53] Mrigank Rochan and Yang Wang. Video summarization by learning from unpaired data. In *CVPR*, 2019. 1

[54] Mrigank Rochan, Linwei Ye, and Yang Wang. Video summarization using fully convolutional sequence networks. In *ECCV*, 2018. 6, 7, 8

[55] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. In *ICLR*, 2017. 4

[56] Aidean Sharghi, Boqing Gong, and Mubarak Shah. Query-focused extractive video summarization. In *ECCV*, 2016. 1, 2

[57] Aidean Sharghi, Jacob S Laurel, and Boqing Gong. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *CVPR*, 2017. 2

[58] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 7

[59] Jie Song, Haofei Zhang, Xinchao Wang, Mengqi Xue, Ying Chen, Li Sun, Dacheng Tao, and Mingli Song. Tree-like decision distillation. In *CVPR*, 2021. 3

[60] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *CVPR*, 2015. 2, 6, 7

[61] 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. 4, 7

[62] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. In *CVPR*, 2021. 3

[63] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 3, 5

[64] Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *CVPR*, 2021. 3

[65] Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. Structured multi-level interaction network for video moment localization via language query. In *CVPR*, 2021. 2

[66] Jingwen Wang, Lin Ma, and Wenhao Jiang. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *AAAI*, 2020. 3

[67] Weining Wang, Yan Huang, and Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *CVPR*, 2019. 3

[68] Huawei Wei, Bingbing Ni, Yichao Yan, Huanyu Yu, Xiaokang Yang, and Chen Yao. Video summarization via semantic attended networks. In *AAAI*, 2018. 7

[69] Jie Wu, Guanbin Li, Si Liu, and Liang Lin. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In *AAAI*, 2020. 3

[70] Shuwen Xiao, Zhou Zhao, Zijian Zhang, Ziyu Guan, and Deng Cai. Query-biased self-attentive network for query-focused video summarization. *TIP*, 2020. 1, 2

[71] Zeyu Xiao, Xueyang Fu, Jie Huang, Zhen Cheng, and Zhiwei Xiong. Space-time distillation for video super-resolution. In *CVPR*, 2021. 3

[72] Caiming Xiong, Victor Zhong, and Richard Socher. Dynamic coattention networks for question answering. In *ICLR*, 2017. 4

[73] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, 2019. 3

[74] Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, and Lei Zhang. Interactive self-training with mean teachers for semi-supervised object detection. In *CVPR*, 2021. 3

[75] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *CVPR*, 2016. 1

[76] Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. Fast and accurate reading comprehension by combining self-attention and convolution. In *ICLR*, 2018. 4

[77] Li Yuan, Francis EH Tay, Ping Li, Li Zhou, and Jiashi Feng. Cycle-sum: cycle-consistent adversarial lstm networks for unsupervised video summarization. In *AAAI*, 2019. 2

[78] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *TPAMI*, 2020. 3

[79] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, 2019. 3

[80] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *CVPR*, 2020. 3

[81] Yawen Zeng, Da Cao, Xiaochi Wei, Meng Liu, Zhou Zhao, and Zheng Qin. Multi-modal relational graph for cross-modal video moment retrieval. In *CVPR*, 2021. 7

[82] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*, 2019. 3

[83] Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Natural language video localization: A revisit in span-based question answering framework. *TPAMI*, 2021. 4, 7

[84] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *ACL*, 2020. 4

[85] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Summary transfer: Exemplar-based subset selection for video summarization. In *CVPR*, 2016. 1, 7

[86] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *ECCV*, 2016. 1, 2, 6, 7, 8

[87] Ke Zhang, Kristen Grauman, and Fei Sha. Retrospective encoders for video summarization. In *ECCV*, 2018. 1, 2

[88] Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. Multi-stage aggregated transformer network for temporal language localization in videos. In *CVPR*, 2021. 2

[89] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, 2020. 3

[90] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *CVPR*, 2020. 3

[91] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *CVPR*, 2020. 3

[92] Bin Zhao, Haopeng Li, Xiaoqiang Lu, and Xuelong Li. Reconstructive sequence-graph network for video summarization. *TPAMI*, 2021. 6, 7, 8

[93] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In *CVPR*, 2018. 1, 2

[94] Bin Zhao and Eric P Xing. Quasi real-time summarization for consumer videos. In *CVPR*, 2014. 2, 7

[95] Hao Zhou, Chongyang Zhang, Yan Luo, Yanjun Chen, and Chuanping Hu. Embracing uncertainty: Decoupling and debias for robust temporal grounding. In *CVPR*, 2021. 2

[96] Kaiyang Zhou, Yu Qiao, and Tao Xiang. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *AAAI*, 2018. 2, 7, 8

[97] Jinguo Zhu, Shixiang Tang, Dapeng Chen, Shijie Yu, Yakun Liu, Mingzhe Rong, Aijun Yang, and Xiaohua Wang. Complementary relation contrastive distillation. In *CVPR*, 2021. 3

[98] Wencheng Zhu, Jiwen Lu, Jiahao Li, and Jie Zhou. Dsnet: A flexible detect-to-summarize network for video summarization. *TIP*, 2020. 4, 7, 8