Global-and-Local Relative Position Embedding for Unsupervised Video Summarization

Yunjae Jung[1], Donghyeon Cho[2], Sanghyun Woo[1], and In So Kweon[1]

[1] Korea Advanced Institute of Science and Technology, Daejeon, Korea
yun9298a@gmail.com, {shwoo93, iskweon77}@kaist.ac.kr
[2] Chungnam National University, Daejeon, Korea
cdh12242@gmail.com

Abstract. In order to summarize a content video properly, it is important to grasp the sequential structure of video as well as the long-term dependency between frames. The necessity of them is more obvious, especially for unsupervised learning. One possible solution is to utilize a well-known technique in the field of natural language processing for long-term dependency and sequential property: self-attention with relative position embedding (RPE). However, compared to natural language processing, video summarization requires capturing a much longer length of the global context. In this paper, we therefore present a novel input decomposition strategy, which samples the input both globally and locally. This provides an effective temporal window for RPE to operate and improves overall computational efficiency significantly. By combining both **G**lobal-and-**L**ocal input decomposition and **RPE** together, we come up with **GL-RPE**. Our approach allows the network to capture both local and global interdependencies between video frames effectively. Since GL-RPE can be easily integrated into the existing methods, we apply it to two different unsupervised backbones. We provide extensive ablation studies and visual analysis to verify the effectiveness of the proposals. We demonstrate our approach achieves new state-of-the-art performance using the recently proposed rank order-based metrics: Kendall's τ and Spearman's ρ. Furthermore, despite our method is unsupervised, we show ours perform on par with the fully-supervised method.

Keywords: Video Summarization, Relative Position Embedding, Unsupervised Learning.

1 Introduction

Video summarization is a task selecting keyframes from the untrimmed whole video, and those selected keyframes should represent entire input video frames. As content videos have recently begun to flood through various video platforms such as Youtube, there is a growing demand for video summarization techniques. In line with this demand, there have been a lot of video summarization-related researches: conventional methods [20,10,25,17,16,18,12,24,13,15,31,6,23,8], supervised learning based methods [5,7,40,39,28,42,43,27], and recent unsupervised methods [36,19,44,27,35,9,26].

As the content video becomes longer, it is difficult to generate a video summary that takes into account the entire story of the video, without considering the long-term dependency on the time axis. Therefore, previous methods that focus mainly on semantic objects, action, motion, and diversity show clear limitations. Even the latest LSTM-based methods [42,43,19] are vulnerable to long-term-term dependency. Inspired by self-attention [33], one of the most widely used technologies in the natural language processing (NLP) field [4,1,37], we try to solve the long-term dependency problem for the content video summarization in this paper. Furthermore, we incorporate relative position information [29] with self-attention (RPE) to overcome shortcomings of self-attention not dealing with the sequential properties in the video.

However, directly applying the self-attention to the entire video brings two unfavorable issues in practice. First, the large dimension of the feature for each frame inhibits an efficient relation computation. Second, due to the lengthy content video, long-term relation modeling becomes very challenging. A natural solution to overcome these intractabilities is the decomposition of input video using sampling. One may attempt to sample a certain amount of input video frames sequentially. Though, in this case, the difference between the relative positions of the last frame in the previous batch set and the first frame in the current batch set could be large even though they are actually very close. Therefore, we instead sample the input video frames, both locally and globally. The standard sequential sampling is the local sampling, whereas the stridden sampling corresponds to global sampling. We see the local and global sampling methods compensate each other and cancel out the errors caused by the previous naive solution. Combining input decomposition with RPE, our method can successfully consider not only the long-term dependency but also the sequential properties of content video effectively. We call the proposed method **G**lobal and **L**ocal-**R**elative **P**osition **E**mbedding (GL-RPE).

As far as we know, this work is the first attempt to apply self-attention with relative position representation for unsupervised video summarization. Our GL-RPE not only shows the state-of-the-art performance in the recently introduced rank order statistics-based evaluation metric [21], but it can also be combined with various backbone networks [19,9] for video summarization. Moreover, we show that GL-RPE with an unsupervised method achieves better performance than conventional supervised method [40]. Our contributions can be summarized as follows:

1. To our best knowledge, it is the first time that self-attention with relative position embedding is explored in the video summarization task.
2. We present a novel method called **GL-RPE** to handle both long-term dependency and sequential properties of the content video effectively. Our proposed method is general, thus can be easily integrated into the existing unsupervised video summarization approaches.
3. We conduct extensive ablation studies and provide intuitive visual analysis to validate the effectiveness of the proposed method.

4. With the GL-RPE, we achieve new state-of-the-art performance in the recently proposed rank order statistics-based metrics: Kendall's τ and Spearman's ρ. Notably, unsupervised learning approaches combined with GL-RPE outperform existing supervised based approach.

2 Related Work

In this section, we review the most relevant works, including recent deep learning-based video summarization approaches, and self-attention.

Supervised With the progress of deep neural network, supervised learning-based video summarization approaches [40,42] emerged as a promising solution and outperformed the previous hand-crafted feature-based methods [30,20,18,13,12]. Zhang *et al.* [40] firstly proposed a deep network for supervised video summarization, using the datasets containing human-made annotations such as TVSum [30], OVP [2] and SumMe [6]. Also, the LSTM-based models were introduced to handle a diverse range of temporal information. The follow-up studies [42,43] proposed a hierarchical recurrent neural network that is more powerful in exploiting long-term temporal dependency among frames.

Unsupervised Recently, unsupervised methods are receiving renewed attention because video summaries are highly subjective and significantly lack human animations in practice. Based on the assumption that features of good summary can reconstruct the features of the full original video, Mahasseni *et al.* [19] proposed a GAN-based method to supervise LSTM networks without human annotations. In [44], Zhou *et al.*considered unsupervised video summarization task as a sequential decision-making process. The authors then proposed an end-to-end deep summarization network (DSN) using reinforcement learning. Jung *et al.* [9] extend the work of [19] by introducing a two-stream network to handle both local and global frames. Besides, the variance loss is designed to avoid a trivial solution (i.e., identity mapping).

Self-attention Vaswani *et al.* [33] first introduced the concept of self-attention. It captures long-range relations by explicitly attending to all the features in the word sequence, which allows the model to build a direct relationship with other long-distance representations. Due mainly to its powerful distant relation modeling ability, it replaces commonly used recurrent architectures and is widely adopted in various natural language processing tasks [4,1,37]. The proposed formulation has been applied to other fields as well: object/action recognition [34], image generation [38], and image restoration [41]. We also attempt to utilize the self-attention for the unsupervised video summarization. However, we empirically observe that directly applying self-attention does not give meaningful improvement due to lengthy video frames. We thus propose to sample the video frames globally and locally to make input compatible with self-attention and improve the computational efficiency at the same time.

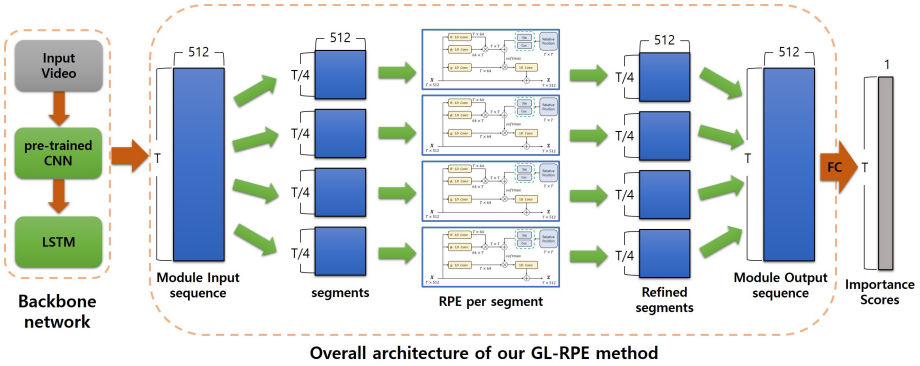

Fig. 1: The overview of our approach, including both the backbone summarization model [19,9], and the proposed GL-RPE. First, the backbone network embeds the T video frames to $T \times 512$ features. We then divide the input sequence into a total of 4 segments globally and locally (we set to 4 for the illustration) (Sec. 3.3). Each feature segment is refined with RPE (Sec. 3.1 and Sec. 3.2). The enhanced features segments are merged back in the original order for the final prediction.

Position Encoding Unlike RNN and LSTM, the self-attention cannot capture position information by design. This is critical, considering that the model is otherwise is entirely invariant to the sequence order, which is harmful for video summarization. To overcome this issue, we adopt relative position embedding [29], which ensures translation-equivariance property and allows the model to generalize unseen sequence length during training. We empirically confirm that relative position indeed helps to capture the sequential properties of video content, improving the video summarization performance further.

3 Proposed Method

In this section, we describe our key solution of the global-and-local relative position embedding module. The module is designed to aggregate the global context non-locally [33], and to be aware of the relative position between frames [29]. An apparent distinction with previous works [33,29,4] is that our target task deals with videos that are relatively longer than the word sequences. In fact, we observe that the direct application of the module to the video brings marginal improvement. To make input compatible, we thus propose to decompose the input video sequences into two scales, i.e., global and local. We then associate the frames with the proposed module. We evaluate our method by integrating the proposed module into the recent state-of-the-art unsupervised video summarization models; VAE-GAN [19] and CSNet [9]. The overview is shown in Fig. 1. We show our approach can successfully extract global and local inter-frame depen-

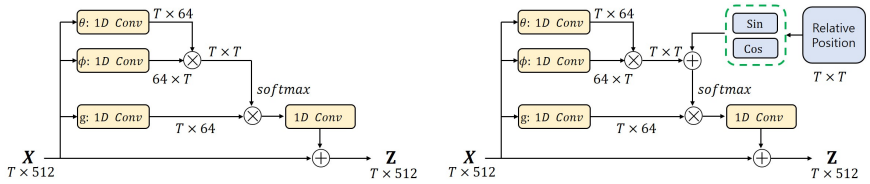

Fig. 2: (a) Self-attention embedding (SAE). The input X is refined by self-attention layers. (b) Relative position embedding (RPE). The input X is reinforced by considering relative positions. The $\oplus$ and $\otimes$ represent addition and matrix multiplication respectively.

dencies, and thus it boosts the baseline performance significantly. We detail our proposals below.

3.1 Video Self-Attention Embedding (SAE)

To capture and utilize the inter-frame relations, we design a module that is based on the scaled dot-product attention [33,34]. The attention layer first transforms an input feature into queries (Q), keys (K), and values (V) using linear embedding matrices. The affinity matrix is then obtained through the matrix multiplication of queries and keys. We then normalize the computed affinity matrix and fetch the values based on it (see Fig. 2-(a)). Note that we squeeze the spatial axis to only focus on extracting the temporal relations. The self-attention embedding module can be expressed as:

$$\begin{aligned} y &= \text{softmax}((W_\theta\, x)^T\, W_\phi\, x)\, W_g\, x, \\ Z &= x + W_z\, y, \end{aligned} \tag{1}$$

where W_θ, W_ϕ, W_g, and W_z denote linear embedding layers. We apply SAE to the outputs of LSTM in the base architectures. Since the past memories of LSTM are diluted as time-step accumulates, we employ SAE to complement this by its long-range temporal relation encoding ability.

3.2 Video Relative Position Embedding (RPE)

The original scaled dot-product attention does not explicitly model relative or absolute position information in its structure. To alleviate the lack of position information, we extend the SPE with relative position representation [29]. The relative position representation satisfies the translation-equivariance property, which is helpful when dealing with the sequence of frames, and also encourages the model to generalize well on the unseen sequence length during training. By incorporating the relative position, the module knows by how far two positions

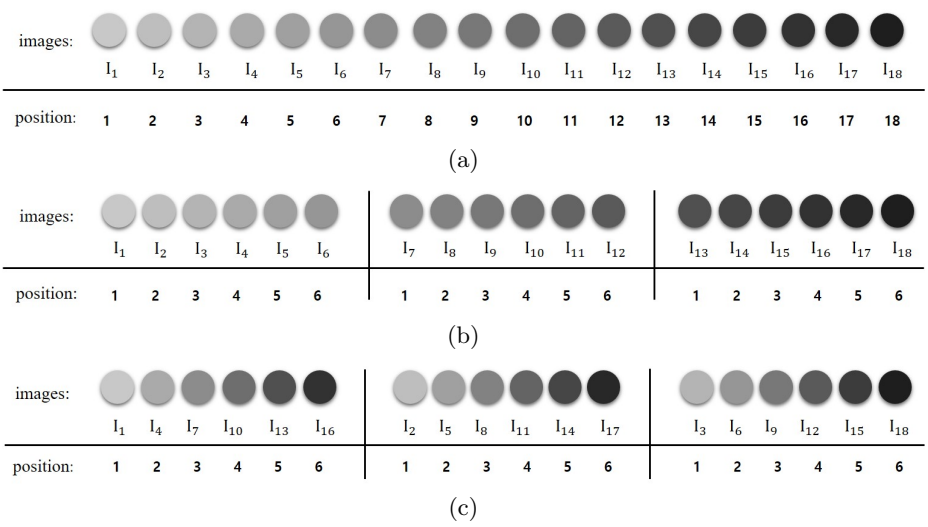

Fig. 3: The concept of proposed global-and-local decomposition. (a) Original video frames with the according indices, (b) Local sampling of video frames, and (c) Global sampling by skipping.

are apart in a sequence. This involves learning a relative position embedding for each possible pairwise distance between query and key. We describe our method below. First, the differences in frame indices between query and key are computed. Then, the different wavelengths of sinusoid functions are utilized to embed the relative distances. Finally, the encoded relative positions are embedded into $T \times T$ matrix as illustrated in Fig. 2-(b). Our relative position embedding (RPE) can be formulated as follow:

$$\begin{aligned} RP_{(rpos,i+j)} &= \sin\left(\frac{rpos}{freq^{((i+j)/d)}}\right), \\ RP_{(rpos,i+j+1)} &= \cos\left(\frac{rpos}{freq^{((i+j)/d)}}\right), \\ a &= (W_\theta\, x)^T\, W_\phi\, x + RP, \\ y &= \mathrm{softmax}(a)\, W_g\, x, \\ Z &= x + W_z\, y, \end{aligned} \tag{2}$$

where $rpos$ denotes relative positions between frames and is calculated as $j - i$. d is set to $2\,T$, and $freq$ is set to 10000. W_θ, W_ϕ, W_g, and W_z are 1-D convolutions.

3.3 Global-and-Local Input Decomposition

To handle very long videos, we present a novel global-and-local input decomposition technique. We begin by illustrating the basic case. Please refer the Fig. 3. Consider the general case of Fig. 3-(a). As the length T of a sequence increases (e.g., long-duration videos), the position embedding matrix of $T \times T$ becomes proportionally large. This makes the model difficult to capture the fine-grained relational features between the distant frames. In other words, the embedding matrix becomes less discriminative. Moreover, in terms of model learning, only the small subset of frames might dominate the learning process when a softmax is used in a large matrix. Thus, we see it suffices to perform the computation over a small fraction of frames. Motivated by our observation, we explore a new approach to tackle these challenges; our key idea is to sample the frames globally and locally and compute the relative position embedding for each separately. More specifically, for the global sampling, we skip the frames given the fixed stride rate as shown in Fig. 3-(c). For the local sampling, we set sampling stride to 1 as described in Fig. 3-(b). Then, we compute the relative position embedding in parallel (i.e., both global and local), and finally, merge them back to obtain the relative position embedding matrix. When the temporal size of an input x is T, the global and local segments can be described as:

$$\begin{aligned} x &= x_1 \oplus x_2 \oplus x_3 \oplus \cdots \oplus x_{T-1} \oplus x_T, \\ x^G_{(n,k)} &= x_k \oplus x_{n+k} \oplus x_{2n+k} \oplus \cdots \oplus x_{T-n+k}, \\ x^L_{(n,k)} &= x_{\lfloor \frac{T}{n}(k-1) \rfloor + 1} \oplus x_{\lfloor \frac{T}{n}(k-1) \rfloor + 2} \oplus \cdots \oplus x_{\lfloor \frac{T}{n} k \rfloor}, \end{aligned} \tag{3}$$

where $\oplus$ indicates the operation of concatenation along the temporal dimension, and n is the total number of segments, and k is a k-th feature in the n-th segment. The global and local segments are denoted to $x^G_{(n,k)}$ and $x^L_{(n,k)}$ respectively. We see the proposed input decomposition not only enhances the computation efficiency but also facilitates the exploitation of global and local inter-frame dependencies effectively.

3.4 Complexity Analysis

We combine **G**lobal-and-**L**ocal input decomposition with **RPE**, and come up with our final model, **GL-RPE**. The global-and-local input decomposition not only provides effective window size for the RPE operation but also improves the computational efficiency. Given a $T \times C$ input feature sequences, where T and C denote the total number of frames and channel dimensions of the feature, respectively, the total computational complexity of RPE is $O(CT^2)$. With the input decomposition of N segments, the complexity significantly reduces to $O(\frac{CT^2}{N})$. In this paper, we set the number of segments to 8 after conducting thorough parameter analysis in the experiment section (see Table 1).

4 Experiments

The implementation details are explained in Section 4.1. The benchmark datasets and evaluation metrics are in Section 4.2 and Section 4.3. Both the F-score [40] and the recently proposed rank-order correlation coefficients: Kendall's τ [11] and Spearman's ρ [45] are detailed. In Section 4.4, extensive ablation studies are carried out. In particular, we evaluate the impact of our major proposals: input decomposition, self-attention, and relative position. Since our approach is general, we can easily apply it to the existing methods. Thus, in Section 4.5, we show that our approach consistently boosts the state-of-the-art baselines with large margins, demonstrating its efficacy. Combined with CSNet [9], we achieve new state-of-the-art performance on TVSum [30] benchmark.

4.1 Implementation Details

We develop the proposed method in our Pytorch platform [22]. The ADAM [14] optimizer is used with the learning rate of 1e-4. It is decreased by 0.1 for every step size 10. The input video is sampled with 2 frames per second, and its spatial resolution is resized to 224×224. Every T frames in the video are forwarded to GoogLeNet [32], which is pre-trained on ImageNet [3]. This results in $T \times 1024$ features. Finally, the 1024 channel dimension is reduced to 512 using bidirectional LSTM. We apply our GL-RPE method on these features (see Fig. 1).

In our relative position embedding (RPE), the internal $T \times T$ matrix is expensive in terms of memory and computation. This induces an inefficient feature learning. The proposed global-and-local concept alleviates these problems effectively by decomposing the input sequence into n segments. As a result, the matrix becomes a size of $(T/n) \times (T/n)$ for each segment. We then apply the proposed self-attention modules to each segment. The enhanced feature segments after the refinement are merged back in the original order. Since our approach is more like a module, experiments are mostly conducted combined with the recent state-of-the-art backbone models [19,9].

4.2 Datasets

We use TVSum [30] and SumMe [6] datasets for the experiments. The TVSum dataset contains 50 videos up to 10 minutes, and 20 users annotate the importance score for each frame. Since each user has a different opinion on how important the frames are, evaluation of the individual user-label is conducted separately. Then, the results are averaged to measure overall performance. The SumMe dataset provides 25 videos up to 6 minutes and is also labeled on a per-frame importance score by a maximum of 18 users. Both TVSum and SumMe provide suitable forms of labels to measure F-score, which measures the intersection of selected frames based on importance scores. On the contrary, the recently suggested Kendall's τ and the Spearman's ρ are directly computed on importance scores, and only TVSum has a proper form of labels for the metrics [21].

4.3 Evaluation metric

F-score Video F-score is formulated in [40]. They suggest three experimental settings: 'Canonical', 'Augmented', and 'Transfer'. First, the 'Canonical' is a plain mode of dividing one dataset into the training and test set. Second, the 'Augmented' setting includes additional data in the training set. Lastly, the 'Transfer' setting excludes the test data, which is used in training in the 'Canonical' setting. After then, lots of follow-up studies [40,44,43,27,26,19,9] benchmarked their approaches using F-score. We detail the F-score formula below.

The kernel temporal segmentation (KTS) [24] is used to produce scene change boundaries. The key-shot is then selected based on the kernel-wise importance scores. For a given video, we consider the predicted key-shot (A) and the ground truth key-shot (B). The precision (P) and the recall (R) are accordingly computed as:

$$\begin{aligned} P &= \frac{\text{overlap of } A \text{ and } B}{\text{duration of } A}, \\ R &= \frac{\text{overlap of } A \text{ and } B}{\text{duration of } B}. \end{aligned} \tag{4}$$

Finally, the F-score is then obtained as follows:

$$\text{F-score} = \frac{2 * P * R}{P + R}. \tag{5}$$

Rank correlation coefficients While KTS-based F-score is known to be effective, recent study [21] points out that a randomly generated summary can, in fact, achieve similar F-score as the state-of-the-art methods. Therefore, as an alternative to the F-score, the rank correlation coefficients are presented. By exploiting well-established statistics that compare the ordinal association, the similarity between ground truth and predicted importance scores are much well evaluated than the F-score. In particular, Kendall's τ [11] and Spearman's ρ [45] correlation coefficients are adopted. With the recently presented metrics [21], the randomized summary now produces 0 scores while the human summary achieves the best. We thus consider rank-based metrics are more reliable than the F-score for the accurate video summary evaluation.

In this work, we benchmark our method using both F-score and rank-based metrics. We show our method achieves new state-of-the-art performance on rank-based metrics.

4.4 Ablation Study

We conduct ablation studies to verify the effectiveness of our major proposals empirically. We first show the impact of adopting self-attention embedding. We then combine it with the global-and-local input decomposition. While adopting self-attention brings positive effect, we observe the marginal improvement

Method	Kendall's τ	Spearman's ρ
Baseline	0.025	0.034
SAE	0.034	0.045
SAE+Global$_2$	0.038	0.050
SAE+Local$_2$	0.040	0.053
SAE+GL$_2$	0.037	0.048
SAE+Global$_4$	0.058	0.076
SAE+Local$_4$	0.057	0.075
SAE+GL$_4$	0.059	0.078
SAE+Global$_6$	0.061	0.079
SAE+Local$_6$	0.063	0.082
SAE+GL$_6$	0.060	0.079
SAE+Global$_8$	0.065	0.085
SAE+Local$_8$	0.065	0.085
SAE+GL$_8$	**0.066**	**0.087**
SAE+Global$_{10}$	0.061	0.080
SAE+Local$_{10}$	0.064	0.083
SAE+GL$_{10}$	0.064	0.084

(a) SAE

Method	Kendall's τ	Spearman's ρ
Baseline	0.025	0.034
RPE	0.033	0.044
RPE+Global$_2$	0.033	0.044
RPE+Local$_2$	0.037	0.049
RPE+GL$_2$	0.039	0.051
RPE+Global$_4$	0.056	0.074
RPE+Local$_4$	0.057	0.075
RPE+GL$_4$	0.058	0.076
RPE+Global$_6$	0.060	0.078
RPE+Local$_6$	0.060	0.079
RPE+GL$_6$	0.062	0.081
RPE+Global$_8$	0.064	0.084
RPE+Local$_8$	0.067	0.088
RPE+GL$_8$	**0.070**	**0.091**
RPE+Global$_{10}$	0.063	0.082
RPE+Local$_{10}$	0.065	0.085
RPE+GL$_{10}$	0.066	0.086

(b) RPE

Table 1: (a) Ablation study for global-and-local self-attention embedding (GL-SAE). (b) Ablation study for global-and-local relative position embedding (GL-RPE). The TVSum [30] dataset is used in both tables. The CSNet [9] is used as a backbone model.

without the input decomposition. Combining both the global-and-local input decomposition and the self-attention, we come up with the GL-RPE method. We show that the proposed GL-RPE dramatically improves the baseline scores. The ablation results are summarized in Table 1.

Baseline We adopt the state-of-the-art unsupervised video summarization method, CSNet [9], as a backbone model for the experiment. It produces scores of 0.025 and 0.034 for Kendall's τ and Spearman's ρ, respectively. We set these scores as a baseline.

Impact of Self-attention embedding We begin by introducing the self-attention embedding (SAE). We see the positive effect of SAE. Specifically, the SAE increases the baseline scores from 0.025 and 0.034 to 0.034 and 0.045. The results show that the long-term, global dependency modeling is crucial for the video summarization task. In the meantime, the relative position embedding (RPE) increases the baseline scores from 0.025 and 0.033 to 0.034 and 0.044. While RPE outperforms the baseline, we do not observe meaningful improvement over the SAE, despite the incorporation of relative position information.

	SumMe			TVSum		
Method	Can.	Aug.	Tr.	Can.	Aug.	Tr.
DPP-LSTM [40]	38.6	42.9	41.8	54.7	59.6	58.7
DR-DSN [44]	41.4	42.8	42.4	57.6	58.4	57.8
HSA-RNN [43]	-	44.1	-	-	59.8	-
SUM-FCN [27]	47.5	51.1	44.1	56.8	59.2	58.2
UnpairedVSN [26]	-	47.5	41.6	-	55.6	55.7
GAN [19]	39.1	43.4	-	51.7	59.5	-
CSNet [9]	**51.3**	52.1	45.1	58.8	59.0	59.2
CSNet+GL+RPE	50.2	-	-	**59.1**	-	-

Table 2: F-score (%) of existing methods including recent state-of-the-art approach.

We see this is because the embedding matrix of $T \times T$ is inefficiently large for the effective position embedding, and thus it brings no remarkable enhancement.

Impact of Input Decomposition We now investigate the impact of input decomposition. In this experiment, we attempt to confirm two main arguments empirically. 1) Input decomposition is essential for the self-attention embedding. 2) Using both the global and local decomposition produces finer representation. We experiment with 5 different numbers of input segments: 2, 4, 6, 8, and 10. We also report 3 different input decomposition methods: global-only (Global_n), local-only (Local_n), global-and-local (GL_n). Regardless of the input segment numbers, in Table 1, we can observe the general tendency of performance improvement with the input decomposition. One interesting point to note is that, as the number of segments increases, the performance improvement becomes large. The performance eventually saturates at 8. This shows that the input decomposition is indeed effective for capturing the inter-frame relations, and there exists an effective processing window size (i.e., $\frac{T}{8} \times \frac{T}{8}$) when using self-attention modules. Meanwhile, we explore the effect of using both the global and local input segments. We observe that the impact of global and local decomposition becomes apparent when using relative position information, and the number of segments increases. The relative position allows the module to be aware of the distance, and this information becomes crucial when dealing with both the global and local segments. The $\text{RPE} + \text{GL}_8$ shows the best results of 0.070 and 0.091 for Kendall's τ and Spearman's ρ. We use this configuration for the following experiments.

As a brief summary, we use self-attention with relative position information (RPE). Moreover, to effectively process the video content, we decompose the input globally and locally (Global-and-local input decomposition). We combine both proposals (GL-RPE) and successfully exploit global dependency and the sequential properties of video content effectively.

Method	Kendall's τ	Spearman's ρ
Random	0.000	0.000
dppLSTM [40]	0.042	0.055
DR-DSN [44]	0.020	0.026
GAN [19]	0.024	0.032
GAN+RPE	0.033	0.044
GAN+GL+RPE	0.064	0.084
CSNet [9]	0.025	0.034
CSNet+RPE	0.033	0.044
CSNet+GL+RPE	**0.070**	**0.091**
Human	0.177	0.204

Table 3: Comparison with the state-of-the-art methods. The TVSum dataset is used in this table.

4.5 Comparison with the state-of-the-art methods

We compare our results with the existing state-of-the-arts using both F-score and rank-based metrics (*i.e.,* Kendall's τ and Spearman's ρ). The results are summarized in Table 2 and Table 3.

F-score As the most existing methods only provide F-scores in their work, we also follow the standard evaluation protocol to benchmark our method. Our CSNet+GL+RPE are measured on both SumMe [6] and TVSum [30] datasets on the 'Canonical' experimental setting. We achieve state-of-the-art performance in the TVSum dataset. In the case of SumMe, our result is comparable to the existing method.

Rank correlation coefficients We now use a more reliable evaluation metric, Kendall's τ and Spearman's ρ, which are recently proposed in [21]. Since the proposed GL-RPE is general, we see it can be easily integrated into the existing networks. Here, we use two different unsupervised models [19,9] to evaluate the impact of GL-RPE. As can be shown in Table 3, RPE improves the baseline performances. With the additional global-and-local input decomposition (GL), the improvement becomes much significant. The tendency holds for both backbones. This again shows that capturing both the global and local inter-frame relations is crucial (RPE), and the impact increases when the input is decomposed into an adequate size (GL). Note that we achieve state-of-the-art results of 0.070 and 0.091 when the GL-RPE is combined with CSNet [9]. Moreover, we outperform the supervised model dppLSTM [40] with a large margin.

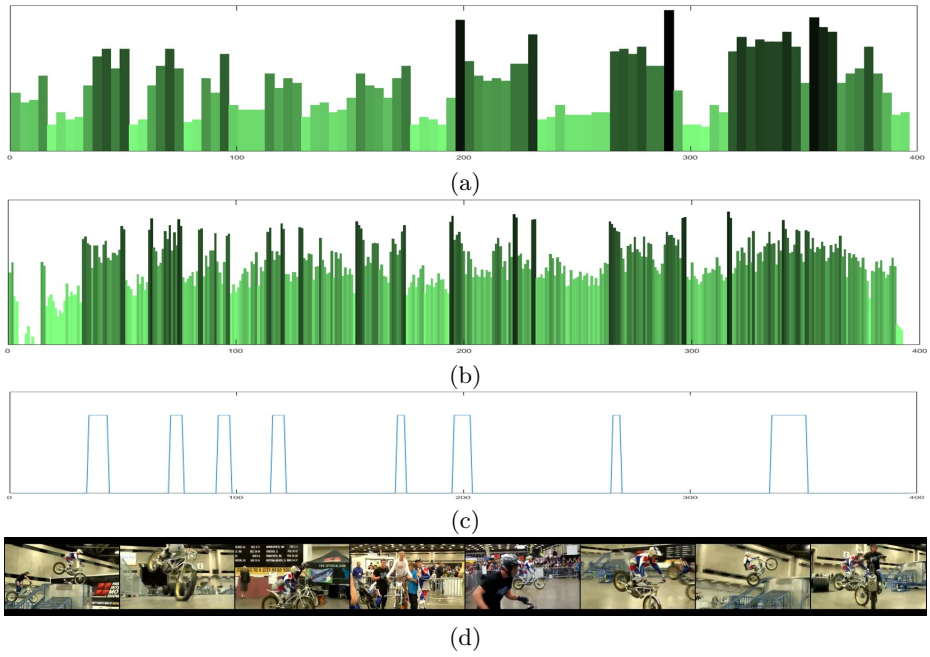

Fig. 4: The qualitative results of importance scores and selected frames. (a) Ground truth importance scores. (b) Predicted importance scores. (c) Post-processed prediction scores with the KTS algorithm. (d) Selected frames. The 42nd video of TVSum dataset is used in this figure.

4.6 Visualization

Here, 1) the frame-level importance scores (Fig. 4) and 2) the embedding matrix in the self-attention (Fig. 5) are visualized for better understanding of our approach. In Fig. 4, we provide **(a)** the ground truth scores, **(b)** predicted scores, **(c)** post-processed scores with KTS algorithm, and **(d)** the selected frames for summary. The frame-level scores are colored by their importance (i.e., the darker, the more important). Despite using unsupervised backbone [9], we can clearly see that our prediction scores well aligns with the ground-truth scores.

To see the actual effect of the proposed relative position embedding (RPE), we visualize the internal embedding matrix, $T \times T$, in the module. In Fig. 5, we show **(a)** self-attention embedding matrix without relative position, **(b)** row-wise softmax of (a) (i.e., SAE), **(c)** predicted importance scores, **(d)** relative position matrix, **(e)** self-attention embedding with relative position, and **(f)** row-wise softmax of (e) (i.e., RPE). As shown in (a) and (b), we see that the self-attention captures key-frames globally. Though, compared to (e) and (f), the difference between the informative and non-informative frames is small. This implies that the relative position information, (d), makes the embedding to be discriminative across different time steps. The effect of relative position becomes significant

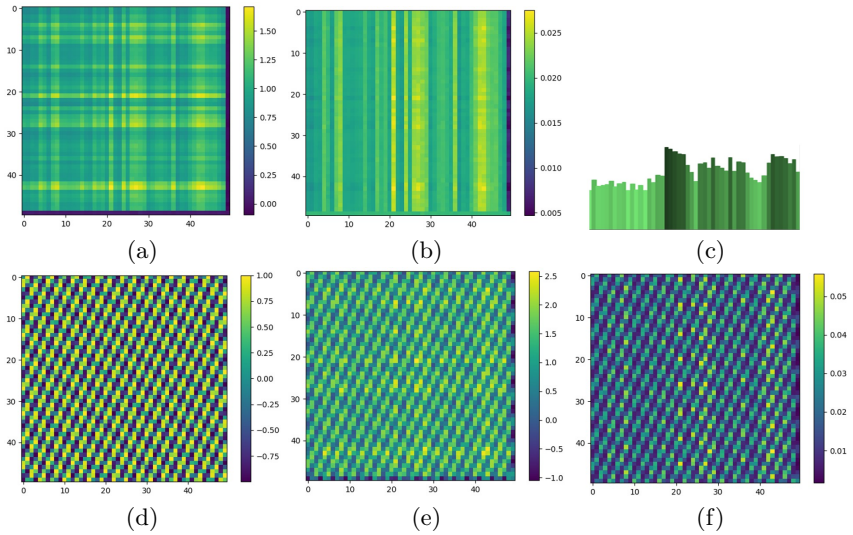

Fig. 5: The visualization of the embedding matrix in the self-attention module. (a) The self-attention embedding matrix without relative position. (b) The row-wise softmax of (a) (SAE). (c) The predicted importance scores. (d) The relative position information. (e) The self-attention embedding matrix with relative position (i.e., (a) + (d)). (f) The row-wise softmax of (e) (RPE).

after the row-wise softmax operation ((b) v.s. (f)). Note that the embedding matrix (f) well aligns with the final prediction scores, which means the model attempt to reflect the captured inter-frame relations in their predictions.

5 Conclusion

In this paper, we have explored the self-attention mechanism with relative position embedding for unsupervised video summarization. Self-attention makes handling long-term dependency among frames possible while relative position embedding provides sequential properties of the input video. We also use a global-and-local strategy to efficiently get the self-attention of a video that has a large and high dimensionality. We demonstrated the effectiveness of the proposed method through extensive ablation experiments. In terms of recently introduced rank order statistics-based evaluation metrics, our method obtains superior results over previous methods, even including supervised learning-based approaches. Also, we provide qualitative visualizations to illustrate that our method well highlights proper key segments in the video without any supervision. We hope many follow-up studies come up with our findings and results.

References

1. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q.V., Salakhutdinov, R.: Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860 (2019) 2, 3
2. De Avila, S.E.F., Lopes, A.P.B., da Luz Jr, A., de Albuquerque Araújo, A.: Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recognition Letters **32**(1), 56–68 (2011) 3
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 248–255. Ieee (2009) 8
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 2, 3, 4
5. Gong, B., Chao, W.L., Grauman, K., Sha, F.: Diverse sequential subset selection for supervised video summarization. In: Proc. of Neural Information Processing Systems (NeurIPS). pp. 2069–2077 (2014) 1
6. Gygli, M., Grabner, H., Riemenschneider, H., Van Gool, L.: Creating summaries from user videos. In: Proc. of European Conference on Computer Vision (ECCV). pp. 505–520. Springer (2014) 1, 3, 8, 12
7. Gygli, M., Grabner, H., Van Gool, L.: Video summarization by learning submodular mixtures of objectives. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 3090–3098 (2015) 1
8. Joshi, N., Kienzle, W., Toelle, M., Uyttendaele, M., Cohen, M.F.: Real-time hyperlapse creation via optimal frame selection. ACM Transactions on Graphics (TOG) **34**(4), 63 (2015) 1
9. Jung, Y., Cho, D., Kim, D., Woo, S., Kweon, I.S.: Discriminative feature learning for unsupervised video summarization. In: Proc. of Association for the Advancement of Artificial Intelligence (AAAI). vol. 33, pp. 8537–8544 (2019) 1, 2, 3, 4, 8, 9, 10, 11, 12, 13
10. Kang, H.W., Matsushita, Y., Tang, X., Chen, X.Q.: Space-time video montage. In: Proc. of Computer Vision and Pattern Recognition (CVPR). vol. 2, pp. 1331–1338. IEEE (2006) 1
11. Kendall, M.G.: The treatment of ties in ranking problems. Biometrika **33**(3), 239–251 (1945) 8, 9
12. Khosla, A., Hamid, R., Lin, C.J., Sundaresan, N.: Large-scale video summarization using web-image priors. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 2698–2705 (2013) 1, 3
13. Kim, G., Xing, E.P.: Reconstructing storyline graphs for image recommendation from web community photos. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 3882–3889 (2014) 1, 3
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proc. of International Conference on Learning Representations (ICLR) (2015) 8
15. Kopf, J., Cohen, M.F., Szeliski, R.: First-person hyper-lapse videos. ACM Transactions on Graphics (TOG) **33**(4), 78 (2014) 1
16. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 1346–1353. IEEE (2012) 1
17. Liu, D., Hua, G., Chen, T.: A hierarchical visual model for video object summarization. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **32**(12), 2178–2190 (2010) 1

18. Lu, Z., Grauman, K.: Story-driven summarization for egocentric video. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 2714–2721 (2013) 1, 3
19. Mahasseni, B., Lam, M., Todorovic, S.: Unsupervised video summarization with adversarial lstm networks. In: Proc. of Computer Vision and Pattern Recognition (CVPR). vol. 1 (2017) 1, 2, 3, 4, 8, 9, 11, 12
20. Ngo, C.W., Ma, Y.F., Zhang, H.J.: Automatic video summarization by graph modeling. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. pp. 104–109. IEEE (2003) 1, 3
21. Otani, M., Nakashima, Y., Rahtu, E., Heikkila, J.: Rethinking the evaluation of video summaries. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 7596–7604 (2019) 2, 8, 9, 12
22. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: Proc. of Neural Information Processing Systems Workshop (NIPS-W) (2017) 8
23. Poleg, Y., Halperin, T., Arora, C., Peleg, S.: Egosampling: Fast-forward and stereo for egocentric videos. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 4768–4776 (2015) 1
24. Potapov, D., Douze, M., Harchaoui, Z., Schmid, C.: Category-specific video summarization. In: Proc. of European Conference on Computer Vision (ECCV). pp. 540–555. Springer (2014) 1, 9
25. Pritch, Y., Rav-Acha, A., Peleg, S.: Nonchronological video synopsis and indexing. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **30**(11), 1971–1984 (2008) 1
26. Rochan, M., Wang, Y.: Video summarization by learning from unpaired data. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 7902–7911 (2019) 1, 9, 11
27. Rochan, M., Ye, L., Wang, Y.: Video summarization using fully convolutional sequence networks. In: Proc. of European Conference on Computer Vision (ECCV). pp. 347–363 (2018) 1, 9, 11
28. Sharghi, A., Laurel, J.S., Gong, B.: Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 2127–2136 (2017) 1
29. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. Proc. of North American Chapter of the Association for Computational Linguistics (2018) 2, 4, 5
30. Song, Y., Vallmitjana, J., Stent, A., Jaimes, A.: Tvsum: Summarizing web videos using titles. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 5179–5187 (2015) 3, 8, 10, 12
31. Sun, M., Farhadi, A., Taskar, B., Seitz, S.: Salient montages from unconstrained videos. In: Proc. of European Conference on Computer Vision (ECCV). pp. 472–488. Springer (2014) 1
32. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 1–9 (2015) 8
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proc. of Neural Information Processing Systems (NeurIPS). pp. 5998–6008 (2017) 2, 3, 4, 5
34. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 7794–7803 (2018) 3, 5
35. Wei, H., Ni, B., Yan, Y., Yu, H., Yang, X., Yao, C.: Video summarization via semantic attended networks. In: Proc. of Association for the Advancement of Artificial Intelligence (AAAI) (2018) 1

36. Yang, H., Wang, B., Lin, S., Wipf, D., Guo, M., Guo, B.: Unsupervised extraction of video highlights via robust recurrent auto-encoders. In: Proc. of International Conference on Computer Vision (ICCV). pp. 4633–4641 (2015) 1
37. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in neural information processing systems. pp. 5754–5764 (2019) 2, 3
38. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318 (2018) 3
39. Zhang, K., Chao, W.L., Sha, F., Grauman, K.: Summary transfer: Exemplar-based subset selection for video summarization. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 1059–1067 (2016) 1
40. Zhang, K., Chao, W.L., Sha, F., Grauman, K.: Video summarization with long short-term memory. In: Proc. of European Conference on Computer Vision (ECCV). pp. 766–782. Springer (2016) 1, 2, 3, 8, 9, 11, 12
41. Zhang, Y., Li, K., Li, K., Zhong, B., Fu, Y.: Residual non-local attention networks for image restoration. In: Proc. of International Conference on Learning Representations (ICLR) (2019) 3
42. Zhao, B., Li, X., Lu, X.: Hierarchical recurrent neural network for video summarization. In: Proc. of Multimedia Conference (MM). pp. 863–871. ACM (2017) 1, 2, 3
43. Zhao, B., Li, X., Lu, X.: Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 7405–7414 (2018) 1, 2, 3, 9, 11
44. Zhou, K., Qiao, Y.: Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In: Proc. of Association for the Advancement of Artificial Intelligence (AAAI) (2018) 1, 3, 9, 11, 12
45. Zwillinger, D., Kokoska, S.: CRC standard probability and statistics tables and formulae. Crc Press (1999) 8, 9