# Supplementary Material
# Global-and-Local Relative Position Embedding for Unsupervised Video Summarization

Anonymous ECCV submission

Paper ID 4920

## 1 Baseline Architecture

Firstly, we elaborate on the baseline architectures such as VAE-GAN and CSNet. The base networks generally sample input videos by 2 fps, and it becomes a sequence of $T$ images with a size of $H \times W$. Note that the $T$ is determined by the length of the video ranging from 200 frames to 1000 frames when using 2 fps. We extract the 1024 sized feature of every single image by GoogLeNet pre-trained on ImageNet. The features are stacked along the temporal dimension and become $T \times 1024$. Next, the feature is embedded to $T \times 256$ using a fully connected (FC) layer followed by the bidirectional LSTM with a hidden size of 256. The output features from both directions are concatenated to $T \times 512$ and reduced to $T \times 1$ using FC and sigmoid function to predict a score per image.

The VAE-GAN framework enables the network to learn a video summarization without ground truth labels. The main idea of unsupervised video summarization is generating original features of a video by giving summarized features through the reconstruction network. The discriminator then distinguishes whether it is from the original feature or the generated feature. If the generator performs enough well to deceive the discriminator, the selected features might have represented the video content successfully, and these frames are considered as keyframes.

The CSNet tackles with mode collapse of the unsupervised framework (i.e., VAE-GAN) and enables more discriminative feature learning. In addition to the feature reconstructing framework of VAE-GAN, they give a constraint that penalizes the low discrepancy of output scores by maximizing a variance of the scores. To ease the temporal dependency problem, they split the features into several temporal segments. Also, a temporal difference in the feature space is utilized to highlight fast scene changes.

## 2 In-depth Analysis of Network Design

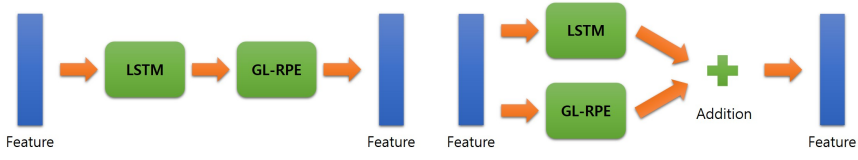
In this section, we present the experimental results which are detailed but less essential to write on the main paper. It can be regarded as a design choice for our global-and-local relative position embedding (GL-RPE).

| Internal size | Kendall's | Spearman's |
|---|---|---|
| 1 | 0.065 | 0.085 |
| 2 | 0.065 | 0.085 |
| 4 | 0.065 | 0.085 |
| 8 | 0.065 | 0.085 |
| 16 | 0.066 | 0.086 |
| 32 | 0.064 | 0.083 |
| 64 | **0.070** | **0.091** |
| 128 | 0.061 | 0.080 |
| 256 | 0.064 | 0.084 |

Table 1: The experiment by varying the feature dimension of RPE. The TVSum dataset is used in this table.

### 2.1 Relational matrix in self-attention embedding

We discover the optimal value of the feature dimension in the relative position embedding in this experiment. As shown in Table 1, the metrics using rank order statistics are measured by varying the feature dimension. The size of 64 shows the best result. Note that $RPE + GL_8$ is used in this experiment.




(a) (b)

Fig. 1: (a) The output of LSTM is passed through the GL-RPE in a serial manner. (b) The output of LSTM is added to the output of GL-RPE in a parallel manner.

### 2.2 Parallel relative position embedding

In the main paper, the feature obtained from long short-term memory (LSTM) is passed to our GL-RPE in a serial manner as shown in Fig. 1-(a). We also tested the parallel case of adding each output of LSTM and RPE as illustrated in Fig. 1-(b), and the corresponding results are shown in Table 2. The serial case shows a better result and is adopted to our main submission.

## 3 Visualization for Keyframes

In our main submission, only a few selected frames are visualized. Here, we present a dense visualization of the predicted summary. As shown in Fig. 2, the representative frames are well selected.