Video Salient Object Detection Using Multi-Scale Self-Attention

Jiahao Liu
CSCE,Graduate School of FSE
Waseda University
Tokyo, Japan
jiahao.liu@akane.waseda.jp

Haoyuan Liu
CSCE,Graduate School of FSE
Waseda University
Tokyo, Japan
liuhaoyuan@akane.waseda.jp

Hiroshi Watanabe
CSCE,Graduate School of FSE
Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

***Abstract*—How to effectively model both spatial information and temporal dynamics is crucial to Video Salient Object Detection (VSOD). Recently, there are some works using self-attention mechanism to capture the spatiotemporal information due to its ability of modeling long-range dependencies of patch tokens. However, these models designate similar receptive fields of the spatiotemporal feature maps, which limits the ability of the models in handling the frames with multiple salient objects of different scales. To address this issue, we propose a Multi-Scale Self-Attention (MSSA) operation to better model the spatiotemporal features of salient objects with different scales. The experimental results demonstrate that our method achieves better performance on challenge datasets by using MSSA operation.**

***Index Terms*—Video Processing, Salient Object Detection**

I. Introduction

As a fundamental task in video processing, Video salient object detection (VSOD) aims at locating and segmenting the most visually distinctive regions in a video clip. This task originates from the studies of human visual attention mechanisms, since human could quickly pay attention to the most informative parts of visual scenes.

Compared with the still image-based tasks, VSOD has an important difference, that is, the motion information between the adjacent frames. When an event that happens in the real world is condensed into seconds, the pixels in different frames could be temporally inconsistent on the time dimension. How to effectively take such dynamic information into consideration makes VSOD very challenging.

Most existing VSOD methods, which can be classified into 3D convolution based, ConvLSTM based and optical-flow based methods. However, using 3D convolution operation or serial structure can bring additional computation cost. The usage of prior optical flow information could make the network not be a real sense of end-to-end network. Some serial processing methods like ConvLSTM based method process the data step by step and may be hard to achieve real-time performance.

More recently, some works intend to use non-local attention-based mechanisms to better explore the pair wise relation in the area of adjacent frames. Fan et al.and Gu et al. [1] propose a saliency-aware-attention module and a Constrained Self-Attention (CSA) operation to better capture the motion cues, respectively. And Su et al [2] first introduce Transformer block to capture the long-range dependencies through self-attention mechanism.

Motivated by the above observations, in this paper, we propose a multi-scale self-attention guided video salient object detection network, which could not only capture the long-range dependencies between adjacent video frames but also improve the model's ability of handling images with multiple objects in different scales.

II. Related Works

A. Self-Attention Mechanism in Vision Field

Self-attention mechanism is proposed to capture long-range dependencies in machine translation. As a sequence model, it works by measuring pair-wise relationships of all patch tokens. Recently, Vision Transformer (ViT) [3] have shown that such self-attention mechanism also has superior performance in visual tasks. In Vision Transformer, an image is viewed as fixed-size patch tokens. Thus, we divide the input image into patch tokens and linearly embed each of them. Then the feature sequence will be fed into self-attention layer to do multi-head self-attention operation. After self-attention layer and LayerNorm, the feature sequence will be fed into FeedForward layer to do Multilayer Perceptron (MLP) operation.

B. Improvement of Fixed-scale Multi-head Self-attention

Prior Transformer models rely on fixed receptive fields of the patch tokens and uniform information granularity within one attention layer. In visual tasks such operations ignore the multi-scale nature of scene objects and could be incapable of extracting features at different scales simultaneously.

PVT (Pyramid Vision Transformer) [4] and Shunted Transformer [5] utilize a spatial reduction attention to merge tokens of key and value. Inspired by their work, we follow the Shunted Transformer to perform spatial reduction of the feature sequence to achieve our multi-scale self-attention.

In our work, since the input is N frames of a video clip, the shape of feature sequence has an additional dimension, number of frames. The self-attention mechanism is utilized to pay more attention to the salient areas and assign more weights to capture the global cues. In order to handle the frames with multiple salient foreground objects, we merge multiscale tokens within one attention layer.

979-8-3503-4017-4/23/$31.00 ©2023 IEEE

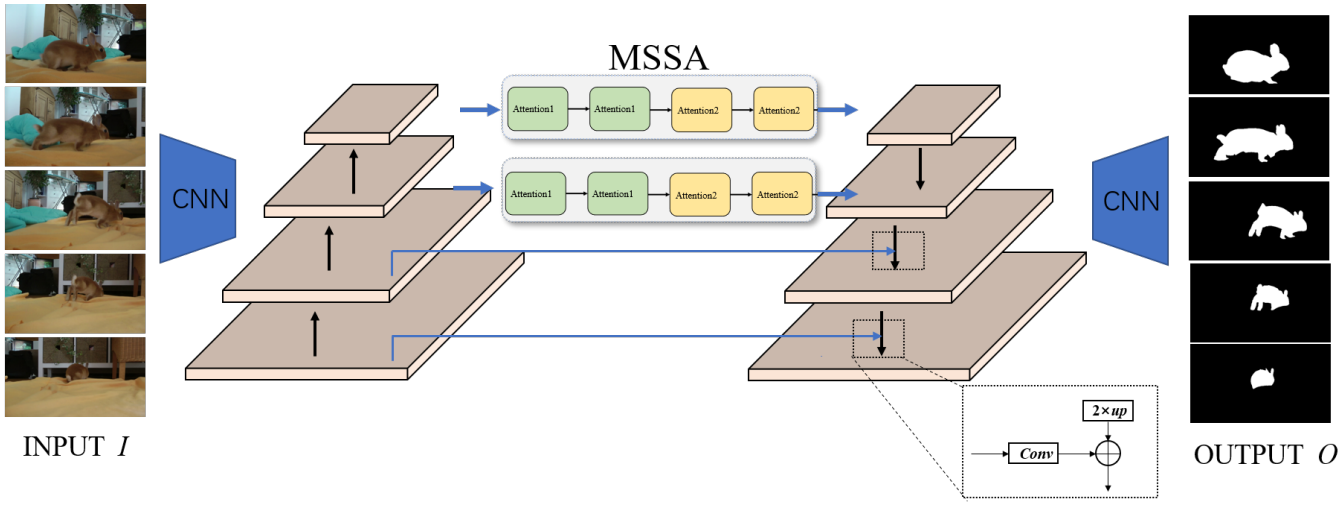

Fig. 1. Overview of our proposed method.

III. Methodology

A. Multi-scale Self-attention

Motivation of proposing multi-scale self-attention for this task:

In VSOD task, because the form of data is a continuous frame of video, the distance between the objects and the camera in the dataset determines their sizes in the frames. Thus, the size of the same saliency object in consecutive frames of the video is different, as shown in Fig.2. The different sizes of the same salient object will bring difficulties to such process. Therefore, we proposed multi-scale self-attention in order to make the model fuse more different scaled feature information.

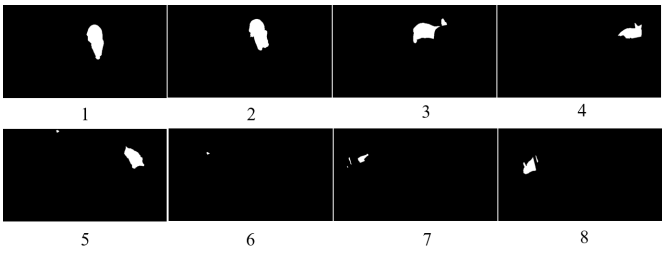

Fig. 2. Illustration of the same salient object with different sizes in consecutive frames.

In our work, since the input is N frames of a video clip, the shape of feature sequence has an additional dimension, that is, number of frames. In order to handle the frames with multiple salient foreground objects, we merge multiscale tokens within one attention layer. The input sequence $F \in R^{h \times w \times frames \times c}$ are firstly projected into query (Q), key (K), and value (V). Different from Shunted Transformer, we do not choose a series of different down-sampling ratios which is specially designed as backbone network. We only use one down-sampling ratio 2 which means reduction of the feature sequence's length by half. The illustration of our multi-scale self-attention operation is shown in Fig.3.

Integrating the keys with different scales within one attention layer enables the network to capture multi-granularity spatio-temporal features. Specifically, we split the attention process into two parts. The sizes of keys K and values V are different in each part for different heads indexed by *i*:

$$Q_i = XW_i^Q, \tag{1}$$

$$K_i = \textit{spatial reduced}(X, r_i)W_i^k, \tag{2}$$

$$V_i = \textit{spatial reduced}(X, r_i)W_i^V. \tag{3}$$

In this work, we use a 3D convolution layer with kernel size and stride of r_i to implement the spatial reduction operations. In practice, we set r_i to 2 or 1 to prevent incurred feature information loss since our feature maps have already down sampled by CNN encoder. The shunted self-attention is calculated by:

$$h_i = \textit{Softmax}(\frac{Q_i K_i^T}{\sqrt{d_h}})V_i. \tag{4}$$

B. Network Architecture

The overview of our proposed method is shown in Fig.1. Given an input of N frames of a video clip, they are firstly fed into CNN backbone to extract multi-scale spatial features. The encoder of our network is built upon VGG16 which I introduced in the related work part. The VGG blocks utilize small 3*3 convolution kernel to increase the number of channels while the pooling layers cut down height and width of feature maps. Such operation makes the encoder learn more deep features while the increase of the calculation amount continues to slow down.

We encapsulate different self-attention schemes into different transformer blocks. For this part of network, we call it MSSA. MSSA is designed to capture spatial temporal information of adjacent frames. They are utilized in the top

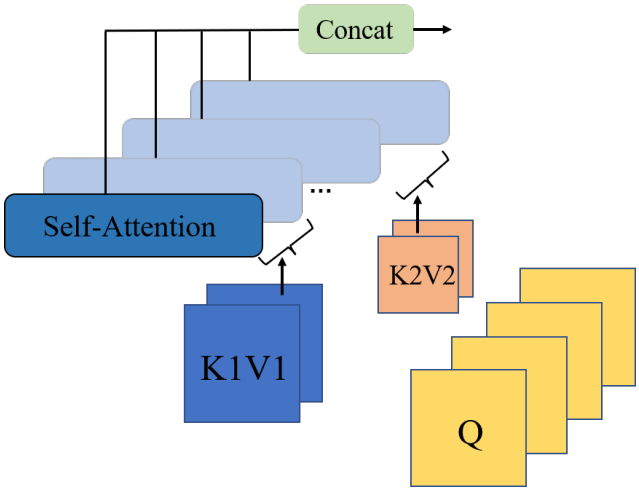

Fig. 3. Illustration of multi-scale self-attention.

two layers in the Fig.1, because the computational cost of the self-attention mechanism depends on the length of feature sequence. To avoid expensive memory consumption and computational cost, we use it on the spatial reduced feature layers to produce the multi-scale self-attention enhanced feature maps. For decoder, we reference the strategies of feature pyramid network. The MSSA enhanced spatiotemporal feature maps are upsampled using bilinear interpolation to match the low-level feature size. Then we concatenate the static spatial features with the spatiotemporal features. Finally, a convolution layer is utilized to give the predicted mask.

IV. Experiments

A. Loss Function

We use a weighted Binary Cross-Entropy loss and a IoU loss for pixel-wise segmentation. We denote the predicted probability as P (i, j) and denote the groundtruth as G(i, j). The ratio of all positive pixels to the total number of pixels in the image is calculated as γ.Then the weight Binary Cross-Entropy loss can be defined as:

$$L_{wbce} = -\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\gamma G(i,j)log(p(i,j)) - (1-\gamma) \\ (1-G(i,j))(log(1-p(i,j))), \tag{5}$$

where H and W denote the height and width of the image.

Besides we also use IoU loss to evaluate segmentation accuracy as follows:

$$L_{iou} = 1 - \frac{\sum_{i=1}^{H}\sum_{j=1}^{W}P(i,j)G(i,j)}{\sum_{i=1}^{H}\sum_{j=1}^{W}[P(i,j)+G(i,j)-P(i,j)G(i,j)]}, \tag{6}$$

The total loss of the framework is formulated as follows:

$$L_{total} = L_{wbce} + L_{iou}. \tag{7}$$

B. Training Scheme

In accordance with the procedure adopted by various studies, we initially conduct pre-training on a static image dataset, and then finetune the whole network on the video dataset. The input contains N = 5 frames, which has fixed size of 224 × 224 for training and testing. The numbers of multi-heads of two different Transformer blocks are set to 4 and 8 respectively, and the hidden dimensions of FeedForward layer are set to 782. All experiments are conducted on two GTX 1080Ti GPUs.

Pretrain phase. We firstly pre-trained the network on static image dataset COCO-SEG which contains 200,000 images and each image has pixel-wise annotations. The Adam optimizer is employed with an initial learning rate of 1e-5, which is reduced by half every 20,000 epochs. The pretraining process takes about 21 hours for 100,000 epochs.

Finetune phase. After pretraining, we finetune the network on the video datasets DAVIS16 and FBMS, which contains 59 video clips in total. We set the batchsize to 4 and use the same learning rate schedule as the pretrain phase. The data augmentation methods contain random rotation, random flip, random crop and color jitter, which are the common practice for data augmentation. The finetune phase takes about 12 hours for 100,000 epochs.

C. Datasets and Evaluation Metrics

We evaluate video salient object detection methods on DAVIS16, FBMS, and SegTrack-V2 benchmarks. DAVIS16 has 50 high-quality video sequences (30 for training and 20 for testing) which have pixel-wise manually created segmentation in the form of binary mask for each frame. SegTrack-V2 dataset is used to evaluate the model because all VSOD methods are not trained with any subsets of them.

Three main metrics are used to evaluate the VSOD method, including mean absolute error MAE [6], F-measure F_β, and Structural measurement (S-m). MAE measures the absolute pixel errors between the predicted mask and groundtruth:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|(G_i - S_i)|, \tag{8}$$

where N denotes the number of all pixels, S_i represents the predicted saliency map value and G_i represents the groud truth value. F-measure is computed as a weighted mean of precision and recall and it defined as follows:

$$F_\beta = \frac{(1+\beta^2)Precision * Recall}{\beta * Precision + Recall}. \tag{9}$$

S-measure assesses the structural similarity between the real-valued saliency map and the binary ground truth. It takes into consideration both object-aware (So) and region-aware (Sr) structural similarities:

$$S = \alpha * S_o + (1-\alpha) * S_r, \tag{10}$$

where α is set to 0.5.

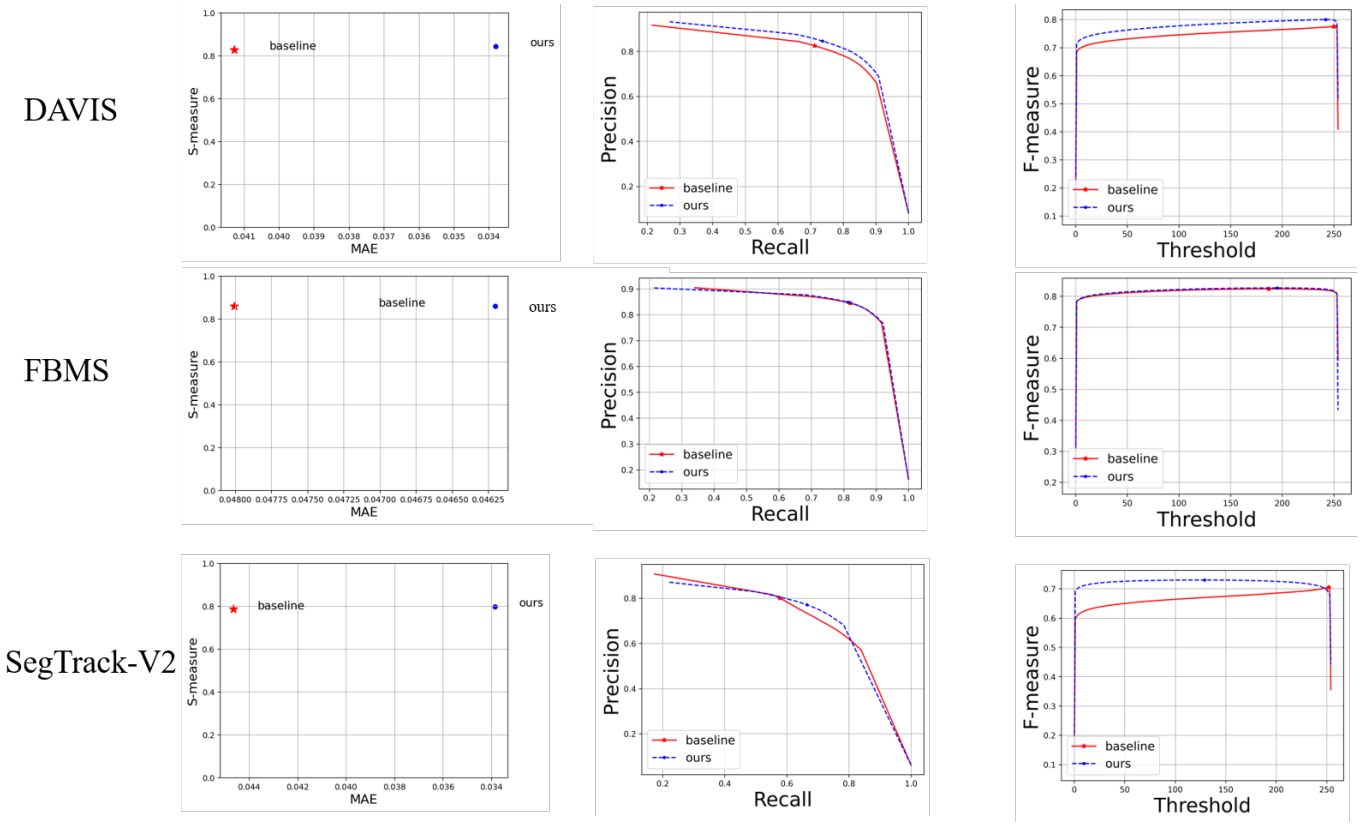

Fig. 4. Analysis of effectiveness of proposed MSSA method.

D. Experimental Results and Comparisons

We investigate the effect of the proposed MSSA branch by making comparisons with the original transformer blocks. The results can be seen in Tab.1, under the same training scheme, the MAE of the baseline with MSSA branch is lower than that with original Transformer blocks. From this table, we can see that our proposed MSSA gives the better performance than that with original Transformer blocks, which shows the effectiveness of MSSA branch in VSOD task.

TABLE I
ANALYSIS OF EFFECTIVENESS OF MSSA BRANCH

Method	DAVIS		SegTrack-V2	
	MAE↓	S-m↑	MAE↓	S-m↑
VGG16+Transformer block	0.041	0.828	0.041	0.788
VGG16+MSSA(ours)	**0.033**	**0.845**	**0.034**	**0.799**

In Fig.4, we compare different metrics on the three main datasets. The first figure gives the comparisons with S-measure and MAE, and our proposed MSSA gives the better performance on these datasets. The Precision-Recall Curves of our proposed method also higher than that with original Transformer blocks. F-measure curves also give the same performance, which proves that our proposed MSSA is effective.

V. CONCLUSION

In this paper, we propose a multi-scale self-attention (MSSA) module for VSOD, which could not only capture the spatiotemporal information but also make the network effectively model the salient objects with different scales. Experiments show that our MSSA achieve better performance than the original self-attention operation in ViT for VSOD task, and the test results on VSOD benchmarks also demonstrate the effectiveness of our method.

REFERENCES

[1] Y. Gu, L. Wang, Z. Wang, Y. Liu, M.-M. Cheng, and S.-P. Lu, "Pyramid Constrained Self-Attention Network for Fast Video Salient Object Detection", Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 10869-10876, Apr. 2020.

[2] Y. Su, J. Deng, R. Sun, G. Lin, H. Su and Q. Wu, "A Unified Transformer Framework for Group-based Segmentation: Co-Segmentation, Co-Saliency Detection and Video Salient Object Detection," in IEEE Transactions on Multimedia, pp.1-13, Apr.2023

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. "An image is worth 16x16 words: Transformers for image recognition at scale," International Conference on Learning Representations (ICLR), May. 2021.

[4] W. Wang, E. Xie, X. Li, D. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," IEEE International Conference on Computer Vision (ICCV), pp.568-578, Oct.2021.

[5] S.Ren, D.Zhou, S. He, J. Feng, X. Wang. "Shunted Self-Attention via Multi-Scale Token Aggregation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10853-10862, Jun.2022.

[6] F. Perazzi, P. Krahenb ¨ uhl, Y. Pritch, and A. Hornung, "Saliency filters: " Contrast based filtering for salient region detection," in 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 733–740. 6, Jun.2012.