

# Interactive Image Segmentation with Temporal Information Augmented –Supplementary Material–

Qiaoqiao Wei  
wqq18@mails.tsinghua.edu.cn

Hui Zhang  
huizhang@tsinghua.edu.cn

Jun-Hai Yong  
yongjh@tsinghua.edu.cn

School of Software  
BNRist  
Tsinghua University  
Beijing, China

## 1 Complementary Ablation Study Results

We provide supplemental ablation study results using ViT (mainly ViT-B) as backbone. Table 4 shows the segmentation results achieved by adding individual modules proposed in this paper. The results demonstrate that each module contributed to the improvement in segmentation quality. The MI and DA modules played a particularly crucial role in enhancing segmentation stability. Table 5 presents the results of our ablation experiments on the hyperparameter  $K$ . The results show that generating object descriptors for background patches is more effective than generating a single object descriptor for the entire background of the image. Table 6 presents the results obtained by setting different capacities for the memory bank. It can be observed that the best performance is achieved when the memory bank capacity is set to 2-4.

| Backbone | IP | MI | DA | NoC@90% | IoU@5(%) | BloU@5(%) | ASSD@5 | NoDC <sub>20</sub> | mDA(%) | CME(%) |
|----------|----|----|----|---------|----------|-----------|--------|--------------------|--------|--------|
| ViT-B    | ✓  |    |    | 4.95    | 89.04    | 81.03     | 7.72   | 1625               | 0.69   | 8.04   |
| ViT-B    |    | ✓  |    | 4.85    | 89.27    | 81.42     | 7.48   | 1589               | 0.52   | 7.69   |
| ViT-B    |    |    | ✓  | 4.89    | 89.10    | 81.21     | 7.80   | 1612               | 0.55   | 7.83   |
| ViT-L    | ✓  |    |    | 4.74    | 89.51    | 83.57     | 6.38   | 1542               | 0.49   | 8.28   |
| ViT-L    |    | ✓  |    | 4.69    | 90.23    | 83.89     | 6.42   | 1535               | 0.40   | 7.68   |
| ViT-L    |    |    | ✓  | 4.72    | 90.35    | 83.62     | 6.44   | 1524               | 0.38   | 7.34   |

Table 4: Supplementary ablation study results for the core components on the DAVIS dataset.

## 2 Visualization of Object Descriptors

To better demonstrate the interpretability of the object descriptors created in the MI branch, we present a series of heatmaps in Figure 6. These heatmaps illustrate the similarity of each pixel of the the MI module’s input features to  $1 + K$  object descriptors generated in the

| $K$     | Berkeley       |                |                 |                 | DAVIS          |                |                 |                 |
|---------|----------------|----------------|-----------------|-----------------|----------------|----------------|-----------------|-----------------|
|         | 1<br>(= 1 × 1) | 4<br>(= 2 × 2) | 16<br>(= 4 × 4) | 64<br>(= 8 × 8) | 1<br>(= 1 × 1) | 4<br>(= 2 × 2) | 16<br>(= 4 × 4) | 64<br>(= 8 × 8) |
| NoC@85% | 1.37           | 1.39           | <b>1.35</b>     | 1.36            | 3.62           | 3.56           | <b>3.54</b>     | 3.55            |
| NoC@90% | 1.69           | 1.66           | <b>1.59</b>     | 1.62            | 4.97           | 4.82           | <b>4.78</b>     | 4.79            |

Table 5: An ablation study for the number of background object descriptors.

| Capacity | Berkeley |             |             |             |      | DAVIS |      |             |      |      |
|----------|----------|-------------|-------------|-------------|------|-------|------|-------------|------|------|
|          | 1        | 2           | 3           | 4           | 5    | 1     | 2    | 3           | 4    | 5    |
| NoC@85%  | 1.37     | 1.36        | <b>1.35</b> | <b>1.35</b> | 1.36 | 3.62  | 3.56 | <b>3.54</b> | 3.55 | 5.21 |
| NoC@90%  | 1.60     | <b>1.57</b> | 1.59        | 1.65        | 1.73 | 4.91  | 4.83 | <b>4.78</b> | 4.79 | 4.85 |

Table 6: An ablation study for the capacity of the memory bank.

previous interaction round, where warmer colors indicate higher similarity and cooler colors indicate lower similarity.  $K$  is assigned to 16 in this paper. Let the features be denoted as  $\mathbf{F} \in \mathbb{R}^{C_D \times H \times W}$ , where  $C_D$ ,  $H$ , and  $W$  denotes the number of channels, height, and width of  $\mathbf{F}$ , and the object descriptors as  $\mathbf{D} \in \mathbb{R}^{(1+K) \times C_D}$ . The similarity maps  $\mathbf{S}_{vis} \in \mathbb{R}^{(1+K) \times H \times W}$  are obtained following this equation:

$$\mathbf{S}_{vis} = \text{softmax}(\mathbf{D} \otimes \mathbf{F}, \dim = 0). \quad (8)$$

We visualize the  $(1 + K)$  similarity maps using colormaps. From these heatmaps, it can be observed that the object descriptors generated in this paper can significantly distinguish between foreground and background in the image.

### 3 Visualization of Interactive Image Segmentation Process

We provide more segmentation results to visualize the interactive segmentation process. The results from Figure 7 indicate that our method achieves superior segmentation results. Even in challenging scenarios such as similar texture between foreground and background, occlusion of the foreground, and poor lighting conditions, our method can produce satisfactory segmentation results with a minimal number of annotated clicks.

### 4 Failure Cases Analysis

In this section, we discuss the possible limitations of our method. As shown in Figure 8, our method performs poorly in some extreme cases. Specifically, for instances whose structures are complex, such as bicycle wheels and tangled boat masts, our method struggles to accurately capture the structure details. The method also fails to correctly infer the desired segmentation granularity when faced with ambiguity. Additionally, our method encounters difficulties in identifying fine details in the structure of small foreground objects, such as the kite strings in kitesurfing. These challenges are also present in other interactive image segmentation methods, constituting inherent difficulties in this field.

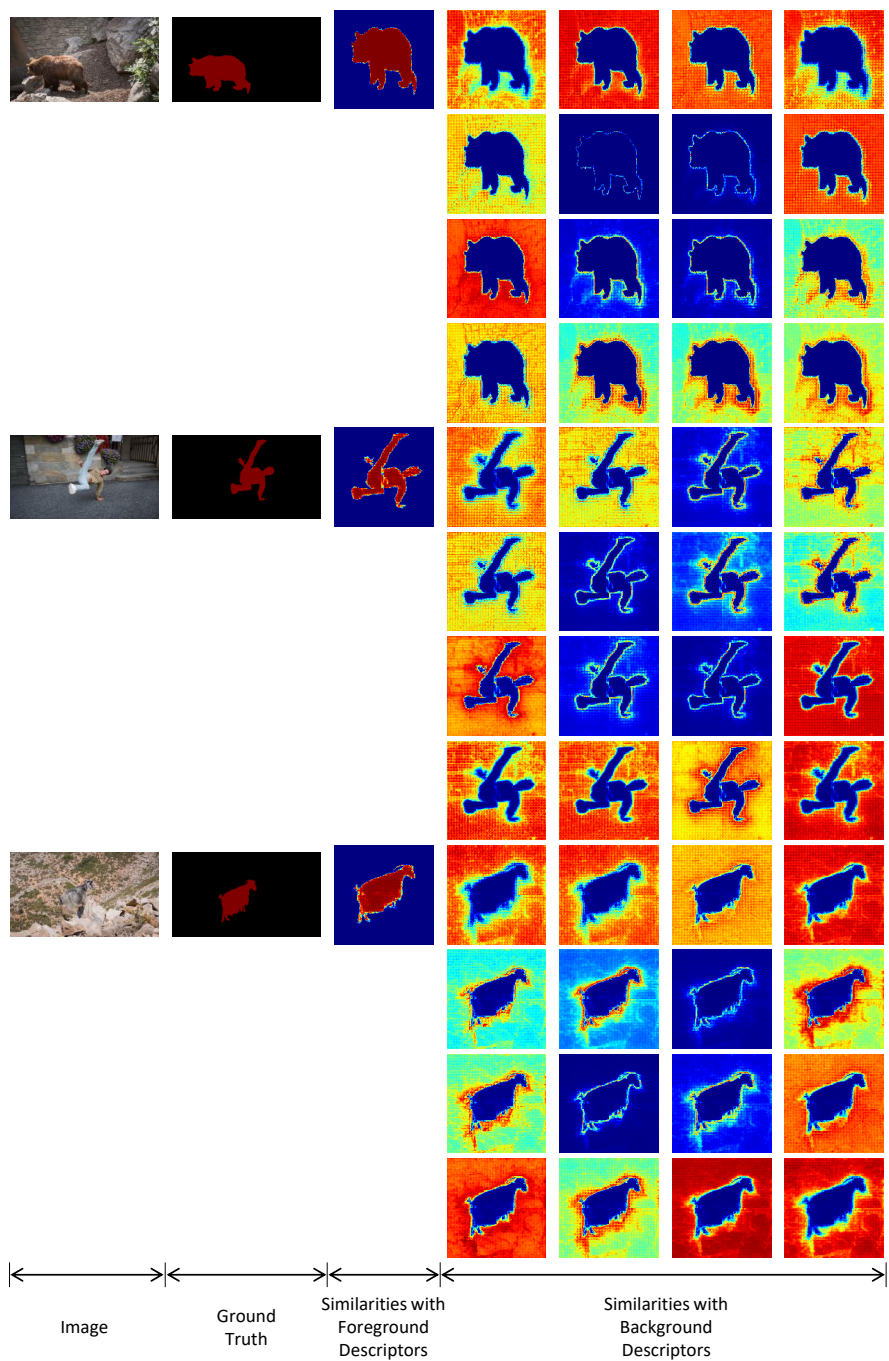


Figure 6: Visualization of the similarities between the input features and the object descriptors generated in the previous interaction round.

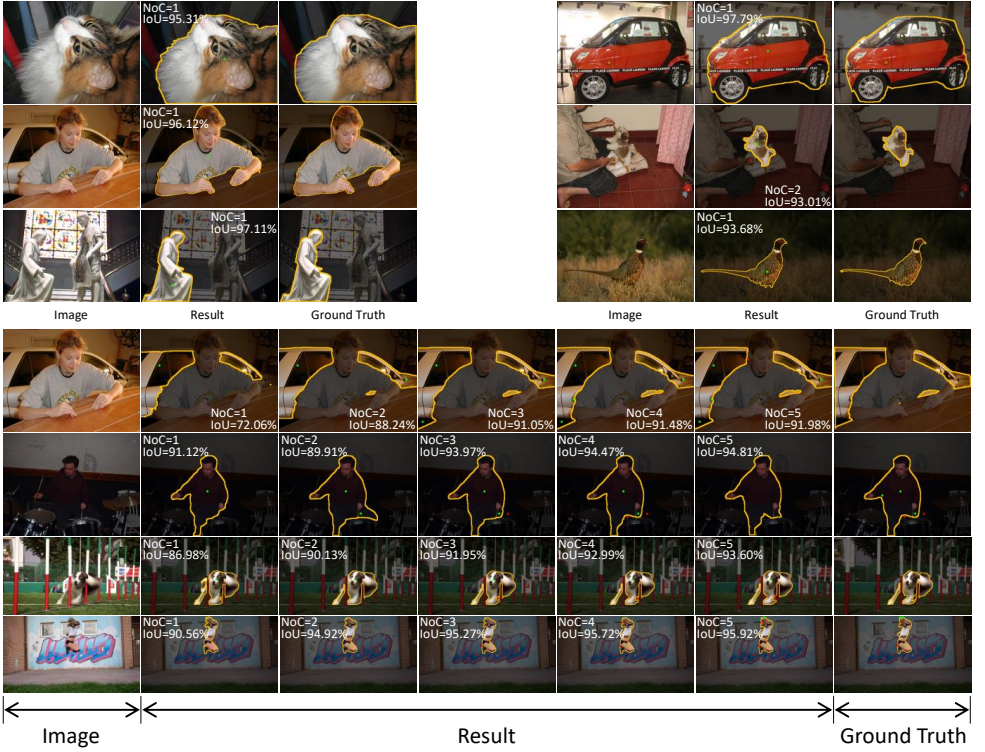


Figure 7: Visualization of the interactive image segmentation process. The green clicks denote foreground clicks, and the red ones denote background clicks.



Figure 8: Possible limitations of the proposed method. The green clicks denote foreground clicks, and the red ones denote background clicks.