

Supplementary Material for Tracking-forced Referring Video Object Segmentation

Anonymous Author(s)*

1 VIDEO ALIGNMENT DETAILS

To intuitively reflect the similarity score distribution of each frame in videos, we conduct an alignment on the validation set of Red-Youtube-VOS. Firstly, we manipulate the normalized similarity scores whose range belongs to $[0,1]$ so as to widen the gap between them. Concretely, Given the normalized similarity scores of a video with T frames $S_n = \{S_i\}_{i=1}^T$, we utilize the power function for obtaining the processed S :

$$S = \{S_i^\lambda | S_i \in S_n\}_{i=1}^T \quad (1)$$

where the value of λ is set as 0.7 to ensure the reasonableness between the scores, it has no impact on the original distribution of the scores. Further, since different video has a different number of frames, we project frames to various stages of videos to prepare for the subsequent alignment. Specifically, we consider the length of a video to be 1, and divide it equally into ten stages such as $(0,0.1]$, $(0.1,0.2]$, and so on. we project the scores to each stage to obtain $S' = \{S'_j\}_{j=0}^{0.9}$, where S'_j is calculated as follows:

$$S'_j = \sum_{i=1}^T S_i, [i/T * 10] / 10 - 1 = j \quad (2)$$

where j represents the end value of each stage. Via this process, the similarity scores of a video are represented by 10 points: where x ranges $[0, 0.9]$ with the step is 0.1, and y is denoted by the stage scores. Next, we calculate the position of the key frame based on its index σ :

$$pos_{key} = \sigma / T \quad (3)$$

we pan the video by $-pos_{key}$ to obtain the relative location of each stage as the final score distribution of a video:

$$\begin{cases} X = [0 - pos_{key} : 0.9 - pos_{key} : 0.1] \\ Y = S' \end{cases} \quad (4)$$

then we use linear interpolation on X and Y to represent the distribution of scores of the video. By moving the key frames of each video to the origin, the alignment of all videos is achieved as shown as in Fig.3(b) in our paper. Besides, we calculate the average value of each stage and mark it with a red line.

2 TRAINING EFFICIENCY

In this section, we conduct quantitative analysis on training efficiency of ReferFormer[2], OnlineRefer[1] and our TF² with ResNet-50 backbone on the training set containing 3,471 videos of the Ref-Youtube-VOS dataset. We make comparisons on their training method, processing steps, time for an epoch, and $\mathcal{J}\&\mathcal{F}$ scores. The results are shown in Fig.1.

We first introduce the implementation details and parameter settings. For clip-level ReferFormer, we follow the settings in its source code to set the clip length to 5, and each clip should contain at least one frame in which the target object is present. i.e. the target mask is not empty. During data loading, if the ground-truth masks

of a clip are all empty, the clip will be reselected. The purpose of this is to ensure the mask loss of each clip is not null. For frame-level OnlineRefer, we also follow its original setup that inputs three frames to the model for online training. To improve the training stability, the number of input frames is set to 2 before the 4th epoch. For our TF², each training unit just contains one frame.

We conduct experiments with 4 NVIDIA RTX A6000. Due to the resource constraint, the maximum batch size of ReferFormer can be set to 1, i.e. one clip. The maximum batch size of OnlineRefer is set to 1, i.e. 2 frames for the first 3th epochs and 3 frames for the last 3th epochs. We take the average value 2.5 for subsequent calculation. Our TF²'s maximum batch size under our resource condition is set to 8 i.e. 8 frames. To make a fair comparison on training steps of each method, we use the least common multiple 40 of 5, 2.5, 8 as the number of frames to be processed for the calculation. Given 40 frames, the processing steps required by clip-level ReferFormer are 8 (40/5). For frame-level OnlineRefer, although these frames can be all input to the model with 16(40/2.5) times, the processing approach of videos is frame by frame, so the processing steps are 40. While our TF² just needs 5 (40/8) processing steps.

Besides, we show the time required to train an epoch for each method. In conjunction with $\mathcal{J}\&\mathcal{F}$ scores, our TF² not only achieves high efficiency through the parallel training way and the training unit just containing one frame, but also achieves the best performance.

Table 1: Comparison of training efficiency

Methods	Parall	Processing Steps	Time/Epoch(s)	$\mathcal{J}\&\mathcal{F}(\%)$
ReferFormer	✓	8	12,060	55.6
OnlineRefer	×	40	20,160	57.3
TF ² (Ours)	✓	5	8,800	64.6

3 CASES

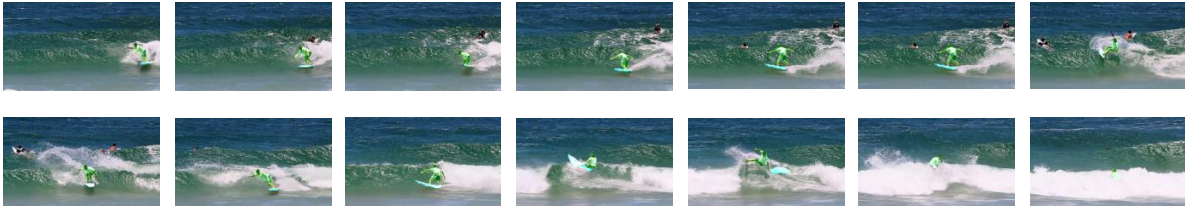
To further verify the effectiveness of our method, We show more cases on the validation set of the Ref-Youtube-VOS dataset with ResNet-50 backbone. These cases source from the supplementary file "submission.zip" in our submitted materials, and this $\mathcal{J}\&\mathcal{F}$ score, which is the results of TF² in Fig.1, is evaluated by the online server¹.

REFERENCES

- [1] Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. 2023. OnlineRefer: A Simple Online Baseline for Referring Video Object Segmentation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 2749–2758.
- [2] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. 2022. Language as Queries for Referring Video Object Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 4964–4974.

¹<https://codalab.lisn.upsaclay.fr/competitions/13520>

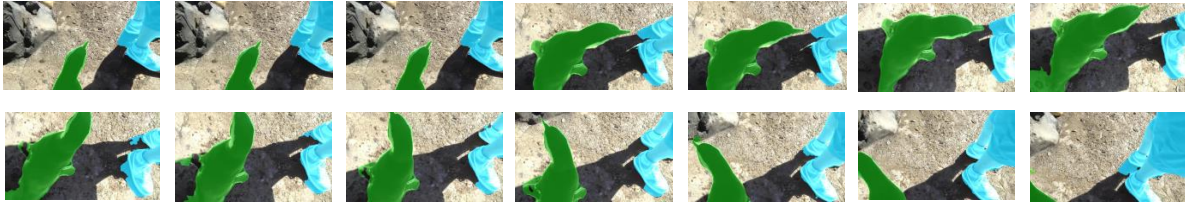
a person wearing a white shirt is in the ocean on a surfboard riding a wave
a white surfboard is being rode by a person in the ocean wearing a white shirt



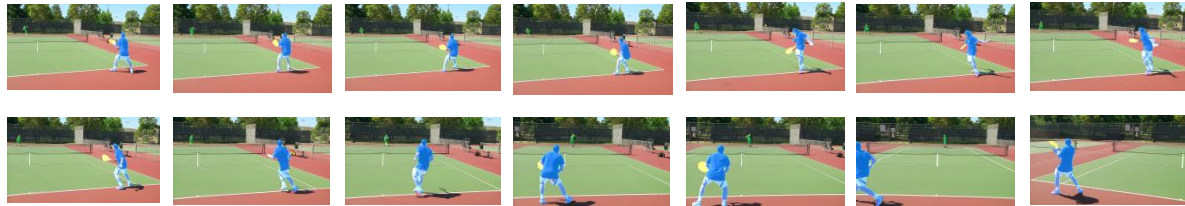
a person wearing white shorts and a red shirt is on the opposite side of the tennis court
a person in a blue shirt and black shorts
a tennis racket in the hand of the man in blue



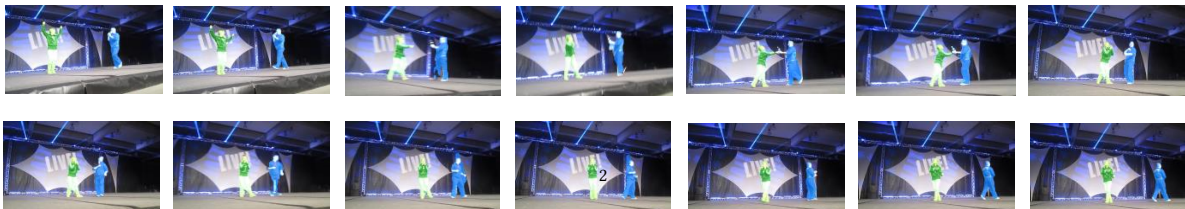
a penguin standing behind a person wearing grey nike shoes
a person wearing gym shoes standing to the right of a penguin



a person on the far side of a tennis court serving a tennis ball
a man in a blue shirt and white shorts
a tennis racket being used by a man wearing red and white shoes



a person dressed in black and white in the center of stage holding a microphone
a person dressed in all black on the right walking toward a person on stage while holding a microphone



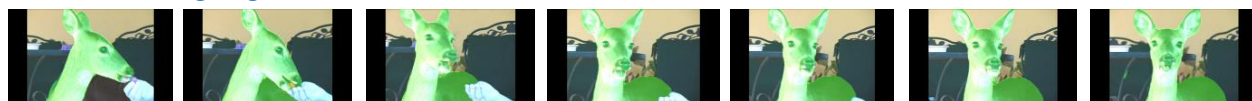
a black hat is being worn by a person riding a horse on the green grass
the tan and white horse is walking in the grass



a white truck driving on a road
a person wearing a white shirt is driving a white truck moving down the road



a brown deer has a hand giving it a red treat and it looks forward chewing
a human hand is giving the brown deer a red treat to eat



a cow on the leftmost side of the view
a black and white cow second from the left



a person dressed in white and gray holding a tennis racket
a tennis racket on the left being held by a person



the brown bear is standing behind a green container and a large rock
a green bucket is behind a brown bear on the grass

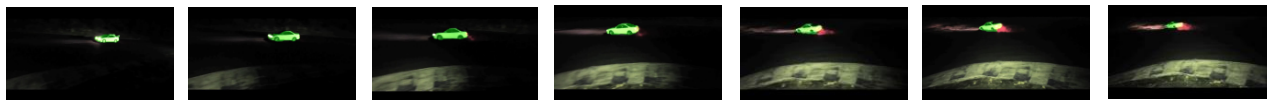


a paddle being used by a man in a green boat
a green boat is carrying a person in the water



a duck is held by a person with her both hands
a person is holding a duck and showing a duck photo



a boat is moving from the right and advancing to the shore**a white sedan moving leftwards through with its headlights on****a brown and white cow walking away from two other cows****a green train on tracks****a small kangaroo standing in the grass****a white boat moving through the water****the black truck with red/white and blue is moving down the road to the right with a crowd behind it****a person is operating the disco musical equipment****a person skiing on a rail**