1 Visualization

We show more target-relevant attention maps on the *search* images in Fig. 1. For 'Base', we replace the backbone of STARK-S with ViT-B/16. 'Ours' adopts the SimTrack framework with ViT-B/16 as the backbone. Both 'Base' and 'Ours' are trained with the same training setting as shown in the paper. While target-relevant attention map can be obtained for 'Ours' directly in the transformer backbone, 'Base' does not have such information since *search* and *exemplar* are processed separately. To obtain the target-relevant attention map for 'Base' model, we get the *exemplar* and *search* features from the l_{th} transformer layer after training and calculate the *search* attention weight $A(s^l)$ through (refer to Equ.(6) in the paper),

$$A(s^l) = softmax\left(\left[a(s^l, e^l), a(s^l, s^l)\right]\right), \tag{1}$$

where $s^l \in \mathbb{R}^{N_x \times D}$, $e^l \in \mathbb{R}^{N_z \times D}$, $A(s^l) \in \mathbb{R}^{N_x \times (N_z + N_x)}$. We select the target-relevant part from $A(s^l) \in \mathbb{R}^{N_x \times N_z}$ and average it along the second dimension to get $A^*(s^l) \in \mathbb{R}^{N_x \times 1}$. Then, we reshape $A^*(s^l)$ to $\frac{H_x}{s} \times \frac{W_x}{s}$ and up-sample it to the same size $(H_x \times W_x)$ with the *search* image. After that, we get the target-relevant attention maps as shown in Fig. 1. As we can see in Fig. 1, 'Ours' can quickly and gradually focus on a more accurate and comprehensive target area because the vital information interaction in the backbone enables the *search* feature learning to 'sense' the designated target.

2 Training Details

The whole training needs 500 epochs with 6×10^4 image pairs in each epoch. The training batch size is 256. All models are optimized with AdamW and the weight decay is 10^{-4}. The initial learning rates of the backbone and head are 10^{-5} and 10^{-4}, which will drop by a factor of 10 after 400 epochs. The loss weights λ_{iou} and λ_{L_1} are 2 and 5 in Equ.(3). For Sim-B/32, we shift the *exemplar* image by 16 pixels (half of the patch size 32) and crop a 64×64 foveal image in the centre of the shifted image. For Sim-B/16, we directly crop a 64×64 foveal image in the centre of the *exemplar* image. For Sim-L/14, to reduce computation cost, the input *exemplar* size is reduced to 84×84. We centre crop a 42×42 image as the foveal image, where the partitioning lines are located in the centre of those on the *exemplar* image.

3 Input Resolution

In the paper, we set the input size of *search* image as 224×224 to be consistent with existing vision transformers. We also evaluate the model performance when we increase the input resolution to 320×320 (the same with STARK-S) and 384×384. The results on LaSOT and TNL2K are shown in Table 1. A higher input resolution helps improve tracking accuracy.

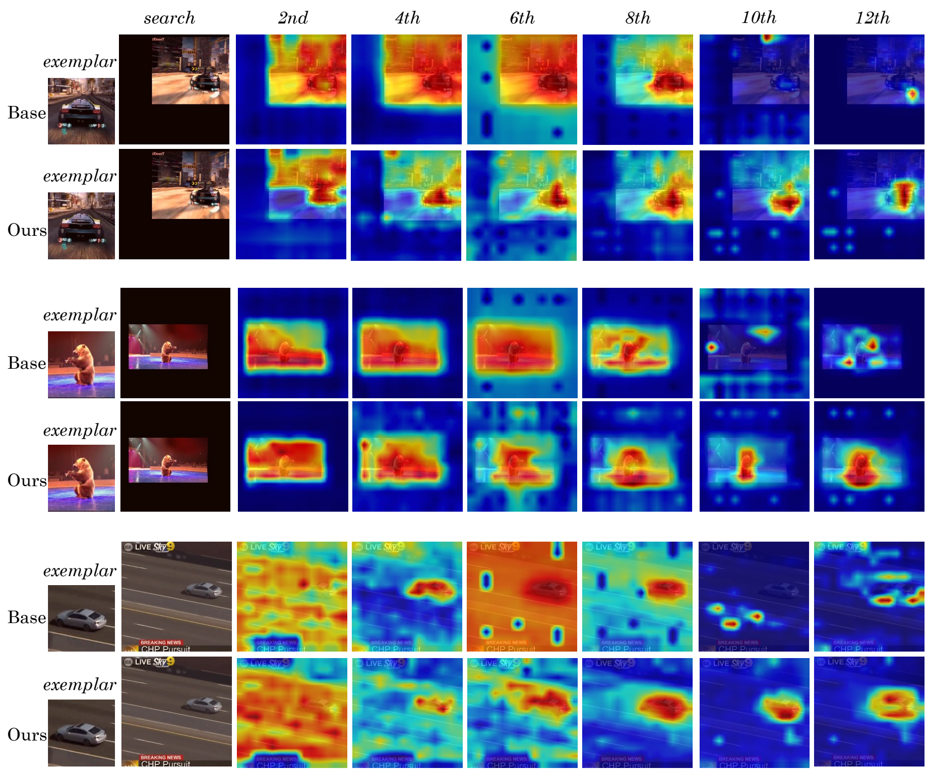

Fig. 1: The images in different columns are the *exemplar* image, the *search* images, the target-relevant attention maps from the $2nd, 4th, 6th, 8th, 10th, 12th(last)$ layer of the transformer backbone. 'Base' denotes the baseline model. 'Ours' is our SimTrack. 'Ours' can quickly and gradually focus on a more accurate and comprehensive target area.

#Num	Input Size	LaSOT			TNL2K	
		AUC↑	P_{norm}↑	P↑	AUC↑	P↑
①	224 × 224	69.3	78.5	74.0	54.8	53.8
②	320 × 320	70.0	79.2	74.8	54.8	54.2
③	384 × 384	70.4	79.3	75.0	55.2	55.2

Table 1: The performance of SimTrack (with ViT-B/16 as backbone) with diverse input sizes. A higher input resolution helps improve tracking accuracy.