Delving Deep into Many-to-many Attention for Few-shot Video Object Segmentation – Supplementary Materials –

Haoxin Chen[1], Hanjie Wu[1], Nanxuan Zhao[2], Sucheng Ren[1], Shengfeng He[1*]
[1] School of Computer Science and Engineering, South China University of Technology
[2] The Chinese University of Hong Kong

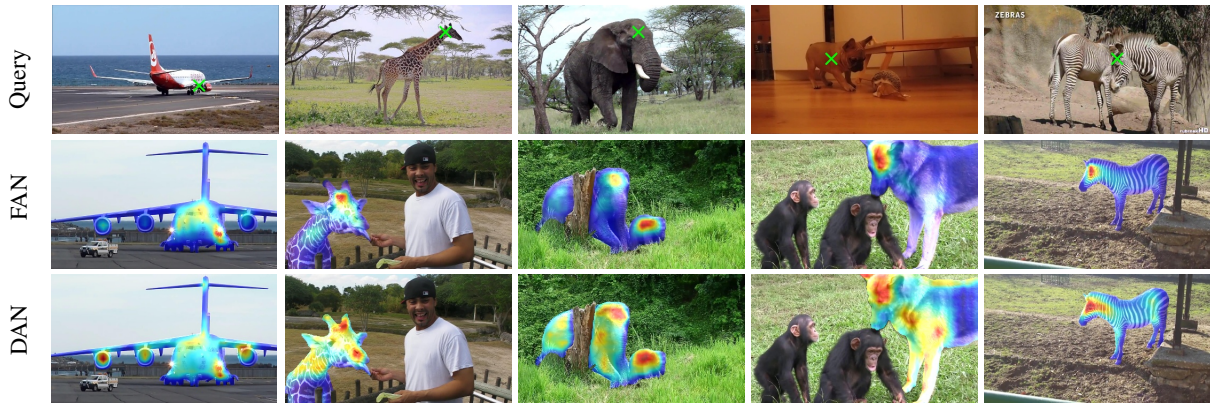

Figure 1. Visualization of attention maps. We compare our proposed domain agent attention network (DAN) with conventional full-rank attention (FAN). We use a green cross to denote query positions for obtaining the attention maps in the top row. And we show the heatmaps of the attention on support images in the bottom rows. We color high response with red color and low response with blue color.

1. Visualization of Attention Maps

Our domain agent attention module adds channel-wise attention into the query features. By building the correlation among channels, the model can better propagate the information from the support set to the query frames. We thus visualize the attention maps in Fig. 1, compared with traditional full-rank attention. Our attention module can associate the relevant regions in a non-local manner, under a broader range. Take the second set of images as an example, our results can associate horns and necks of the giraffes on the support images to the query position for guiding segmentation, while FAN only focuses on small local regions.

2. The Effect of Query Frames Number

We have shown that the computational cost of our DAN achieves a linear growth as the number of both support and query images increases. One may raise a question that is the efficiency an outcome of sacrificing the quality of results?

Query frames		Fold-1	Fold-2	Fold-3	Fold-4	Mean
$\mathcal{F}$	5	40.3	62.3	60.2	59.4	55.6
	10	40.3	62.4	59.6	58.6	55.2
	20	40.1	62.0	59.9	58.9	55.2
	40	40.0	62.5	59.5	59.1	55.3
$\mathcal{J}$	5	41.5	64.8	61.3	61.4	57.2
	10	41.3	64.8	60.0	60.7	56.7
	20	41.2	64.3	60.8	60.9	56.8
	40	41.1	64.8	60.3	61.4	56.9

Table 1. The effect of the number of query frames during the testing phase. The number of query frames only influences the quality of results slightly, while the computational cost growing linearly.

We thus test on a different number of query frames in the testing phase with a range between 5 to 40, and we train the model with 5 query images. As shown in Tab. 1, increasing the number of query frames only influence the quality of results slightly. The accuracy of both edges and regions is only reduced by 0.3% while the number of query frames increasing from 5 to 40. This indicates that our model can

*Corresponding author (hesfe@scut.edu.cn).

balance well between the computational efficiency and the quality of results, and able to process longer video robustly.

3. The Effect of Support Images Number

In this experiment, we investigate how the number of support images influences the model's performance. We test on 10 categories containing the largest number of videos in the dataset as it is easier to select more support images. We also use online learning to enhance feature representation with more support images. As shown in Tab. 2, increasing the number from 5 to 10 enhances the performance of all the tested categories. Besides, adding more support images has a huger effect on the category with relatively lower performance.

Support frames	$\mathcal{F}$		$\mathcal{J}$	
	5-shot	10-shot	5-shot	10-shot
Deer	63.6	63.9	66.6	67.3
Duck	61.4	62.4	58.8	59.0
Giant panda	63.4	65.7	72.4	74.8
Motorbike	43.5	44.2	51.6	52.9
Person	36.2	48.2	29.4	42.8
Skateboard	36.8	39.7	16.3	16.9
Snake	43.7	46.2	47.3	51.8
Snowboard	23.2	34.3	5.4	11.8
Train	23.4	24.6	38.4	42.3
Truck	44.3	49.3	62.8	67.1
Mean	43.9	**47.9**	44.9	**48.7**

Table 2. The effect of the number of support frames. The performance enhances for all the tested categories when increasing the number of support images from 5 to 10.

4. The Effect of Domain Agent Position

In the main paper, we use the middle frame of the query video as the agent. To study the effect of the position of the agent, we sample the agent at different positions: beginning (the first one), middle, and end (the last one). To manifest the performance gap, we test on 40 query frames.

Agent position		Fold-1	Fold-2	Fold-3	Fold-4	Mean
$\mathcal{F}$	Beginning	40.3	61.9	59.8	59.1	55.2
	Middle	40.0	62.5	59.5	59.1	55.3
	End	40.1	62.0	60.4	58.7	55.3
$\mathcal{J}$	Beginning	41.3	64.3	60.3	61.1	56.8
	Middle	41.1	64.8	60.3	61.4	56.9
	End	41.3	64.1	61.8	60.6	56.8

Table 3. The effect of domain agent position. Selecting the agent at different positions make almost no difference on the final results.

As shown in Tab. 3, the influence of different agent position is small, and choosing the middle one is slightly better than the other two on average.

5. The Effect of Support Class

In this experiment, we verify whether our method learns to segment different objects based on the class of the support set through meta-learning. We obtain uncorrelated support sets by randomly sampling images of different classes. As shown in Tab. 4, the randomly sampled support sets can not segment the target objects well, which validates that our method is to segment query videos based on the novel class indicated by the support set.

Support class		Fold-1	Fold-2	Fold-3	Fold-4	Mean
$\mathcal{F}$	Random	29.7	46.7	38.6	43.5	39.6
	Same	40.3	62.3	60.2	59.4	55.6
$\mathcal{J}$	Random	22.5	44.8	34.0	42.8	36.1
	Same	41.5	64.8	61.3	61.4	57.2

Table 4. The effect of support class. Support sets of the same class as the target objects can effectively segment the objects compared to support sets of random classes.

6. The Effect of Domain Agent Content

Theoretically, the domain agent contains more representative information about the query video, the model achieves better performance. To verify whether the performance is affected if the domain agent does not contain the object-of-interest. We set the value of domain agent to zero during the test. As shown in Tab. 5, domain agent containing semantic content lead to performance gains.

Agent content		Fold-1	Fold-2	Fold-3	Fold-4	Mean
$\mathcal{F}$	Zero	39.5	60.8	58.3	58.6	54.3
	Semantic	40.3	62.3	60.2	59.4	55.6
$\mathcal{J}$	Zero	36.0	62.6	58.8	60.8	54.5
	Semantic	41.5	64.8	61.3	61.4	57.2

Table 5. The effect of domain agent content. The agent does not contain semantic content causing performance degradation.

7. More Qualitative Results

We show more qualitative results in Fig. 2. Our model can segment the target objects accurately across different frames.

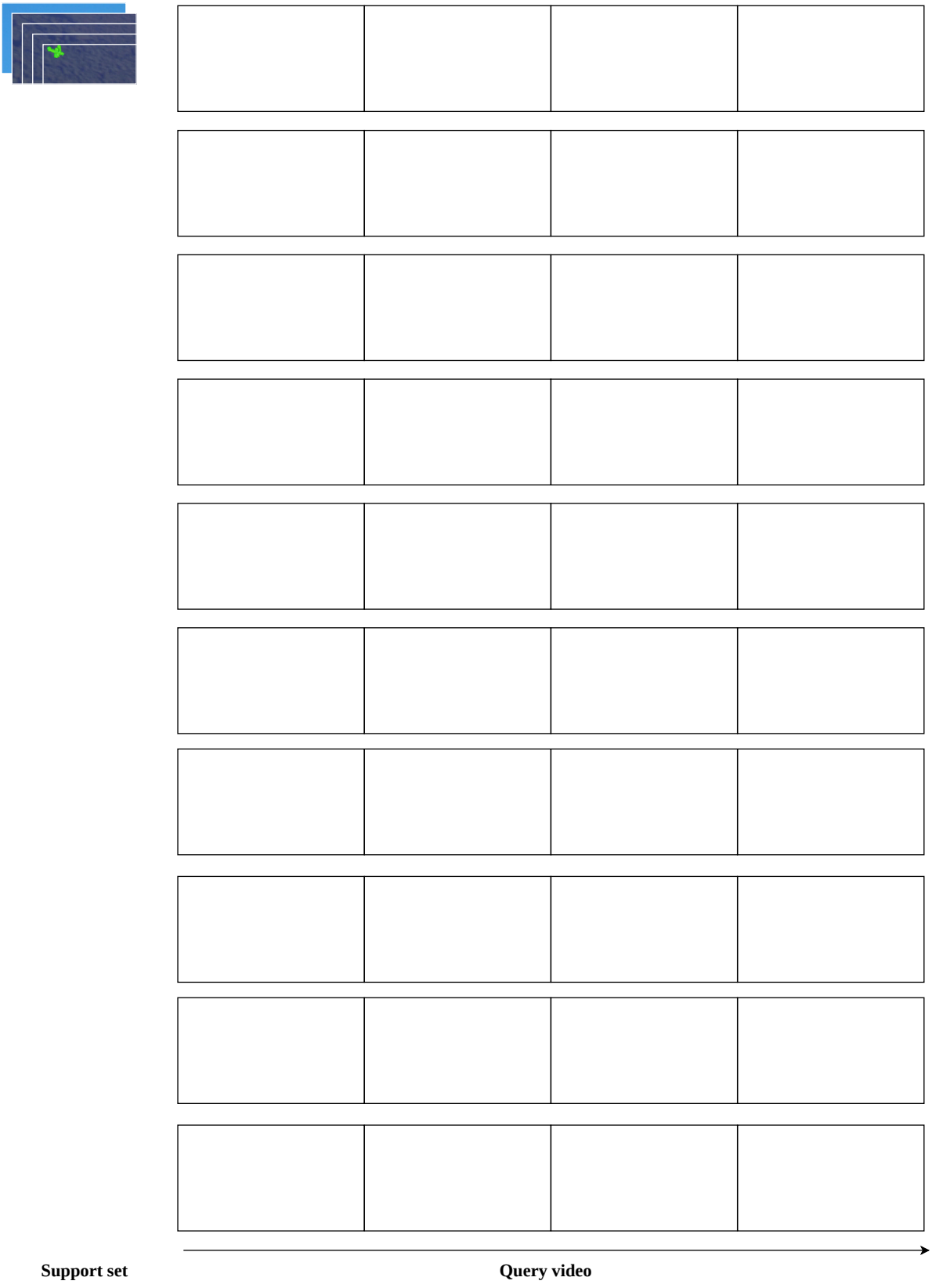

Figure 2. More qualitative results of our model.