DeVOS: Flow-Guided Deformable Transformer for Video Object Segmentation

Supplementary Material

1. More Implementations Details

1.1. Matching details

The proposed multiscale flow-guided deformable cross-attention block effectively processes frame embeddings E^i, utilizing movement information from flow embeddings E^f. This information is incorporated into the original frame embeddings E^i and normalized to form queries and keys.

$$LayerNorm(x) = \frac{x - E(x)}{Var(x) + \epsilon} * \phi + \beta$$

$$Q = LayerNorm(E^i_{curr} + WE^f_{forward})$$

$$K = LayerNorm(E^i_{prev} + WE^f_{backward})$$

Flow-based and semantic-based offsets are computed and combined to form sampling offsets that cater to extracting keys and values from different scales. Afterwards, the correlation between keys and queries is computed. The output features pyramid consists of the weighted sum of joined mask and image embeddings from the previous frame.

1.2. Training details

For pre-training, we use an initial learning rate of 4e-4 and a weight decay of 0.03 for 100000 steps, similarly to [1]. Our main training is divided into two phases, first utilizes ground truth mask as previous frame prediction, and the other propagates predicted mask through sampled sequence. On the first phase, the model is trained for 50K optimization steps, while the second phase takes 100K optimization steps. Initial learning rate for main training equals 2e-4 and the weight decay is 0.05. Mask and flow encoders are frozen before second phase. For the whole training, length of our sampled sequence is 5.

During our main training, we use curriculum sampling strategy similarly to [2]. We employ a combination of DAVIS 2017 [3] train and Youtube-VOS [4], [5] train datasets in 5:1 proportion. Additionally, we study adopting MOSE 2023 [6] as additional training data, with mixture of DAVIS, Youtube-VOS and MOSE with proportion $5 : k : p$ where $k + p = 1$. Initial fraction of Youtube-VOS $k_{start} = 0.75$ linearly decays to a final value of $k_{end} = 0.15$. To additionally prevent overfitting and increase average number of objects present on the scene, for the DAVIS and Youtube-VOS dynamic merge augmentation with probability of 0.4 is applied. Considering complexity of the MOSE dataset even without dynamic merge augmentation, we employ importance sampling augmentation specifically for this dataset. Further details regarding this will be elaborated in Section 2.

We adopt AdamW optimizer [7] with a one-cycle learning rate schedule. Initial learning rate for both stages declines in polynomial manner with 0.9 decay factor to a final value of 1e-5. We also use learning rate warm-up for 5000 iterations. To address overfitting of our encoders, we set the learning rate for them to 0.1 of a total learning rate. Following [8], we use bootstrapped cross-entropy and dice losses with equal weighting. For both stages, we use a batch size of 16. DeVOS-L model training is distributed across four Tesla A100 GPUs, while for DeVOS-B we use four RTX 3090 GPUs. The entire training process takes around 80 hours for the large model and 60 hours for the basic one.

2. MOSE 2023

2.1. Training

Classical VOS datasets [3]–[5] lack sequences with a large number of objects present. To address this issue, a dynamic merge augmentation is introduced, which involves merging two videos with a certain probability, denoted as p. However, due to the high complexity of scenes in the MOSE dataset, the utilization of the same augmentation technique is deemed suboptimal as it would make the scenes too challenging for the model to effectively learn meaningful information. In order to maximize the benefits of incorporating the MOSE dataset as additional training data for the VOS task, a novel augmentation technique called importance sampling is proposed. This technique involves

Table 1. The results of training with MOSE 2023. IS - Importance Sampling.

MOSE	IS	D_{17V}	Y_{19}
✗	✗	86.1	85.2
✓	✗	86.0	85.2
✓	✓	86.4	85.4

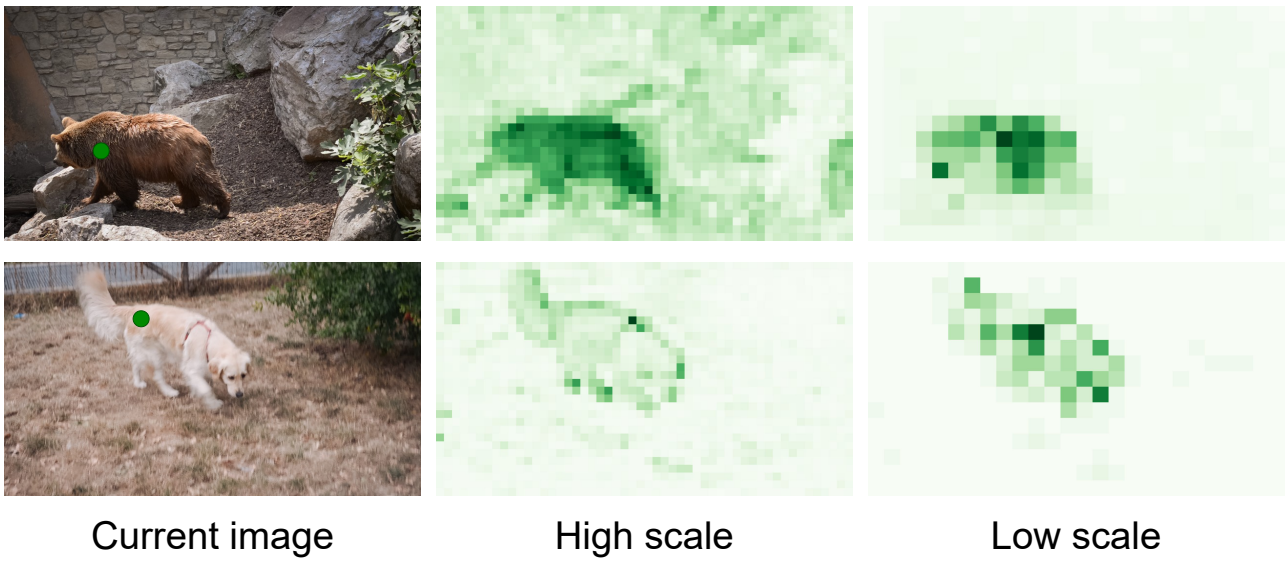

Figure 1. Attention maps (1/16, 1/32) for long-term memory matching branch.

assigning a probability of sampling, denoted as p_i, to each sequence i from the dataset, which consists of n sequences. The probability p_i is determined using the following formula:

$$p_i = \frac{o_i^T}{\sum_{j=1}^{n} o_j^T}$$

where o_i represents the number of objects in the i-th sequence, and T is a temperature parameter that controls the sharpness of the probability distribution. For our training, we use $T = 1$. The adoption of importance sampling provides a boost in performance, which can be observed in the Tab. 1.

2.2. Quantitative comparison with other methods on MOSE 2023

The quantitative comparison on MOSE 2023 validation is presented in Tab. 2.

3. Illustration of Multi-Scale Attention

We argue that multi-scale matching allows our approach to benefit both from lower scale features that capture more semantic information and higher-scale features that contain richer spatial and fine-grained details. To support our claim, we visualize attention maps from the long-term matching branch for a given query on different scales. As we see in Fig. 1, low scale attention maps exhibit high certainty regarding particular semantic areas, allowing to conduct better matching for semantic categories, while attention maps on higher scale pay more attention to edges. All of this results in effective and accurate matching of objects even on complex sequences where occlusion is present.

4. Additional qualitative comparisons

4.1. DAVIS 2017 Test set

We conduct additional qualitative comparisons with recent relevant VOS methods on DAVIS 2017 test set as it features more challenging sequences. The results show that our method gives superior results under rapid movement, scale, and appearance changes. This is the most prominent for the ”giant-slalom” sequence, where XMem fails completely, likely because the working memory fails to model objects with large motion blur. On the contrary, our approach succeeds in such cases due to advanced short-term matching capabilities. Moreover, we argue that thanks to

Table 2. The quantitative evaluation on MOSE 2023.

Methods	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
AOT	53.1	61.3	57.2
STCN	46.6	55.0	50.8
RDE	44.6	52.9	48.8
SWEM	46.8	54.9	50.9
XMem	53.3	62.0	57.6
DeAOT	55.1	63.8	59.4
DeVOS-B	54.2	62.4	58.3
DeVOS-L	58.3	67.1	62.7

the proposed multi-scale matching, we successfully match objects on different scales. Thanks to large-scale pretraining of the image backbone we utilize, we are able to get better performance on the sequences with challenging illumination - "tractor", "people-sunset", "deer". The general notion of what objects are helps with sequences with challenging occlusions as well ("salsa").

4.2. YouTube VOS

In order to demonstrate the robustness of our method, we also provide its predictions on challenging sequences from the YouTube VOS dataset. We carefully examine the qualitative results within three primary categories of complex scenes: fast motion (such as surfer and bike), occlusions (illustrated by the deer example), and scenarios with poor image quality, including strong blurring (depicted by the spider) and unfavorable lighting conditions (as observed in the night car scene).

5. Future work & Ethical considerations

As mentioned previously, our approach is independent and complementary to the methods proposed by XMem, ISVOS, and DeAOT. Therefore, it is logical to further enhance our approach by incorporating a more intelligent memory scheme, integrating instance understanding from a separate instance-segmentation branch, and decoupling image features accordingly. To improve the quality of similarity search regions for specific queries, it is also worth exploring the implementation of kernelized memory reads, following [9].

Video object segmentation plays a crucial role in various applications, such as video editing and augmented reality. At the same time, it is essential to consider the potential misuse of this technology, including unauthorized surveillance or malicious alternation of videos. The presence of bias in VOS can lead to unfair outcomes and perpetuate societal inequalities. Taking all of this into consideration, the models trained with our approach on real-world datasets should undergo ethical review to ensure that it is usable and beneficial for everyone and is not used for the application, including but not limited to illegal surveillance.

References

[1] Z. Yang, Y. Wei, and Y. Yang, "Associating objects with transformers for video object segmentation," *arXiv preprint arXiv:2106.02638*, 2021.

[2] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video object segmentation using space-time memory networks," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9225–9234. DOI: 10.1109/ICCV.2019.00932.

[3] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Computer Vision and Pattern Recognition*, 2016.

[4] N. Xu, L. Yang, Y. Fan, *et al.*, *Youtube-vos: A large-scale video object segmentation benchmark*, 2018. arXiv: 1809.03327 [cs.CV].

[5] N. Xu, L. Yang, Y. Fan, *et al.*, *Youtube-vos: Sequence-to-sequence video object segmentation*, 2018. arXiv: 1809.00461 [cs.CV].

[6] H. Ding, C. Liu, S. He, X. Jiang, P. H. Torr, and S. Bai, "Mose: A new dataset for video object segmentation in complex scenes," *arXiv preprint arXiv:2302.01872*, 2023.

[7] I. Loshchilov and F. Hutter, *Decoupled weight decay regularization*, 2019. arXiv: 1711.05101 [cs.LG].

[8] H. K. Cheng and A. G. Schwing, "XMem: Long-term video object segmentation with an atkinson-shiffrin memory model," in *ECCV*, 2022.

[9] H. Seong, J. Hyun, and E. Kim, "Kernelized memory network for video object segmentation," *ArXiv*, vol. abs/2007.08270, 2020.

[10] J. Wang, D. Chen, Z. Wu, *et al.*, *Look before you match: Instance understanding matters in video object segmentation*, 2022. arXiv: 2212.06826 [cs.CV].

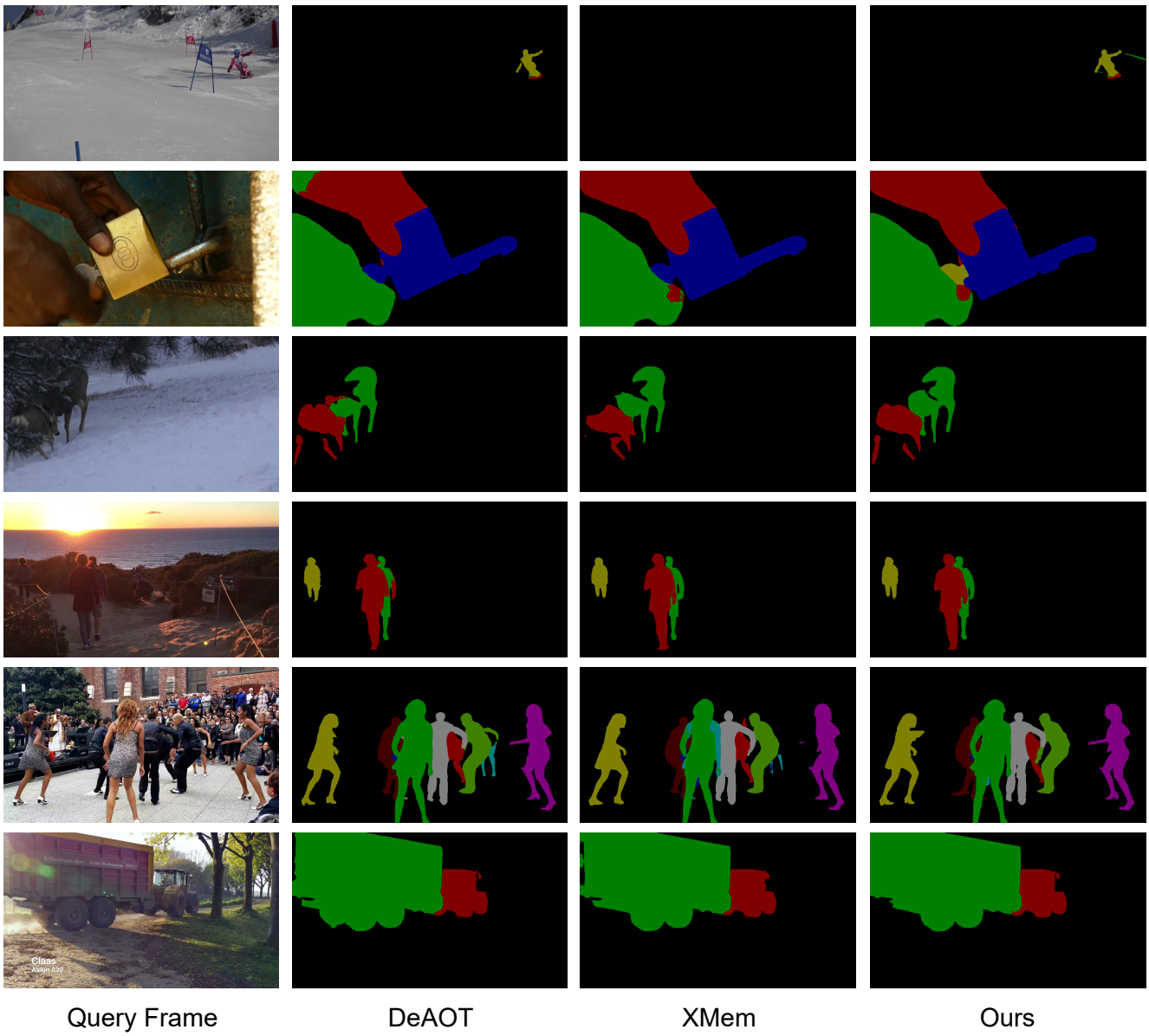

Figure 2. Qualitative comparison between DeVOS and some state-of-the-art VOS methods on DAVIS 2017 Test set. Best viewed in zoom. We don't include ISVOS [10] since there is no source code available. For all methods we used DAVIS2017 test-dev sequences in 480p.

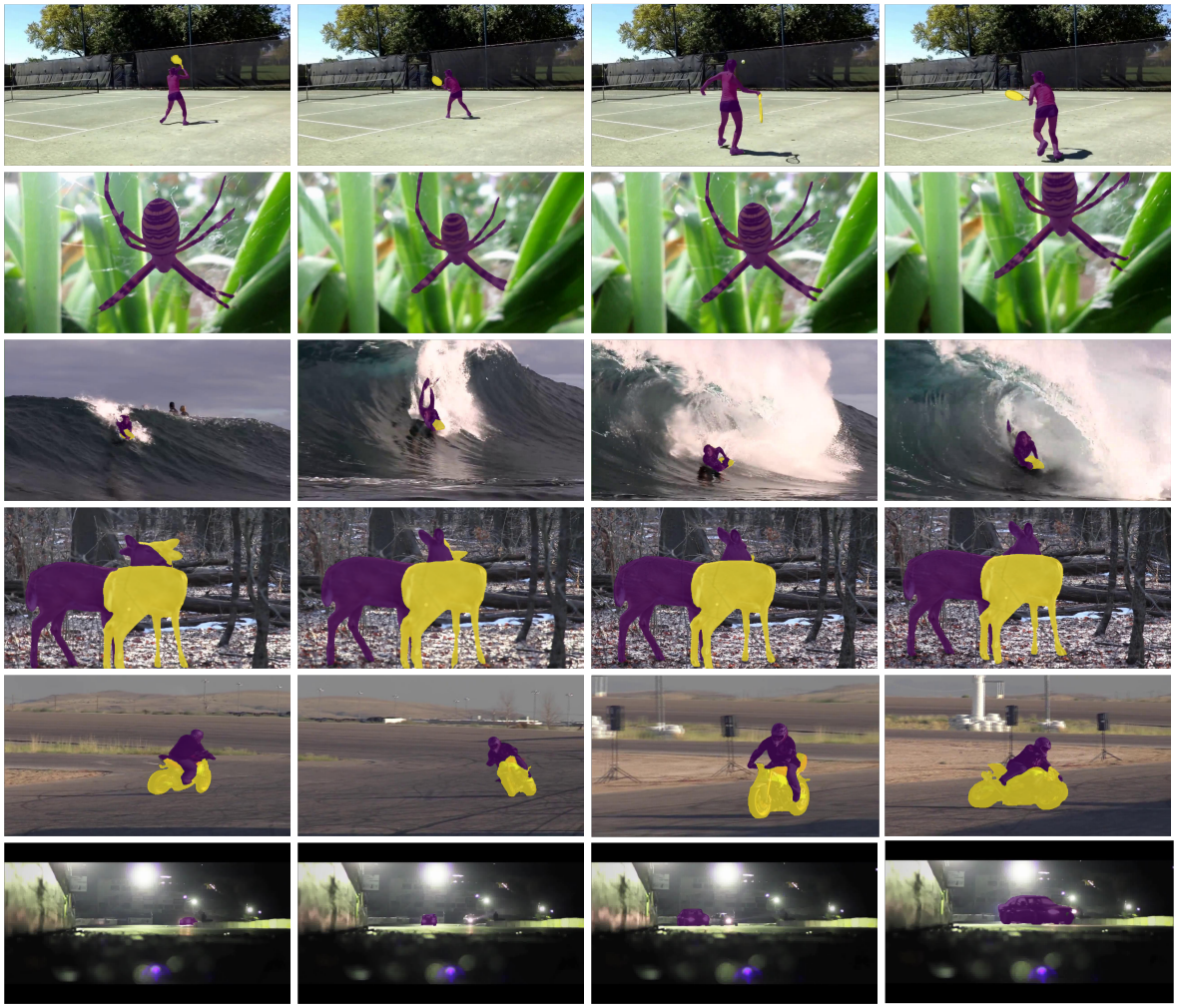

Figure 3. Qualitative results on the validation split of YouTube-VOS 2019.