DeVOS: Flow-Guided Deformable Transformer for Video Object Segmentation

Volodymyr Fedynyak[1], Yaroslav Romanus[1,2], Bohdan Hlovatskyi[1], Bohdan Sydor[1],
Oles Dobosevych[1], Igor Babin[1,2], Roman Riazantsev[2]

[1]Ukrainian Catholic University, [2]ADVA Soft

{v.fedynyak, yaroslav.romanus, bohdan.hlovatskyi, b.sydor, dobosevych}@ucu.edu.ua

{ihor.babin, roman.riazantsev}@adva-soft.com

Abstract

The recent works on Video Object Segmentation achieved remarkable results by matching dense semantic and instance-level features between the current and previous frames for long-time propagation. Nevertheless, global feature matching ignores scene motion context, failing to satisfy temporal consistency. Even though some methods introduce local matching branch to achieve smooth propagation, they fail to model complex appearance changes due to the constraints of the local window. In this paper, we present DeVOS (Deformable VOS), an architecture for Video Object Segmentation that combines memory-based matching with motion-guided propagation resulting in stable long-term modeling and strong temporal consistency. For short-term local propagation, we propose a novel attention mechanism ADVA (Adaptive Deformable Video Attention), allowing the adaption of similarity search region to query-specific semantic features, which ensures robust tracking of complex shape and scale changes. DeVOS employs an optical flow to obtain scene motion features which are further injected to deformable attention as strong priors to learnable offsets. Our method achieves top-rank performance on DAVIS 2017 val and test-dev (88.1%, 83.0%), YouTube-VOS 2019 val (86.6%) while featuring consistent run-time speed and stable memory consumption.

1. Introduction

Video Object Segmentation (VOS) is a fundamental task of video understanding. In a semi-supervised approach, it is formulated as the identification and segmentation of objects through the video sequence given the ground truth annotation masks for the first and, optionally, some other frames.

Previous VOS methods [1]–[5] focus on distilling the information from past frames into a *feature memory* storage and then perform a dense memory matching to identify objects on the current frame. Some approaches [6]–[8] suggest enhancing memory-based matching with mask propagation to achieve smooth predictions and improve temporal consistency. Yang *et al.* in their work "Associating Objects with Transformers for Video Object Segmentation" (AOT) [9] proposed using image attention mechanism [10] to perform hierarchical propagation and matching, employing global attention for memory readouts and local windowed attention for short-term propagation. In DeAOT [11], the architecture was further improved by decoupling processing of visual and object information.

Recently, Wang, Chen, Wu, *et al.* in ISVOS [5] noticed that existing methods suffer from performance degradation in scenarios of substantial shape deformations and appearance changes caused by camera and scene motion. The authors propose to utilize instance discriminative features while performing dense matching with the memory bank, ensuring selecting of the correct object from past frames and avoiding false positives. ISVOS achieves state-of-the-art performance on most of the benchmarks, outperforming the methods specifically designed for long-time videos, *e.g.*, XMem [12] and AFB-URR [13] on the Long-time Video dataset. However, the main research effort of the aforementioned approach lies in determining *how to improve features for matching* without focusing on *how exactly to perform the matching*.

The temporal evolution of an object's appearance depends on the semantic properties, *i.e.* rigidity. Thus it's natural to adapt the similarity search region to specific semantic features of the query point. Some existing implementations of matching logic construct a global affinity matrix between current and previous features and use similarity score as a matching objective. STCN [2] and ISVOS [5] adopt negative L2 distance for this purpose, treating all possible search locations equally. XMem [12] proposes anisotropic L2 similarity, allowing query-specific importance interpretation. AOT [9] shrinks the search space by using windowed cross-attention for short-term matching, while the query-specific importance is assured by learnable relative posi-

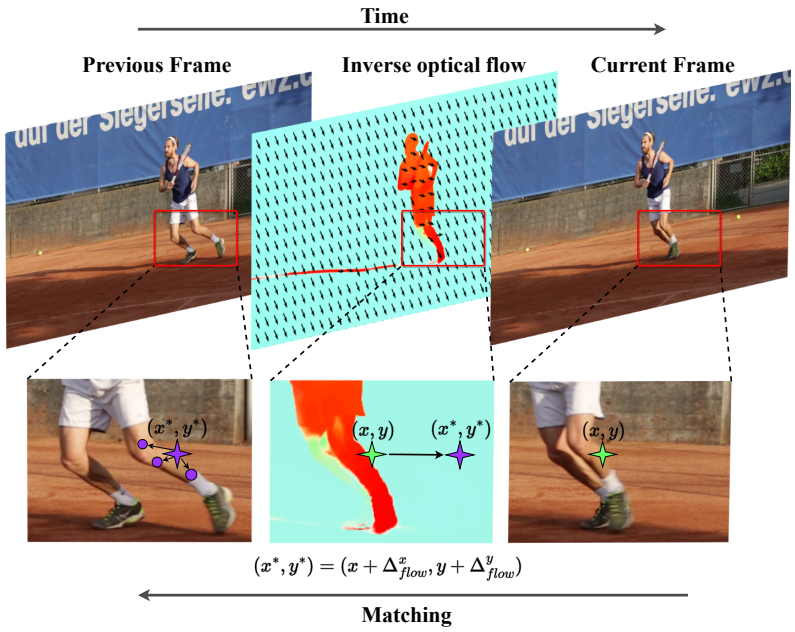

Figure 1. The process of matching features between the current and preceding frames is divided into two steps: flow-based displacement adjustment and semantics-driven deformable attention

tion bias. In existing methods the query-specific adaptation is limited to only tweaking importance over the query-agnostic set of spatial locations (often limited to a region around the spatial location of a query). In terms of handling motion, global matching leads to degenerating of temporal-spatial consistency, while windowed matching fails to capture rapid movement.

We argue that adapting similarity search region to specific query semantic properties is crucial to perform propagation robust to appearance, scale, and shape change. To further enhance the performance, we propose to decouple motion and semantics during matching, adopting a global scene displacement field as an initial offset of the search region.

In this spirit, we present DeVOS, a novel architecture for VOS introducing a new attention-based short-term matching mechanism ADVA (Adaptive Deformable Video Attention). Inspired by Deformable DETR [14] and DAT [15], we adopt multi-scale deformable cross-attention capable of sampling search locations on the previous frame based on motion and query-specific semantic features of the current frame. More specifically, given some reference location, we predict initial global offset using the scene motion features, *positioning* the search region. Consequently, we use corresponding query features to predict several local offsets, *shaping* the search region. Finally, keys and values are sampled from predicted locations using the previous frame and passed to multi-head attention. Comparing to previous methods, we present formulation of deformable cross-attention for video-related tasks, while preserving efficient query offset modelling. ADVA is described in details in Sec. 3.1. Furthermore, we enhance the keys and queries of the matched video frames with motion features to achieve strong temporal consistency, which is described in details in Sec. 3.2. The short-term and long-term memory matching results are fused and passed to the decoder producing the final propagated object mask. To obtain motion features a generic optical flow estimation network is used.

We conduct experiments on the standard DAVIS [16] and YouTube-VOS [17] benchmarks. We optionally conduct additional training on the large-scale MOSE 2023 [18] dataset to achieve robustness under complex VOS scenarios. Conducted experiments demonstrate that DeVOS achieves top-ranked performance while enjoying consistent run-time speeds. It is worth noting that our research direction is orthogonal to those in ISVOS [5], DeAOT [11], and XMem [12] and can further benefit from the ideas presented in those works.

2. Related Work

2.1. Optical Flow Estimation

Optical flow estimation is crucial for modeling global motion. Initial studies focused on optimization problems, emphasizing visual similarity and regularization [19]–[22]. The introduction of deep neural networks, especially convolutional networks, significantly advanced this field.

The RAFT model [23] introduced a significant upgrade

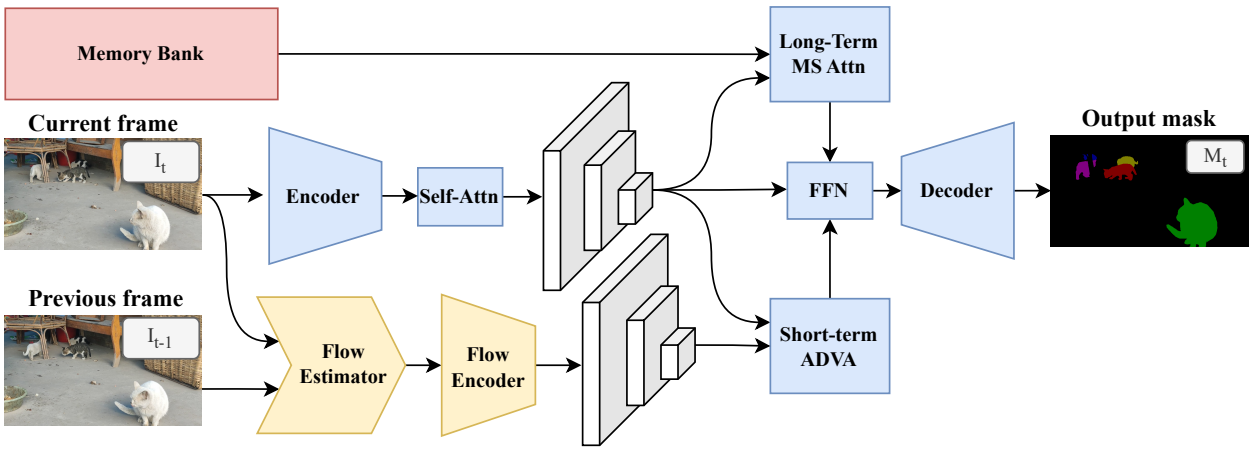

Figure 2. The overview of DeVOS architecture. The current frame is processed through encoder and self-attention block. After that, optical flow between current and previous frames is computed for the adaptive deformable video attention between current and previous frame features. Information from a memory bank containing frames for long-term memory is incorporated through a long-term multi-scale deformable attention block.

to optical flow estimation, incorporating the multi-scale search window through the recurrent module. Following the introduction of RAFT, subsequent studies like GMA [24] and DEQ-Flow [25] further improved accuracy and computational efficiency. FlowFormer [26] extends RAFT by utilizing a transformer-based strategy for aggregating cost volume in latent space, building on Perceiver IO [27]. It pioneered the use of transformers [10] for long-range relationships in optical flow, achieving top-tier performance.

Recently, Fedynyak, Romanus, Dobosevych, *et al.* in WarpFormer [28] showed that employing an optical flow estimator to support a generic VOS architecture by warping the past frames into the current frame domain could be benefitial for smooth propagation.

2.2. Video Object Segmentation

A key approach in the field of Video Object Segmentation (VOS) is AOT (Associating Objects with Transformers for VOS) [9]. This method uses a Long Short-Term Transformer (LSTT) block that incorporates short-term attention and long-term attention mechanisms to extract features from input imagery. Long-term attention gathers information from extended memory frames, while short-term attention disseminates information from the previous frame. The outputs of both attention units are integrated into a feed-forward network and then passed to the decoder to predict the current object mask.

DeAOT [11] builds on the hierarchical propagation concept of AOT for semi-supervised video object segmentation, introducing a dual-branch propagation for object-agnostic and object-specific embeddings.

XMem [12] is a Video Object Segmentation (VOS) architecture designed for long videos. It utilizes the Atkinson-Shiffrin memory model to create an architecture with multiple independent, interconnected feature memory stores. It incorporates a sensory memory, a working memory, and a long-term memory. A memory potentiation algorithm is used to consolidate working memory elements into long-term memory, preventing memory overload and maintaining performance for long-term prediction.

The paper ISVOS [5] further highlights the importance of instance understanding in VOS. While recent memory-based methods have achieved impressive results in VOS through dense matching between current and past frames, these methods often falter when confronted with large appearance variations or viewpoint changes caused by object and camera movements. To mitigate these issues, the authors propose a two-branch network for VOS, which incorporates a query-based instance segmentation (IS) branch to delve into the instance details of the current frame. This approach allows the integration of instance-specific information into the query key, facilitating instance-augmented matching. These works collectively underscore the importance of instance understanding in VOS and propose solutions that effectively integrate this concept into existing VOS methods.

2.3. Vision Transformers and Deformable Attention

Transformers have gained traction in computer vision, yet their large receptive fields pose computational and memory challenges. Deformable attention, introduced in Deformable DETR [14] and Deformable Attention Transformer (DAT) [15], addresses these issues by focusing on a small set of key sampling points, reducing computational load and enhancing performance.

Deformable DETR applies deformable attention in the

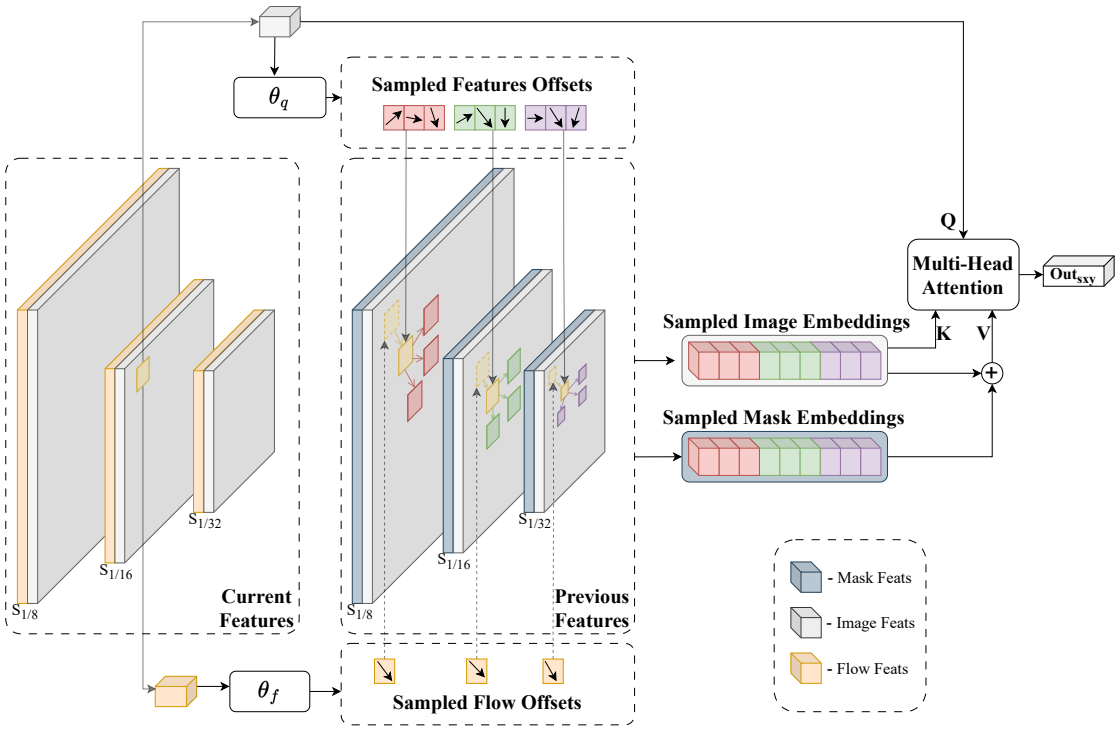

Figure 3. Adaptive deformable video attention. The multi-scale flow-based feature matching consists of two steps: offsets prediction for features alignment and multi-head attention. Two types of offsets are used: flow-based offsets for movement compensation and semantic-based offsets to extract previous frame image and mask embeddings. Multi-head attention combines the previous frame mask, and image embeddings based on the correlation of the previous frame sampled features and query image embedding vector.

detection head to improve performance on small objects and speed up convergence. DAT introduces a deformable self-attention module in the vision backbone, enabling data-dependent selection of key and value pairs, thus efficiently capturing more informative features and modeling long-range relations.

3. Method

To describe our method, let's consider a video sequence denoted as $V = [X_1, X_2, ..., X_T]$, along with the annotation mask of the first frame. Our approach processes the frames sequentially, storing the predicted results in memory to inform future predictions. Firstly, we extract features from the current image, X_t, using a backbone encoder, resulting in a feature map I_t. Subsequently, the feature map of the current image is compared with the memory frames to perform semantic matching and propagate the mask. Finally, the matching result, combined with features from the encoder at multiple scales, is fed into the decoder, which restores the object mask in the original resolution (see the full architecture in Figure 2).

3.1. Adaptive Deformable Video Attention

The classic attention operation is defined as follows:

$$\text{Att}(Q, K, V) = \text{Corr}(Q, K)V = \text{softmax}\Big(\frac{QK^T}{\sqrt{C}}\Big)V$$

where Q, K, and V denote the attention queries, keys, and values, respectively, while C is the embedding space dimension. In order to choose only the relevant spatial locations, a data-driven approach is used for selecting some predefined number N_k of key and value pairs. For each query and its corresponding reference point, N_k offsets are learned to indicate the specific locations from which values and keys should be sampled. These offsets are obtained based on the query features, ensuring that they capture and represent semantic information:

$$\Delta p_{qk} = \theta_{\text{offset}}(QW^Q),$$

where θ_{offset} - sub-network for offset generation. To stabilize the training, the predicted offsets are scaled to fit into window with size σ: $\Delta p_{qk} \leftarrow \sigma \cdot \tanh(\Delta p_{qk})$. After offsets are sampled attention is computed as in classical formulation:

$$\text{DfAttn}(Q, K, V, p_q) = \text{softmax}\Big(\frac{QK_p^T}{\sqrt{C}}\Big)V_p,$$

$$p \in \{p_q + \Delta p_{qk} : k \in K\}$$

where subscript k refers to an index of learned offset for a given query. Such formulation allows learning sparse regions to attend to each of the queries and can be naturally extended to high-resolutions as it is linear with respect to spatial resolution. Following the [14] and [10], we extend this formulation with multi-scale feature maps and multi-head attention correspondingly. For each resolution, we add both learnable positional embeddings π and scale-level embeddings ω. Moreover, the window size σ is dynamically adjusted in proportion to the scale.

Motion becomes a crucial factor when designing deformable cross-attention for multiple frames. As the nature of the movement is isotropic, in the same naive formulation, query offsets would be forced to learn windowed attention. This is unwanted as it undermines offsets' ability to learn query-specific information and thus - similarity search region adaptation. To mitigate this issue, we propose to decouple *motion* and *semantic* information, creating separate offset branches for them:

$$\Delta p_{qk}^q = \theta_{\text{offset}}^q(QW^Q), \Delta p_{qk}^f = \theta_{\text{offset}}^f(F_{inv}W^F),$$

where F_{inv} denotes inverse optical flow and $\theta_{\text{offset}}^q, \theta_{\text{offset}}^f$ - sub-networks for offset generation based on queries and flow respectively. Besides, we normalize the predicted flow offsets to fit into the image: $\Delta p_{qk}^f \leftarrow D \cdot \tanh(\Delta p_{qk})$, where D denotes spatial dimension size. Afterward, the total offsets are computed as the combined sum of semantic(query-based) and motion(flow-based) offsets:

$$\Delta p_{qk} = \Delta p_{qk}^q + \Delta p_{qk}^f$$

This novel type of attention, called adaptive deformable video attention (ADVA), adapts search region for cross-attention based on the motion and semantic features, thus showing superior perfomance on VOS benchmarks. We believe that is can be applied to various video-related tasks beyond VOS as well.

3.2. QK-flow

To further leverage motion information, we explore the possibility of integrating it into the semantic feature map. We argue that this enhancement helps in distinguishing between different instances of the same semantic class, as they naturally have distinct motion patterns. Formally, we denote the direct and inverse flow between the previous and current frames as F_{dir} and F_{inv}. To integrate motion information, we augment our queries (Q) with the linearly projected flow towards the previous frame: $Q_m = Q + W_{inv}F_{inv}$. Similarly, we augment our keys (K): $K_m = K + W_{dir}F_{dir}$. Here, the subscript m indicates that our queries and keys have been enriched with motion information.

3.3. Multi-scale matching

To benefit from the sparsity of the proposed attention formulation to its fullest, we propose to conduct semantic matching with memory bank on multi-scale feature maps. We argue, that it helps dealing with overlapping objects that share a similar appearance thanks to effective utilization of high-resolution features. Formally, our backbone encoder generates features at multiple scales, denoted as $[I_t^{(1)}, I_t^{(2)}, I_t^{(3)}]$, corresponding to scales of $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$, respectively. To ensure consistent matching, we map the features on different scales into the same embedding space with linear projections.

Subsequently, our multi-scale features are passed to a self-attention block, implemented as deformable self-attention [14]. For short-term matching, we employ sparse attention in the form of the ADVA, which is described in Sec. 3.1. During the long-term matching, we stack the flattened encoder feature maps on the spatial dimensions of $\frac{1}{16}$ and $\frac{1}{32}$, then perform an attention-based global matching with the memory bank.

3.4. Network details

To study performance capabilities and contributions impact, we introduce two variants of network architecture. Namely, DeVOS-B (Base) is a baseline implementation of the proposed method featuring consistency with previous approaches and considerable runtime speeds. Alternatively, DeVOS-L (Large) is a larger-scaled configuration for which we adopt more advanced building blocks and inject more complex architecture decisions.

Encoder & Decoder To achieve fairness in comparison and to keep consistency with previous works [5], [9], [11], [12], we equip our basic model DeVOS-B with ImageNet1K [29] pre trained ResNet50 [30] image feature encoder. Meanwhile, with the aim of enhancing instance understanding logic, our bigger model DeVOS-L is equipped with ViT-B [31] encoder pre trained on Segment Anything Dataset [32]. We assume that large-scale pre-training of transformer encoder on supervised instance segmentation is more suitable for video object segmentation as it allows the backbone to learn the notion of what objects actually are. We leave this fact, though, for further research. FPN [33] decoder with Group Normalization [34] is used in both DeVOS-B and DeVOS-L.

Table 1. The quantitative evaluation on multi-object benchmarks YouTube-VOS 2019 and DAVIS 2017. * denotes training on MOSE 2023. † denotes replacing ResNet50 with Swin-B encoder. ‡ denotes FPS retimed on our hardware. Top-3 results are denoted in bold font.

	YouTube-VOS 2019 Val					DAVIS 2017 Val			DAVIS 2017 Test			
Methods	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	Avg	$\mathcal{J}$	$\mathcal{F}$	Avg	$\mathcal{J}$	$\mathcal{F}$	Avg	FPS
STCN	81.1	85.4	78.2	85.9	82.7	82.2	88.6	85.4	72.7	79.6	76.1	19.5
AOT-L	83.5	88.1	78.4	86.3	84.1	82.3	87.5	84.9	75.9	83.3	79.6	18.0
AOT-L†	84.0	88.8	78.4	86.7	84.5	82.4	88.4	85.4	77.3	85.1	81.2	12.1
DeAOT-L	84.6	89.4	80.8	88.9	85.9	82.2	88.2	85.2	76.9	84.5	80.7	34.0‡
DeAOT-L†	85.3	90.2	80.4	88.6	**86.1**	83.1	89.2	86.2	78.9	86.7	**82.8**	21.1‡
XMem	84.3	88.6	80.3	88.6	85.5	82.9	89.5	86.2	77.4	84.5	81.0	34.4‡
ISVOS	85.2	89.7	80.7	88.9	**86.1**	83.7	90.5	**87.1**	79.3	86.2	**82.8**	-
DeVOS-B	84.5	89.5	79.4	87.4	85.2	83.4	88.8	86.1	77.2	84.7	81.0	36.7‡
DeVOS-L	85.2	90.1	80.7	89.0	**86.3**	84.2	91.2	**87.7**	79.4	86.4	**82.9**	24.7‡
DeVOS-B*	84.7	89.7	79.4	87.8	85.4	83.5	89.3	86.4	77.4	84.9	81.2	36.7‡
DeVOS-L*	85.4	90.3	80.8	89.3	**86.6**	84.4	91.8	**88.1**	79.4	86.6	**83.0**	24.7‡

Object masks Following [2], [5], we adopt a lightweight ResNet18 [30] network to encode one-hot object masks into the multi-scale embedding space. The number of input channels to mask encoder is set to 15, matching the maximal object number in benchmarks. To achieve homogeneous and simultaneous learning of segmentation mask representation while training, the input one-hot mask is zero-padded to have 15 channels, and the objects (i.e., channels) are then randomly shuffled.

Flow representation Optical flow field is used to capture the motion context between consecutive frames. For this, we employ GMA [24] network due to its favorable performance and flexibility in adjusting run-time speed by tweaking the number of refinement updates. Even though the original paper suggests performing 12 updates, we find that four is enough to provide a strong displacement prior to matching. Notably, our model is designed to be independent of the actual flow estimator implementation. To construct a multi-scale motion representation from estimated optical flow, a lightweight ResNet18 [30] is used.

4. Experiments

4.1. Implementation details

Training details Similarly to [1], [2], [4], [5], [9], [11], [12], we split the training of DeVOS into two stages. During the first stage, we adopt pretraining on synthetic sequences derived from static image datasets [35]. Consequently, we conduct main training on DAVIS 2017 [16], YouTubeVOS 2019 [17], and optionally MOSE 2023 [18]. A more detailed description of training is provided in Supplementary.

Evaluation In order to evaluate our models, we use traditional VOS metrics as proposed in [36]. We evaluate our method on DAVIS 2016 & 2017 using the default 480p 24FPS videos, not benefiting from higher resolutions or test-time augmentations. The impact of multi-scale inference [37] augmentation is studied in Supplementary. While evaluating our method on YouTube-VOS 2019 validation split, we exploit all intermediate frames of the videos to benefit from smooth motion implying more accurate optical flow. Even though we use 24 FPS sequences during evaluation, the 6FPS version is used during training and for metric computation.

Inference Following [1], [2], [5], [12], we maintain feature memory by memorizing every fifth frame during inference. To keep consistent run-time speeds and stable memory consumption, the memory bank is implemented as a FIFO queue with a maximum size of 16. Meanwhile, the

Table 2. The quantitative evaluation on DAVIS 2016.

Methods	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
AOT-T	86.1	87.4	86.8
DeAOT-T	87.8	89.9	88.9
STCN	90.8	92.5	91.6
XMem	90.4	92.7	91.5
ISVOS	**91.5**	93.7	92.6
Swin-B AOT-L	90.7	93.3	92.0
Swin-B DeAOT-L	91.1	94.7	92.9
DeVOS-B	90.8	93.0	91.9
DeVOS-L	91.0	**95.8**	**93.5**

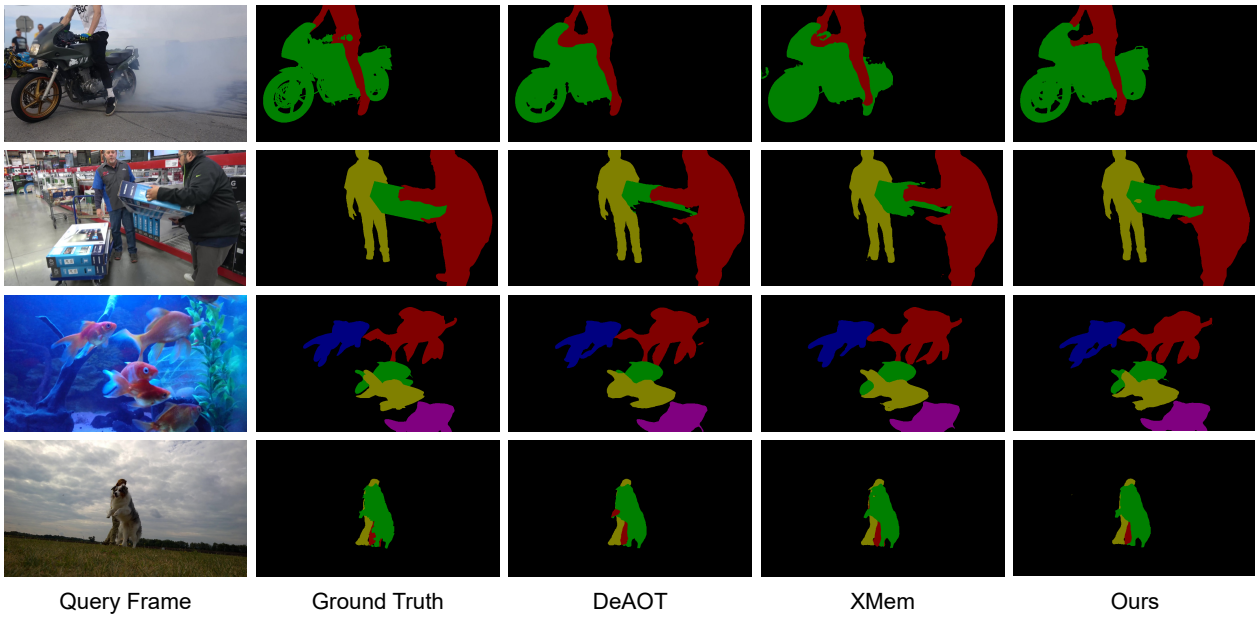

Figure 4. Qualitative comparison between DeVOS and some state-of-the-art VOS methods. Best viewed in zoom. We don't include ISVOS [5] since there is no source code available. For all methods we used DAVIS2017 val sequences in 480p.

first frame is always kept in the memory [5]. We don't use top-k filtering [3] or kernelized memory readouts [38] as we rely on short-term matching for smooth propagation and on QK-flow for temporal consistency.

4.2. Comparison with State-of-the-art Methods

Quantitative comparison Table 1 presents a comparison of DeVOS with other state-of-the-art methods on DAVIS 2017 validation, DAVIS 2017 test-dev, and Youtube-VOS 2019 validation. The quantitative comparison on DAVIS 2016 validation is listed in Table 2.

We can see that without BL30K [3] for pretraining and MOSE [18] for main training, our ViT-B DeVOS-L achieves state-of-the-art performance scoring **87.7%** $\mathcal{J}\&\mathcal{F}$ on DAVIS 2017 validation set, **82.9%** $\mathcal{J}\&\mathcal{F}$ on DAVIS 2017 test set and **86.3%** $\mathcal{J}\&\mathcal{F}$ on Youtube-VOS 2019 validation. The integration of the MOSE dataset further enhances our metrics, resulting in improved performance: **88.1%** $\mathcal{J}\&\mathcal{F}$ **83.0%** $\mathcal{J}\&\mathcal{F}$, **86.6%** $\mathcal{J}\&\mathcal{F}$ on the DAVIS 2017 validation test-dev, and on the Youtube-VOS 2019 validation.

The DeVOS-B model exhibits robust performance on the principal benchmark, scoring **86.1%** $\mathcal{J}\&\mathcal{F}$ on DAVIS 2017 validation set, **81.0%** $\mathcal{J}\&\mathcal{F}$ on DAVIS 2017 test set and **85.2%** $\mathcal{J}\&\mathcal{F}$ on Youtube-VOS 2019 validation. Despite employing a less complex memory mechanism in contrast to [12], omitting the direct injection of instance information as [5], and foregoing hierarchical propagation like [11], our method achieves commendable outcomes in both accuracy and, notably, FPS. This highlights the efficacy of incorporating multi-scale matching and motion-guided attention, as it contributes to the enhancement of matching performance.

Qualitative comparison Fig. 4 displays the qualitative comparison of our method with recent state-of-the-art approaches. As shown by the gold-fish sequence, our approach demonstrates superior performance under complex shape and appearance changes. Additionally, mbike-trick sequence demonstrates that our design results in strong performance under rapid motion. These findings highlight the effectiveness of the proposed approach in handling various challenging conditions.

4.3. Discussion

Training with MOSE 2023 MOSE 2023 [18] (Co**M**plex video **O**bject **SE**gmentation) is a novel VOS benchmark featuring extreme scenarios of the video sequence which are not handled good enough by existing VOS methods. The main features of introduced videos include a large number of crowded and similar objects, heavy occlusions by similar-looking objects, extremely small-scale objects, and reference masks covering only a small region of the whole object.

With adopting MOSE 2023 as training data, performance on the classic benchmarks experiences only a small boost (Table 1), likely because they don't feature any similar extreme scenarios. However, DAVIS and Youtube-VOS focus on circumstances with a large number of object classes and classes unseen during training, along with a wide variety of challenging environments, while MOSE 2023 lacks

Table 3. Ablation study. The experiments are based on DeVOS-B. MS: multi-scale matching. ADVA: adaptive deformable video attention. FP: flow priors to offset prediction in ADVA. QK: query-key flow enhancement. N_k: number of offsets per head / scale in deformable attention. Iters: number of flow refinement iterations of GMA. ω: scale embedding. θ: offset normalization. Note: in the final configuration QK-flow is used only in DeVOS-L.

(a) Multi-scale matching

MS	ADVA	D_{17V}	D_{17T}	Y_{19}	#param	FPS
✗	✗	84.7	79.2	83.7	35.4M	32.4
✓	✗	85.2	80.5	84.4	38.1M	12.9
✓	✓	86.1	81.0	85.2	40.3M	36.7

(b) Motion injection

FP	QK	D_{17V}	D_{17T}	Y_{19}	#param	FPS
✗	✗	84.9	79.6	84.2	31.1M	52.1
✓	✗	86.1	81.0	85.2	40.3M	36.7
✓	✓	86.5	81.2	85.3	40.4M	29.4

(c) Number of offsets

N_k	D_{17V}	Y_{19}	FPS
2	82.5	81.8	37.1
4	86.1	85.2	36.7
6	86.3	84.9	36.1
8	86.0	84.6	35.5

(d) Optical flow iters

Iters	D_{17V}	Y_{19}	FPS
0	81.3	81.0	45.5
1	85.5	84.9	41.4
4	86.1	85.2	36.7
12	86.3	85.3	24.8

(e) Scale emb & offset norm

ω	θ	D_{17V}	Y_{19}
✗	✗	84.9	84.1
✓	✗	85.0	84.3
✗	✓	85.8	85.0
✓	✓	86.1	85.2

such flexibility. Wrapping up, even minor improvements on classic benchmarks while training with MOSE 2023 indicate the high robustness and performance capacity of the proposed method. The quantitative comparison with other methods on MOSE 2023 validation is studied in Supplementary.

Impact of multi-scale matching We argue that matching conducted solely on 1/16 of the input resolution does not convey enough spatial information and fine-grained details to perform instance discrimination effectively. This limitation becomes particularly crucial when dealing with overlapping objects that share a similar appearance. Conversely, the short-term branch can leverage smaller feature map resolutions, specifically 1/8. To validate these hypotheses, we remove multi-scale matching and evaluate the performance of the resulting architecture in Table 3a. Multi-scale matching increases $\mathcal{J}\&\mathcal{F}$ by **0.5%** and ADVA matching further boosts the performance by **0.9%** $\mathcal{J}\&\mathcal{F}$ while featuring $\times 3$ run-time speed boost on multi-scale. The importance of the number of sampled offsets per attention head and scale is studied in Table 3c. Selecting $N_k = 4$ provides optimal performance-efficiency tradeoff. Scale embedding and offset normalization drastically improve training stability thus lead to better final performance, which is reflected in Table 3e.

Impact of optical flow guidance We assert that to make matching emphasize semantic features and instance discrimination it is necessary to inject global motion understanding prior to the matching process. To accomplish this, we enhance offset prediction with optical flow. Additionally, we study the effect of QK-flow, which directly injects motion information into the query and value feature maps. We argue that this ensures strong cycle consistency. From Table 3b, we can see that removing QK-flow results in a reduction of **0.4%** $\mathcal{J}\&\mathcal{F}$ on the DAVIS 2017 Validation set. Additionally, removing optical flow-based offset prediction results in a reduction in $\mathcal{J}\&\mathcal{F}$ by **1.2%**.

Impact of ViT backbone To further evaluate the impact of ViT-backbone, we train the same architecture but with Swin-B [39] transformer used as the backbone. This results in decrease in $\mathcal{J}\&\mathcal{F}$ by **0.2%**. We leave evaluation whether this small improvement comes from backbone architecture or SAM [32] pre-training for further research.

Limitations One practical limitation is that the framework depends on a pre-trained optical flow estimator. We believe, though, that it is quite common that both optical flow and VOS are required simultaneously. Moreover, our approach works with different flow estimating architectures thus provides flexibility of actual choice (without need of retraining). Ablation on the number of optical flow iterations of GMA [24] (Table 3d) shows that quality of optical flow is not crucial in the overall performance of our framework and thus any method performing good enough would work fine.

5. Conclusion

This paper proposes DeVOS (Deformable VOS), an architecture that incorporates adaptive deformable video attention. DeVOS combines memory-based matching with motion-guided propagation, resulting in robust matching under challenging appearance changes and strong temporal consistency. DeVOS achieves state-of-the-art performance while maintaining top-rank FPS.

References

[1] 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.

[2] H. K. Cheng, Y.-W. Tai, and C.-K. Tang, "Rethinking space-time networks with improved memory coverage for efficient video object segmentation," in *NeurIPS*, 2021.

[3] H. K. Cheng, Y.-W. Tai, and C.-K. Tang, "Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion," in *CVPR*, 2021.

[4] H. Seong, S. W. Oh, J.-Y. Lee, S. Lee, S. Lee, and E. Kim, "Hierarchical memory matching network for video object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[5] 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].

[6] H. Xie, W. Wang, X. Li, L. Xie, Y. Zhang, and Q. Tian, "Rmnet: Equivalently removing residual connection from networks," *arXiv preprint arXiv:2111.00687*, 2021.

[7] Z. Yang, Y. Wei, and Y. Yang, "Collaborative video object segmentation by foreground-background integration," in *European Conference on Computer Vision*, Springer, 2020, pp. 332–348.

[8] Z. Yang, Y. Wei, and Y. Yang, "Collaborative video object segmentation by multi-scale foreground-background integration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021. DOI: 10.1109/TPAMI.2021.3081597.

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

[10] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[11] Z. Yang and Y. Yang, "Associating objects with transformers for video object segmentation," *arXiv preprint arXiv:2210.09782*, 2022.

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

[13] Y. Liang, X. Li, N. Jafari, and J. Chen, "Video object segmentation with adaptive feature bank and uncertain-region refinement," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 3430–3441. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/234833147b97bb6aed53a8f4f1c7a7d8-Paper.pdf.

[14] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[15] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 4794–4803.

[16] 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.

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

[18] 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.

[19] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[20] M. J. Black and P. Anandan, "A framework for the robust estimation of optical flow," in *Proceedings of the 4th International Conference on Computer Vision*, IEEE, 1993, pp. 231–236.

[21] A. Bruhn, J. Weickert, and C. Schnörr, "Lucas/kanade meets horn/schunck: Combining local and global optic flow methods," *International journal of computer vision*, vol. 61, no. 3, pp. 211–231, 2005.

[22] D. Sun, S. Roth, and M. J. Black, "A quantitative analysis of current practices in optical flow estimation and the principles behind them," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 115–137, 2014.

[23] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *European Conference on Computer Vision*, Springer, 2020, pp. 402–419.

[24] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. Hartley, "Learning to estimate hidden motions with global motion aggregation," *arXiv preprint arXiv:2104.02409*, 2021.

[25] S. Bai, Z. Geng, Y. Savani, and J. Z. Kolter, "Deep equilibrium optical flow estimation," *arXiv preprint arXiv:2204.08442*, 2022.

[26] Z. Huang, X. Shi, C. Zhang, *et al.*, "Flowformer: A transformer architecture for optical flow," *arXiv preprint arXiv:2203.16194*, 2022.

[27] A. Jaegle, S. Borgeaud, J.-B. Alayrac, *et al.*, "Perceiver io: A general architecture for structured inputs & outputs," *arXiv preprint arXiv:2107.14795*, 2021.

[28] V. Fedynyak, Y. Romanus, O. Dobosevych, I. Babin, and R. Riazantsev, "Global motion understanding in large-scale video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jun. 2023, pp. 3152–3161.

[29] O. Russakovsky, J. Deng, H. Su, *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015. DOI: 10.1007/s11263-015-0816-y.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[32] A. Kirillov, E. Mintun, N. Ravi, *et al.*, "Segment anything," *arXiv:2304.02643*, 2023.

[33] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944. DOI: 10.1109/CVPR.2017.106.

[34] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, *Feature pyramid networks for object detection*, 2017. arXiv: 1612.03144 [cs.CV].

[35] Y. Wang, Z. Xu, X. Wang, *et al.*, "End-to-end video instance segmentation with transformers," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021.

[36] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, "The 2017 davis challenge on video object segmentation," *arXiv:1704.00675*, 2017.

[37] S. Chandra and I. Kokkinos, *Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs*, 2016. arXiv: 1603.08358 [cs.CV].

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

[39] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.