From ViT Features to Training-free Video Object Segmentation via Streaming-data Mixture Models

Supplemental Material

Roy Uziel
Ben-Gurion University of the Negev, Israel
uzielr@post.bgu.ac.il

Or Dinari
Ben-Gurion University of the Negev, Israel
dinari@post.bgu.ac.il

Oren Freifeld
Ben-Gurion University of the Negev, Israel
orenfr@cs.bgu.ac.il

Contents

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

A vMF Mixtures

This section contains the expressions for **vMF Normalizer** and the **Approximation of** $\log C_d(\tau)$.

The **vMF Normalizer** is

$$C_d(\tau) = \frac{\tau^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\tau)} \tag{1}$$

where $I_{d/2-1}$ is the modified Bessel function of the first kind with order $d/2-1$.

The well-known **approximation of** $\log C_d(\tau)$ [5] is

$$\begin{aligned}\log C_d(\tau) \approx & \frac{d-1}{4} \log \left(\frac{d-1}{2} + \sqrt{\left(\frac{d-1}{2}\right)^2 + \tau^2} \right) \\ & + \frac{d-1}{4} \log \left(\frac{d-1}{2} + \sqrt{\left(\frac{d+1}{2}\right)^2 + \tau^2} \right) \\ & - \frac{1}{2} \sqrt{\left(\frac{d+1}{2}\right)^2 + \tau^2} \,.\end{aligned} \tag{2}$$

B Implementation Details

This section offers insights into the implementation details of our method.

Feature extraction: Our feature extraction utilized the official weights of XCiT-small with patch size of 8, trained in accordance with the self-supervised learning approach detailed in [2]. For DAVIS-2017, we evaluated images at a 480p resolution, following the standard practice. To encourage accurate clustering and avoid numerical errors caused by small clusters, we upsampled the features by a factor of two through bilinear interpolation.

vMF distribution: The implementation of our vMF distribution is based on [3].

PAC-CRF Model: Our PAC-CRF model is comprised of two parallel 5×5 Position-Adaptive Convolution (PAC) kernels with dilation factors of 64 and 16. The selection of the PAC kernel size and dilation factors was guided by the objective of capturing both local and global context information effectively. We determined that these specific kernel sizes and dilation factors strike a balance between the receptive field size and computational efficiency, resulting in optimal performance.

Hyperparameters: The number of components in each mixture, denoted as $(k_s)_{s=1}^S$, was determined based on the size and complexity of the regions. To adaptively adjust the number of components, we evenly distributed k_s between the minimal and maximal values based on the value of S. Specifically, for objects, we evenly distributed between 2 (the minimal) and the square root of the number of pixels in the object mask in the first frame. Similarly, for the background, we evenly distributed between 50 (the minimal value) and the square root of the number of pixels in the background mask in the first frame. This approach ensures that the number of components in the mixture aligns with the complexity of the scene, allowing our model to effectively capture diverse appearances. Additionally the weight w_ρ for positional embeddings was set to 15.

To enforce spatial constraints while ensuring computational efficiency, we utilized an indicator function with thresholds $(r_s)_{s=1}^S$. This function determines the membership of a pixel by evaluating the similarity between its appearance and the learned appearance of the model. The thresholds r_s are evenly distributed in the range of 0.68 to 0.78, depending on the number of mixtures. Increasing the number of mixtures requires tighter similarity constraints, enabling more precise capture of finer details in the segmentation process.

To enable efficient processing and effective model adaptation, we stored and utilized the sufficient statistics from the last previous 15 frame. That is, at time t, we are using the sufficient statistics from times $(t-15, t-14, \ldots, t-1, t)$. These statistics provide valuable information about the previous frames, allowing our method to maintain contextual knowledge and adapt to dynamic changes in the video. By retaining this limited-legnth history of sufficient statistics, we strike a balance between memory utilization and the ability to capture long-term dependencies within the video sequence.

Furthermore, for computational efficiency, we applied spatial constraints by considering a fixed rectangle of size 18×36 for each pixel. We observed that removing this constraint did not impact the performance of our method. This finding demonstrates that our current spatial constraints effectively capture the relevant information within the defined radius.

C Results

This section presents additional qualitative results to complement the main paper.

For video demonstrations, please visit the project page: https://github.com/BGU-CS-VIL/Training-Free-VOS.

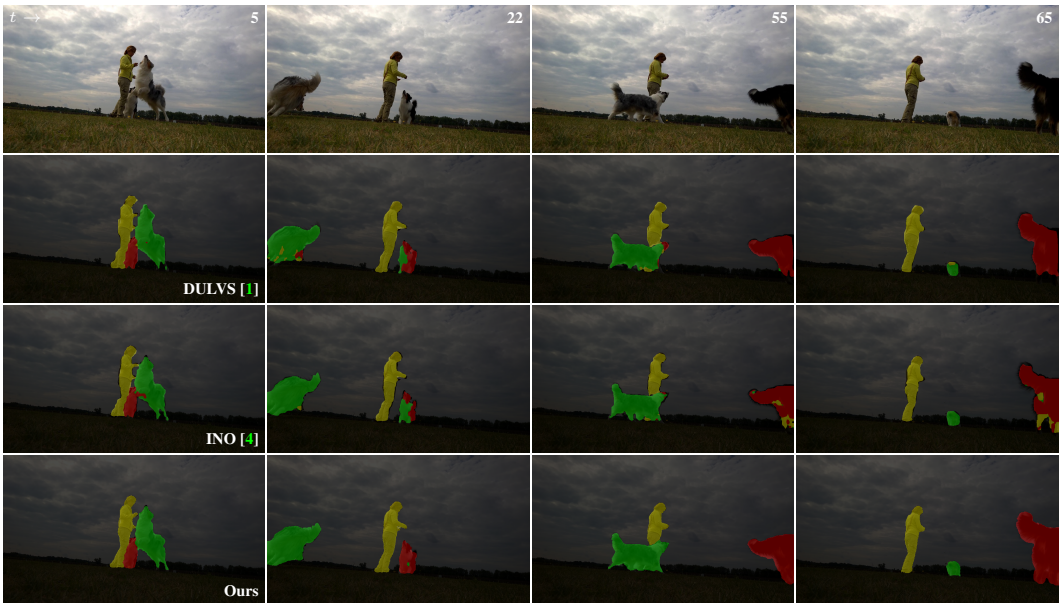

Figure 1: **Qualitative examples on DAVIS-2017**

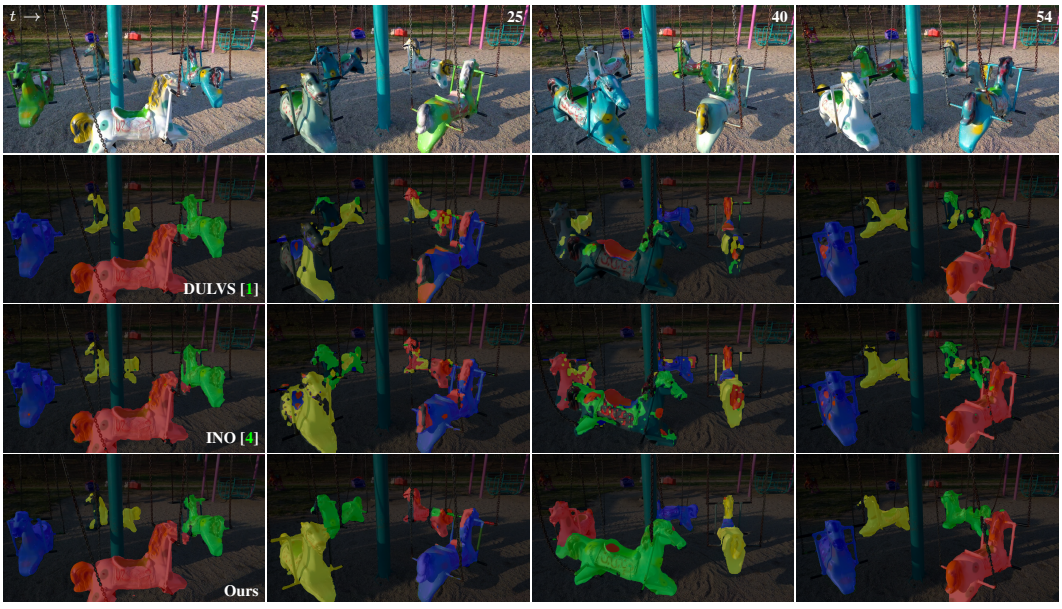

Figure 2: **Qualitative examples on DAVIS-2017**

Figure 3: **Qualitative examples on DAVIS-2017**

Figure 4: **Qualitative examples on DAVIS-2017**

Figure 5: **Qualitative examples on DAVIS-2017**

D Examples of Failure Cases

The figure below shows typical failure cases of our proposed method. For more details about the associated limitations (related to either re-identification or motion blur), see the paper.

Figure 6: **Failure cases**. Rows 1 and 3 display two frames from different scenarios in the YouTube-VOS 2018 dataset, while rows 2 and 4 show our segmentation predictions for each scenario, respectively.

References

[1] Nikita Araslanov, Simone Schaub-Meyer, and Stefan Roth. Dense unsupervised learning for video segmentation. *NeurIPS*, 2021. 4, 5, 6

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3

[3] Minyoung Kim. On pytorch implementation of density estimators for von mises-fisher and its mixture. In *arXiv preprint*, 2021. 3

[4] Xiao Pan, Peike Li, Zongxin Yang, Huiling Zhou, Chang Zhou, Hongxia Yang, Jingren Zhou, and Yi Yang. In-n-out generative learning for dense unsupervised video segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 4, 5, 6

[5] Tyler R Scott, Andrew C Gallagher, and Michael C Mozer. von mises-fisher loss: An exploration of embedding geometries for supervised learning. In *ICCV*, 2021. 2