MobileVOS: Real-Time Video Object Segmentation Contrastive Learning meets Knowledge Distillation

Roy Miles* Mehmet Kerim Yucel Bruno Manganelli Albert Saà-Garriga
Samsung Research UK

1. Supplementary Material

1.1. Qualitative results

We provide an extensive qualitative evaluation of both the ResNet and MobileNet variants. Firstly, we explore the performance of MobileVOS on an out-of-domain video and compare its predictions to XMem, whereby we observe that, although some segmentation errors occur, they are far less significant. Secondly, we provide real-time predictions of a single object, given on a mobile device. We expose this object to severe occlusions to highlight the robustness of our models in the wild. Finally, we show the generality of SVOS in its application with video inpainting.

Out-of-domain The mask predictions given in figure 1 demonstrate the robustness of MobileVOS to domain shifts, unseen classes, and camera shot changes. We compare these predictions to those given by XMem and observe 3 distinct failure modes that are unique to each of these model, where these failure modes are tied to the underlying architectures and memory models used. Although some segmentation errors occur only on MobileVOS, and not XMem, we expect that this is simply a trade-off imposed by the smaller network capacity, and the other types of failure modes (observed only in XMem) are much more detrimental.

1. **Similar features** The second frame shows some poor segmentation on the wrong object, which we attribute to the smaller network capacity that is unable to learn sufficiently discriminative features. This is not observed in XMem due to the much larger backbones.

2. **Shot changes** XMem can fail to segment the correct objects under camera shot changes since the model is matching features to a long sequence of intermediate frames which do not include the main object.

3. **Drift** After XMem makes this first mistake, the model then begins to drift. MobileVOS does not suffer from this problem due to only storing the first and most recent frames/masks in memory. This drift leads to XMem poorly segmenting later frames in the video, including segmenting the wrong object.

We have included the full length videos alongside this supplementary document. This video example highlights limitations of the YouTube and DAVIS evaluation datasets, which do not consider domain shifts or camera shot changes.

Poly Loss	DAVIS 2016	DAVIS 2017
with	89.8	80.1
without	89.8	79.9

Table 1. Evaluating the impact of the poly loss on both the DAVIS 2016 and DAVIS 2017 datasets, where $\epsilon = 1$ and the query encoder uses a MobileNetV2 backbone wo/ ASPP.

On-device Long term occlusion We demonstrate the robustness of our most efficient MobileVOS model (MobileNet V2 wo/ ASPP) under severe occlusion in real-time and on a mobile device. The results can be seen in the attached OnDeviceOcclusion.mp4 video and show almost no segmentation errors.

Image inpainting A practical use case of SVOS is video inpainting. This task often requires per-frame masks, which indicate the areas to be inpainted. We use the MobileVOS ResNet18 variant to generate these per-frame masks, and then perform inpainting using FuseFormer [3]. We use the original operating resolution of FuseFormer (240p), where the segmentation masks are downsized to this resolution accordingly. In videos with multiple objects, we merge the masks into a single binary mask and use this as the inpainting input. We show an example of this application on a video provided in the YouTube validation split (see the attached video or figure 3). Despite significant object movement, our model is still able to provide highly accurate per-frame masks that lead to visually appealing inpainting results.

*Work done while being an intern at Samsung Research UK

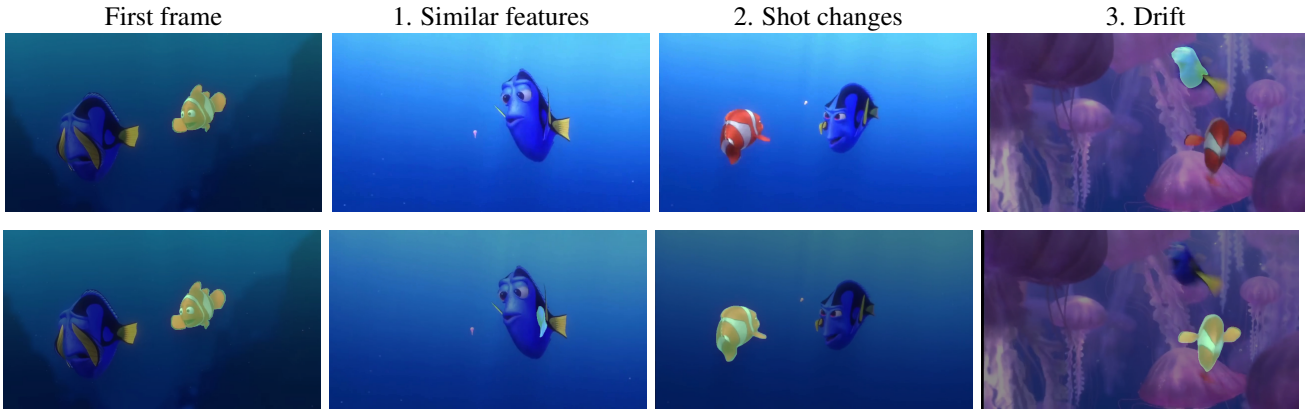

Figure 1. Comparison to XMem on a long out-of-domain single object segmentation task. Top row are the predictions by XMem, while the bottom row is from the ResNet MobileVOS. We categorise and highlight 3 distinct segmentation errors that can occur. The colour shift is just an artifact of how the masks are overlayed on the frames and is unrelated to the segmentation.

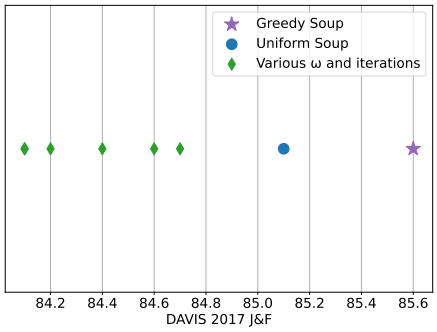

Figure 2. Model soups can improve the accuracy on both the DAVIS and YouTube datasets, without any additional inference costs.

1.2. Poly loss ablation

Table 1 shows the $\mathcal{J}\&\mathcal{F}$ results on both the DAVIS 2016 and DAVIS 2017 validation splits with and without using the additional poly loss component. In these experiments, we train the MobileNet (wo/ ASPP) backbone with no distillation or contrastive learning and observe that the poly loss can be safely removed without impacting the models performance.

1.3. Model soups

We use models soups [9] as an alternative to multi-scale inference, which is typically adopted in the SVOS literature [1]. Unlike model soups, multi-scale inference can incur significant additional inference costs due to multiple forward passes at different resolutions. Figure 2 shows the accuracy of a few checkpoints with varying values of ω and at different iterations of training. By simply averaging the weights of all of these models, we achieve a significant increase in the DAVIS 2017 validation accuracy. However, by adopting a greedy selection process, we are able to achieve a much more significant increase. One noticeable observation from this process is that no additional data is needed for selecting the model to be included in the soup - they are simply conditionally added based on the observed training accuracy.

1.4. Background - Kernel Perspective

Rényi's α-entropy [5] of order $\alpha \in (0,1) \cup (1,\infty)$ provides a natural extension of Shannon's entropy. Consider a random variable X with probability density function (PDF) $f(x)$ in a finite set $\mathcal{X}$, the α-entropy $\mathbf{H}_\alpha(X)$ is defined as:

$$\mathbf{H}_\alpha(f) = \frac{1}{1-\alpha} \log_2 \int_{\mathcal{X}} f^\alpha(x) dx \tag{1}$$

Where the limit as $\alpha \to 1$ is the well-known Shannon entropy. [7,8] propose a set of quantities that closely resemble Rényi's entropy and omit the need for evaluating the underlying probability distributions. These information quantities are estimated directly from the data and are based on the theory of infinitely divisible matrices. Their usage leverages the representational power of reproducing kernel Hilbert spaces (RKHS), which is a concept that has been widely studied and adopted in classical machine learning. These estimators have been successfully applied in the context of knowledge distillation for image classification, reading comprehension, and binary network classification [4].

For completeness, we now provide definitions of these entropy-based quantities and their connections with positive semidefinite matrices. This idea then leads to a multivariate extension using Hadamard products, from which conditional and mutual information can be defined. For brevity, we omit the proofs and connections with Rényi's axioms, which can be found in [7,8].

Definition 1: Let $X = \{x^{(1)}, \dots x^{(n)}\}$ be a set of n data points of dimension d and $\kappa : X \times X \to \mathbb{R}$ be a real-valued positive definite kernel. The Gram matrix $\mathbf{K}$ is obtained

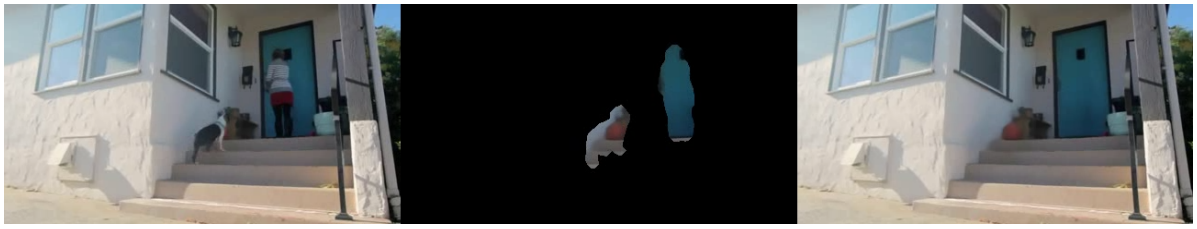

Figure 3. Using MobileVOS in conjunction with video in-painting to remove selecting objects.

from evaluating κ on all pairs of examples, that is $K_{ij} = \kappa(x^i, x^j)$. The matrix-based analogue to Rényi's α-entropy for a normalized positive definite (NPD) matrix $\mathbf{A}$ such that $tr(\mathbf{A}) = 1$, can be given by the following functional:

$$\mathbf{S}_\alpha(\mathbf{A}) = \frac{1}{1-\alpha}\log_2(tr(\mathbf{A}^\alpha)) \tag{2}$$

$$= \frac{1}{1-\alpha}\log_2\left[\sum_{i=1}^{n}\lambda_i(\mathbf{A}^\alpha)\right] \tag{3}$$

where $\mathbf{A}$ is the kernel matrix $\mathbf{K}$ normalised to have a trace of 1 and $\lambda_i(\mathbf{A})$ denotes its i-th eigenvalue. This estimator can be seen as a statistic on the space computed by the kernel κ, while also satisfying useful properties attributed to entropy.

Definition 2: Let X and Y be two sets of data points. After computing the corresponding Gram matrices $\mathbf{A}$ and $\mathbf{B}$, the joint entropy is then given by:

$$\mathbf{S}_\alpha(\mathbf{A}, \mathbf{B}) = \mathbf{S}_\alpha\left(\frac{\mathbf{A}\circ\mathbf{B}}{tr(\mathbf{A}\circ\mathbf{B})}\right) \tag{4}$$

where $\circ$ denotes the Hadamard product between two matrices. Using these two definitions, the notion of conditional entropy and mutual information can be derived. We focus on the mutual information, which is given by:

$$\mathbf{I}_\alpha(\mathbf{A};\mathbf{B}) = \mathbf{S}_\alpha(\mathbf{A}) + \mathbf{S}_\alpha(\mathbf{B}) - \mathbf{S}_\alpha(\mathbf{A},\mathbf{B}) \tag{5}$$

1.5. Decomposing the representation loss

In this section, we provide an intricate connection between the proposed loss, mutual information, and contrastive learning. By bridging between these two training regimes, we find that models can benefit from minimising a linear weighting of these two objectives. We hope that this abstract lense can provide additional insights into the training dynamics for learning, and specifically in the context of very practical dense prediction tasks.

Relating $\mathcal{L}_{repr}$ to mutual information *More formally, in the case where $\omega = 1$, we show that minimising $\mathcal{L}_{repr}$ is equivalent to maximising the pixel-wise mutual information between the student and teacher representations.*

Given $\mathbf{A}$ is a real symmetric matrix, then $\|\mathbf{A}\|_F^2 = tr(\mathbf{A}\mathbf{A}^T) = \sum_{i=1}^{n}\lambda_i(\mathbf{A^2})$. This follows from the definition of the Frobenius norm of a matrix, $\|\mathbf{A}\|_F^2 = \sum_{ij}\mathbf{A}_{ij}^2$. The trace term can be expanded as follows $tr(\mathbf{A}\mathbf{A}^T) = \sum_i\left(\mathbf{A}\mathbf{A}^T\right)_{ii} = \sum_i\sum_j\mathbf{A}_{ij}\mathbf{A}_{ji}$. Since $\mathbf{A}$ is symmetric, $\mathbf{A}_{ij} = \mathbf{A}_{ji}$ and thus $tr(\mathbf{A}\mathbf{A}^T) = \sum_i\sum_j\mathbf{A}_{ij}^2 = \|\mathbf{A}\|_F^2$. Finally, the equality between the trace of a matrix and the sum of eigenvalues is a known relation in linear algebra.

The representations, $\mathbf{Z}_S$ and $\mathbf{Z}_T$, are $L2$ normalised and thus the correlation matrices $\mathbf{C}$ will have 1s along their leading diagonal. These matrices are real and symmetric, which allows use to use the relation derived above.

The representation loss $\mathcal{L}_{repr}$ can be decomposed into the difference of two information-theoretic quantities, namely the entropy and joint entropy.

$$\mathcal{L}_{repr} = \frac{1}{|\,\mathbf{C}_s\,|}\left(\log_2\|\mathbf{C}_s\|^2 - \log_2\|\mathbf{C}_s\odot\mathbf{C}_t\|^2\right) \tag{6}$$

$$= \frac{1}{|\,\mathbf{C}_s\,|}\left(-\mathbf{S}_2(\mathbf{Z}_S) + \mathbf{S}_2(\mathbf{Z}_S;\mathbf{Z}_T)\right) \tag{7}$$

Equation 7 follows from 6 using the definitions for the entropy estimators in equation 3 and 4 with $\alpha = 2$. Maximising the mutual information can be given as follows:

$$\mathcal{L}_{mi} = -\mathbf{I}_2(\mathbf{Z}_S;\mathbf{Z}_T) \tag{8}$$

$$= \cancel{-\mathbf{S}_2(\mathbf{Z}_T)} - \mathbf{S}_2(\mathbf{Z}_S) + \mathbf{S}_2(\mathbf{Z}_S;\mathbf{Z}_T) \tag{9}$$

Where the first entropy term can be omitted since no gradients flow through the teachers representation. From an optimisation perspective, these two losses are then equivalent. The only distinction is in the pixel-wise sampling strategy, where we opt to select only the boundary pixels, which leads to much faster model convergence.

Relating $\mathcal{L}_{repr}$ to contrastive learning *In the case where $\omega = 0$, minimising $\mathcal{L}_{repr}$ is a pixel-wise contrastive objective.*

The loss can be deconstructed into the sum of positive and negative pixel-wise pairings. In the case where $\omega = 0$, $\mathcal{C}_{ty} = \mathcal{C}_y = \mathbf{Y}\mathbf{Y}^T$.

$$\mathbf{C}_y = \left(\mathbf{Y}\mathbf{Y}^T\right)_{ij} = \begin{cases} 1 & j \in \mathcal{P}_i \\ 0 & j \in \mathcal{N}_i \end{cases} \tag{10}$$

where $\mathcal{P}_i, \mathcal{N}_i$ denote the set of positive and negative indices for the i-th sample. The loss is then decomposed as follows.

$$\mathcal{L}_{repr} = \frac{1}{\mid \mathbf{C}_s \mid} \left(\log_2 \|\mathbf{C}_s\|^2 - \log_2 \|\mathbf{C}_s \odot \mathbf{C}_y\|^2 \right) \tag{11}$$

$$= \frac{1}{\mid \mathbf{C_s} \mid} \log_2 \sum_i \left(\sum_{j \in \mathcal{P}_i} (\mathbf{C}_s)_{ij}^2 + \sum_{j \in \mathcal{N}_i} (\mathbf{C}_s)_{ij}^2 \right) \tag{12}$$

$$- \frac{1}{\mid \mathbf{C_s} \mid} \log_2 \sum_i \sum_{j \in \mathcal{P}_i} (\mathbf{C}_s)_{ij}^2$$

This loss can be further simplified by the log identity $log(a) - log(b) = log(a/b)$.

$$\mathcal{L}_{repr} = \frac{1}{\mid \mathbf{C}_s \mid} \log_2 \frac{\sum_i \left(\sum_{j \in \mathcal{P}_i} (\mathbf{C}_s)_{ij}^2 + \sum_{j \in \mathcal{N}_i} (\mathbf{C}_s)_{ij}^2 \right)}{\sum_i \sum_{j \in \mathcal{P}_i} (\mathbf{C}_s)_{ij}^2} \tag{13}$$

$$= -\frac{1}{\mid \mathbf{C}_s \mid} \log_2 \frac{\sum_i \sum_{j \in \mathcal{P}_i} (\mathbf{C}_s)_{ij}^2}{\sum_i \left(\sum_{j \in \mathcal{P}_i} (\mathbf{C}_s)_{ij}^2 + \sum_{j \in \mathcal{N}_i} (\mathbf{C}_s)_{ij}^2 \right)} \tag{14}$$

$$= -\frac{1}{\mid \mathbf{C}_s \mid} \log_2 \sum_i \frac{\sum_{j \in \mathcal{P}_i} (\mathbf{C}_s)_{ij}^2}{\sum_k (\mathbf{C}_s)_{ik}^2} \tag{15}$$

The original supervised contrastive loss [2] is given as follows.

$$\mathcal{L}_{SupCon} = -\sum_i \log \frac{1}{\mid \mathcal{P}_i \mid} \frac{\sum_{j \in \mathcal{P}_i} sim(z_i, z_j)}{\sum_k sim(z_i, z_j)} \tag{16}$$

where $i, k \in \{1 \ldots |\mathbf{C}_s|\}$ index the set of all sampled pixels. In the case where we define $sim(z_i, z_j)$ to be the cosine similarity between the two vectors z_i and z_j, these two losses are very similar. The only distinction between the two lies in switching the position of the normalisation and summation with respect to the logarithm. It is also worth noting that we use base 2 for the logarithm, as is convention in the information theory literature. In essence, the numerator in this loss pushes positive terms together, while the denominator repels negative pairs.

1.6. Comparison with other distillation methods

We trained the ResNet model with two different distillation losses and observed a significant drop in attainable performance on the DAVIS16 benchmark, which can be seen in figure 4. Our method outperforms others by over 1 $\mathcal{J}\&\mathcal{F}$.

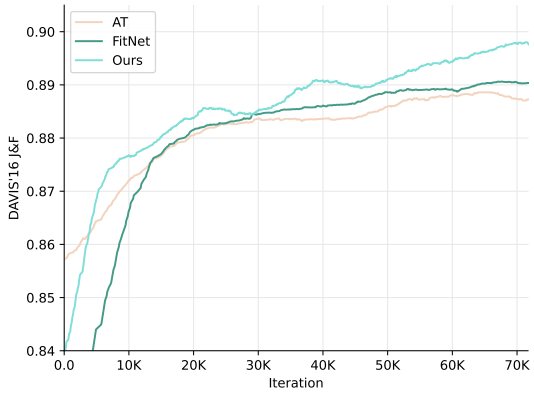

Figure 4. Comparing the performance of the ResNet model trained with hints (FitNets) [6] and Attention Transfer [10].

1.7. Loss ablation with the MobileNet backbone

Additional experiments demonstrating the effectiveness of our proposed loss on the MobileNet architectures are given in table 2, whereby we observe a consistent improvement in $\mathcal{J}\&\mathcal{F}$ across both the DAVIS16 and DAVIS17 datasets with and without ASPP.

Model	distillation	DAVIS16	DAVIS17
wo/ ASPP	✗	89.2	80.5
wo/ ASPP	✓	**90.1**	**81.8**
w/ ASPP	✗	89.6	81.6
w/ ASPP	✓	**90.5**	**82.2**

Table 2. Evaluating the effectiveness of our proposed distillation loss on the MobileNet backbone architecture.

References

[1] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *NeurIPS*. 2

[2] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *NeurIPS*, 2020. 4

[3] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in

transformers for video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14040–14049, 2021. 1

[4] Roy Miles, Adrian Lopez Rodriguez, and Krystian Mikolajczyk. Information Theoretic Representation Distillation. *BMVC*, 2022. 2

[5] Alfréd Rényi. On Measures of Entropy and Information. *Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability*, 1960. 2

[6] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints For Thin Deep Nets. *ICLR*, 2015. 4

[7] Luis Gonzalo Sanchez Giraldo, Murali Rao, and Jose C. Principe. Measures of entropy from data using infinitely divisible Kernels. *IEEE Transactions on Information Theory*, 2015. 2

[8] Paul L. Williams and Randall D. Beer. Nonnegative Decomposition of Multivariate Information. 2010. 2

[9] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *ICML*. PMLR, 2022. 2

[10] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2019. 4