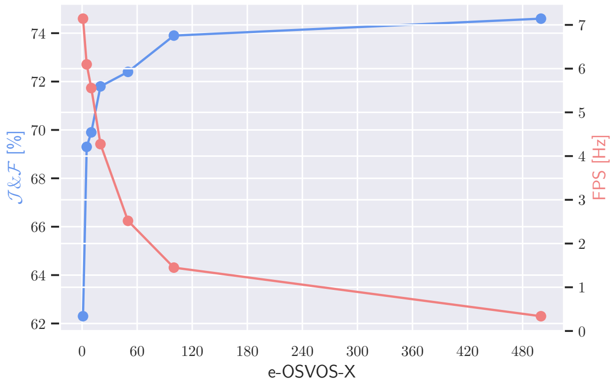
We thank all reviewers for their constructive and valuable feedback and are delighted to receive an overall positive acceptance of our work. Furthermore, we will integrate all changes according to your suggestions and questions.

Effectiveness (R1, R2) In the context of our work, efficiency refers to the runtime of the test time optimization which directly translates to inference time. In order to achieve high segmentation performance previous fine-tuning approaches suffer from unfeasible high runtime due to many fine-tuning epochs (up to 1000). Our approach reduces the number of epochs drastically by meta learning the initialization and learning rates. Furthermore, in Figure 1 (which will be included in the supplementary) we illustrate the performance gains and invertedly reduction in FPS of our e-OSVOS framework for increasing number of fine-tuning epochs. For many fine-tuning epochs, the actual inference time on the frames is neglectable with respect to the fine-tuning. For longer sequences the amortization of the fine-tuning time is higher, however, these sequences usually face additional challenges due to changing object appearance which can be tackled by the presented online adaptation.

State-of-the-art (R1, R4) We understand any potential doubt of our overall impact on the field of VOS as we merely provide a new state-of-the-art for fine-tuning approaches. However, considering publications of the recent years, the VOS community has collectively deemed fine-tuning as unfeasible. Therefore, we hope our demonstration of an efficient fine-tuning approach enabled through meta learning will have a substantial impact and ignite future research based on the release of our code base.

Figure 1: Evaluation on the DAVIS 2017 validation set.



Learning (neuron-level) learning rates (R1) We predict test time learning rates for fully-connected and convolutional layers, each consisting of neurons with a weight vector and a scalar bias. To account for convolutional neurons with a spatial weight tensor (also referred to as kernels), we will rename *weight vector* to *weight tensor*. (R3) Learning and deriving a gradient with respect to the learning rates $\boldsymbol{\lambda}$ is analogous to the model initialization $\boldsymbol{\theta}_f^0$. Hence, we formulate the meta optimization of the optimization g for the joint set of its parameters $\boldsymbol{\theta}_g = \{\boldsymbol{\theta}_f^0, \boldsymbol{\lambda}\}$. The gradient flows from the loss (Equation (4)) to each of the T parameter updates (Equation (5)) of the inner fine-tuning loop. The connection between the inner and outer optimization is also illustrated in Algorithm 1 of the supplementary. As the SGD update consists only of differentiable operations, gradients with respect to the learning rates can be derived analogous to the derivation for the initial parameters. It should be noted, that these gradients are with respect to a different loss ($\mathcal{L}_{seg}(\mathcal{D}_{test}, \boldsymbol{\theta}_f^T)$) as the ones of the inner gradient ($\mathcal{L}_{seg}(\mathcal{D}_{train}, \boldsymbol{\theta}_f^t)$). It is a common approach to meta learn the initialization and learning rate(s) jointly. We refer the reader to [2, 1] for further insights. **(R2)** At test time, the set of learning rates does not change and is the same for all sequences. However, it is interesting to observe which neurons are updated with particularly small, e.g., biases of last layers, or large, e.g., FC6 of the box head, learning rates. The FC6 layer of the box head prepares the spatial bounding box features for the regression and classification heads and benefits from a strong adaption to each individual given object. To further illustrate e-OSVOS, we will add a summarized visualization of the overall more than 20000 neuron-level learning rates to the supplementary.

Other comments/suggestions (R1) In the ablation study (first row of Table 1), we present a Mask R-CNN baseline without meta learning which was pre-trained on ImageNet, COCO segmentation, YouTube-VOS and DAVIS 2017. However, this baseline is not representative for state-of-the-art fine-tuning approaches as we omitted any additional handcrafted test time improvements. For a state-of-the-art fine-tuning approach without meta learning but with bells and whistles, e.g., online adaption, we compare to OnAVOS [3] in Table 3. **(R2)** Upon acceptance we will publish our results on the official DAVIS challenge webpage which provides a tool for visual comparison per sequence, e.g., blackswan. **(R3)** The *mitigate* in line 7 refers to *shortcomings* and we will improve the understandability of the abstract. The fine-tuning epochs in Table 1 refer to a single update with one image. However, to improve the generalization over a sequence, we train on batches of random transformations of that image. The superscript of $\boldsymbol{\theta}_f$ always implies how the parameters were optimized, e.g., for T update steps or by optimization g. In our formulation, the optimization g describes a model initialization, a set of learning rates and number of steps T. We will clarify this overloading of the superscript in the final version. Furthermore, we observed overfitting without the YouTube-VOS dataset and therefore train on a combination of all three datasets.

[1] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your MAML. In *International Conference on Learning Representations*, 2019.

[2] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conf. on Machine Learning*, ICML'17, pages 1126–1135. JMLR.org, 2017.

[3] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. In *BMVC*, 2017.