SUPPLEMENTARY MATERIAL Strike the Balance: On-the-Fly Uncertainty based User Interactions for Long-Term Video Object Segmentation

In this supplementary document we provide additional quantitative and qualitative experiments alongside insights on the current limitation and future directions for improvements.

A Additional Evaluations

A.1 Quantitative Results (Perfect Mask Initialization)

Tab. S1 reports the evaluation of semi-automatic Video Object Segmentation (sVOS) and lazy interactive Video Object Segmentation (ziVOS) methods on the LVOS validation set [8], using ground-truth annotations to indicate which object to segment in the sequence (as in sVOS). We re-evaluated each method, and compute the robustness metric $R@\tau_{IoU}$, expect for DDMemory [8] as the code is unavailable at the time of writing. Similarly to Table 1 (refer to the paper), Lazy-XMem† with only pseudo-interaction achieves competitive results to State-of-the-Art (SOTA) sVOS methods. However, by including user interactions on-the-fly to aid Lazy-XMem, we manage to improve the results robustness for the cost of 315 interactions (about 1.02% of the total number of frames in LVOS).

Table S1: Quantitative evaluation of sVOS and ziVOS methods on the LVOS validation set [8], when initialized with the ground-truth annotations (curated masks as in sVOS).

		Robustness			**User-Workload**		
Method	$\mathcal{J}\&\mathcal{F}$	R@0.1	R@0.25	R@0.5	$\mathcal{N}_{\Delta}\mathcal{I}$	$No\mathcal{I}$	$\Delta\mathcal{I}$
sVOS							
QDMN [9] (ECCV 2022)	48.2	54.0	50.1	41.5	-	-	-
XMem [3] (ECCV 2022)	53.7	54.6	51.7	41.3	-	-	-
DDMemory [8] (ICCV 2023)	60.7	-	-	-	-	-	-
DEVA [2] (ICCV 2023)	58.2	65.3	62.7	56.8	-	-	-
Cutie-base [1] (CVPR 2024)	60.3	62.9	62.0	58.3	-	-	-
Cutie-small [1] (CVPR 2024)	59.0	61.3	59.0	56.5	-	-	-
Lazy-XMem (ours)	57.2	60.3	58.5	49.6	-	-	-
ziVOS							
Rand-Lazy-XMem (ours)	60.3	66.3	64.3	58.8	5.05	320	18.2
Lazy-XMem (ours)	63.5	70.0	68.3	63.1	4.86	315	18.9

A.2 Additional Ablations

Table S2 tabulates the results when relying directly on the masked entropy $S_{\mathcal{R}_c}$ and its respective derivative $\Delta S_{\mathcal{R}_c}$ as a condition to request user help. To isolate the influence each strategy for calling the user's help, we discard the mask refiner and the pseudo interaction. We only consider user interactions and rely directly on the ground-truth annotations to correct the model's predictions, instead of the mask refiner. We can see in table Tab. S2, that both strategies enhance the robustness and the accuracy, especially when updating the memory of the sVOS baseline (XMem [3]) with the refined masks through the Interaction Driven Update (IDU). However, by issuing an interaction based on the derivative $S_{\mathcal{R}_c}$, we manage to significantly reduce the number of user calls from 787 to 327 calls

Table S2: Results for Lazy-XMem when requesting user corrections through $S_{\mathcal{R}_c}$ or $\Delta S_{\mathcal{R}_c}$ (note that for this table we discard the pseudo-interaction). We initialize each method with perfect masks. UDU denotes Uncertainty Driven Update

		Robustness			User-Workload		
Configuration	$\mathcal{J}\&\mathcal{F}$	R@0.1	R@0.25	R@0.5	$\mathcal{N}_{\Delta}\mathcal{I}$	$No\mathcal{I}$	$\Delta\mathcal{I}$
XMem [3] (baseline)	53.7	54.6	51.7	41.3	-	-	-
Call user corrections based on $S_{\mathcal{R}_c}$							
XMem + UDU	54.7	56.3	54.5	50.0	56.1	3647	1.9
XMem + UDU + IDU	63.5	67.6	66.1	61.7	12.1	787	8.5
Call user corrections based on $\Delta S_{\mathcal{R}_c}$							
XMem + UDU	55.6	58.2	56.4	51.8	7.80	507	12.6
XMem + UDU + IDU	62.9	67.8	66.2	60.9	5.05	327	18.3

B Implementation Details

For our sVOS baseline, we rely on the original weights provided by the authors of XMem [3], which is trained on the static and DAVIS 2017 training set [10]

Deep Ensemble variant: We experiment with an ensemble approach that combines three XMem models The first model is trained on the static [5] and DAVIS 2017 training set [10]. The second model (which we use as a baseline in Lazy-XMem) is trained similarly to the first model but also includes the synthetic dataset BL30K [4]. The third model is trained like the first model but with the addition of the MOSE [7] dataset.

Monte Carlo variant: We rely on spatial pooling [12] applied to the key-projection of XMem [?], with a dropout ratio of 0.2 for our Monte Carlo Dropout variant during training, which is maintained during inference. For more details, we refer the reader to the original paper [3].

Thresholds: We using the training set of the LVOS dataset [8] to identify the values for $\tau_u = 0.5$, $\tau_p = 0.2$ and $\tau_m = 0.8$.

Hardware: All experiments are performed on an Nvidia GeForce GTX 1080 Ti.

C Qualitative Results

In this section we provide qualitative results that highlight both success and failure cases whenever Lazy-XMem issues either pseudo- or user-corrections to generate a refined mask. Fig. S1 and Fig. S2 displays success and failure cases, respectively, for generating a refined mask through pseudo-corrections. Fig. S3 and Fig. S4 show results when a refined mask is generated via a simulated user-correction, as described in **??**.

We indicate a ground-truth mask in yellow, the original prediction in turquoise, the refined mask in orange or purple after a pseudo-and user-correction respectively. We mark the location of a pseudo- or user-corrections through a yellow star ☆.

For small objects, we provide a cropped version to better visualize the different predictions. In these cases, a small image of the original image is shown on the first column, surrounded by a red border. Note that in Fig. S4, we do not display refined masks for the third, fourth and fifth rows, as Lazy-XMem missed for those instances the generation of either a user- or pseudo-corrections.

C.1 Pseudo-Corrections

Through the pixel wise uncertainty estimation, we are able to identify confusing and confident regions, helpful for the generation of pseudo-corrections, allowing us to correct the segmentation whenever a distractors is present and anticipate when the method is likely to fail as shown in Fig. S1. We can observe that our proposed pseudo-correction generation strategy successfully recovers the original object of interest in the presence of distractors (e.g., rows two, three, and four). Additionally, objects that are about to be lost are also recovered (e.g., rows one, three, and five).

Note that for small objects (refer to Fig. S2), the mask refinement incorrectly generates masks, although the pseudo-correction location's lies on the target, as seen in rows two, three, and five. In the first row, the small gorilla (target) is lost in favor to the adult gorilla, since the uncertainty is lower the method fails to issue correct pseudo-corrections or request a user-corrections. Ideally, the method should detect the transition from the small gorilla to the adult gorilla, while the pixel level uncertainty for both objects is still high, to indicate confusion. In row 6, we note that the pixel uncertainty for the foot region and the ball (target) are very similar, consequently the method is unable to find a correct location for the pseudo-correction generation as both object are as likely considered as the actual object to track by the sVOS baseline, here the method failed to actually issue a user-correction.

Fig. S1: Qualitative results on the validation set of LVOS [8] when refining the mask through pseudo-corrections (Success cases).

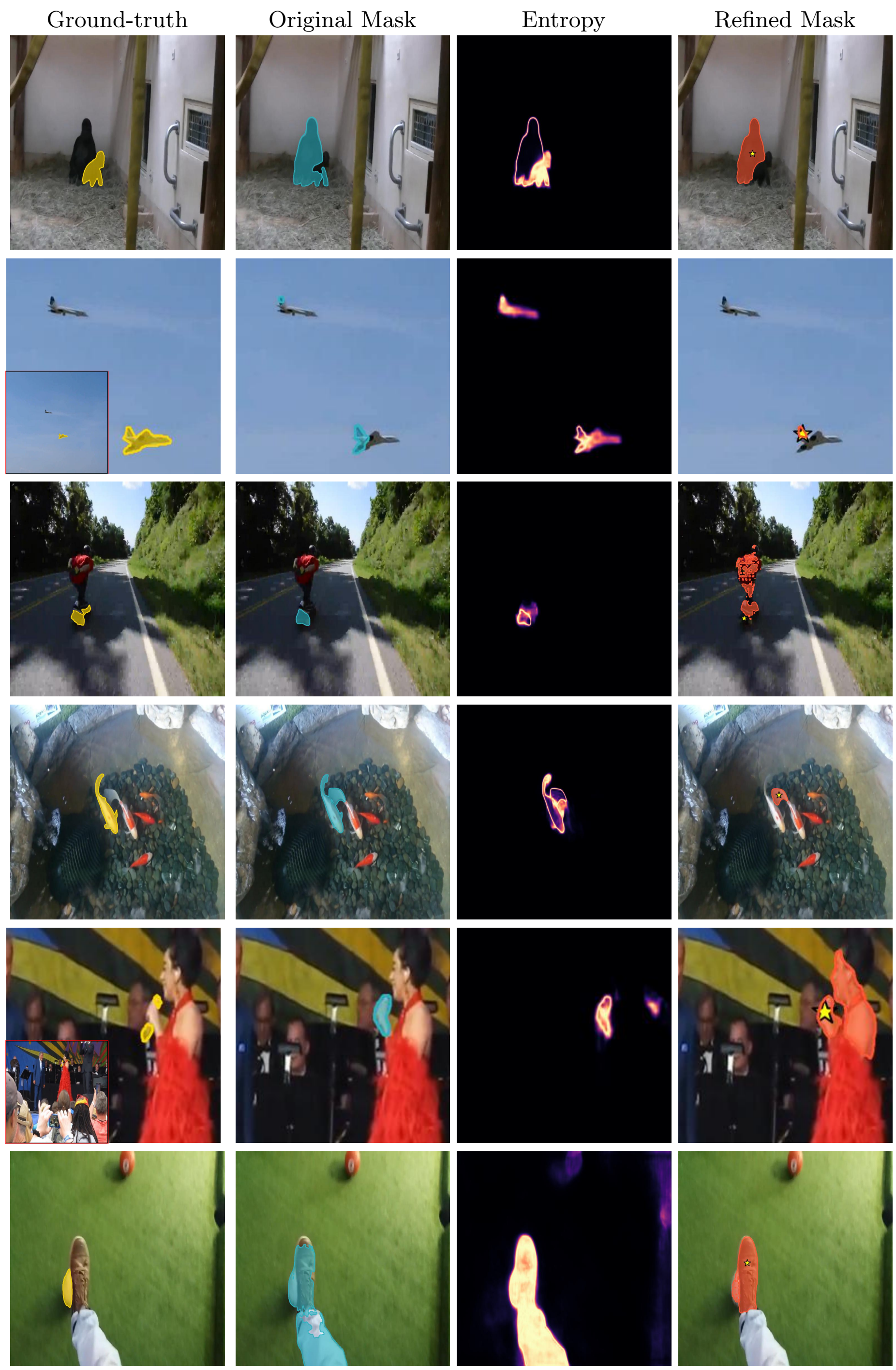


Fig. S2: Qualitative results on the validation set of LVOS [8] when refining the mask through pseudo-corrections (Failure cases).

C.2 User-Corrections

In the first and last rows of Fig. S3, we note that the method correctly issues a user interactions, as only the ear of the sheep and the back of the zebra are still segmented, preventing the loss of the target. Similarly, in the second and third rows, the method manages to issue and interaction to the user while losing the target in favor to a distractors. Note that in the third row, the method correctly issues a user-correction instead of a pseudo-correction, as otherwise the pseudo-correction would be generated on the wrong sheep.

In Fig. S4, we observe that the method sometimes unnecessarily calls for user interaction even when a good portion of the object is correctly predicted (*i.e.*, first and second row), and where a pseudo-correction would be more appropriate (first row).

Additionally, there are instances where a user (or pseudo) correction is missed, as seen in rows three, four and five. In the fourth row, the tracker confidently segments a distractor after the disappearance of the object of interest, while indicating the actual object with some uncertainty. Lastly, when the SVOS backbone loses track of the object of interest, it is unable to recover it, as shown in the fifth row.

D Kernel Size for Dilating the Mask

In Fig. S5, we present the distribution of the Spearman correlation coefficient [11] on the DAVIS dataset [10]. Our experiments use a kernel size of 2. However, as shown in Fig. S5, this choice is rather permissive, as larger kernel sizes (and none for the first case) yield similar outcomes.

E Limitations and Future Directions

Currently, Lazy-XMem generates only click-based pseudo-corrections, which are fed to the mask-refiner without including the predicted mask. This approach limits the impact of the initial mask proposed by the sVOS pipeline.

This bottleneck is inherent to SAM-based models, as they do not consider masks as prompts in practice. An alternative approach, explored by Delatolas *et al.* [6], involves iteratively prompting the mask-refiner with pseudo-prompts generated from the initial mask until a certain level of alignment is achieved between the SAM-predicted mask and the original sVOS initial mask. However, this method assumes that the initial mask (from the sVOS pipeline) is accurate enough to serve as a reliable base for further prompting the mask-refiner with uncertainty-based prompts.

An additional direction to follow in future work is to incorporate other types of prompts, like bounding-boxes or scribble-type, which might add more context to the prompt. Additionally, while we mostly rely on positive pseudo-clicks, including negative interactions could further enhance the method's capabilities.

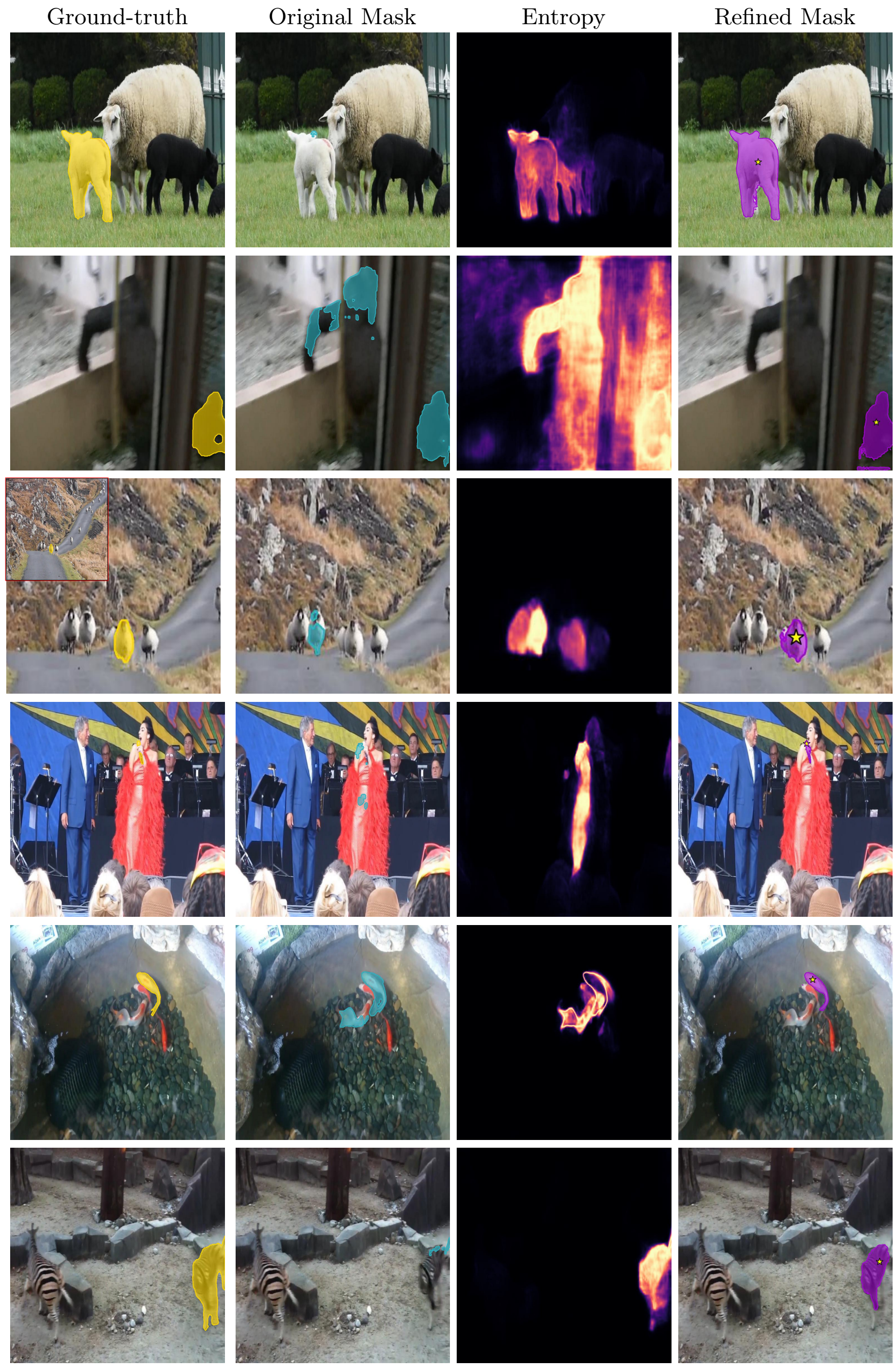


Fig. S3: Qualitative results on the validation set of LVOS [8] when refining the mask through user-corrections (Success cases).

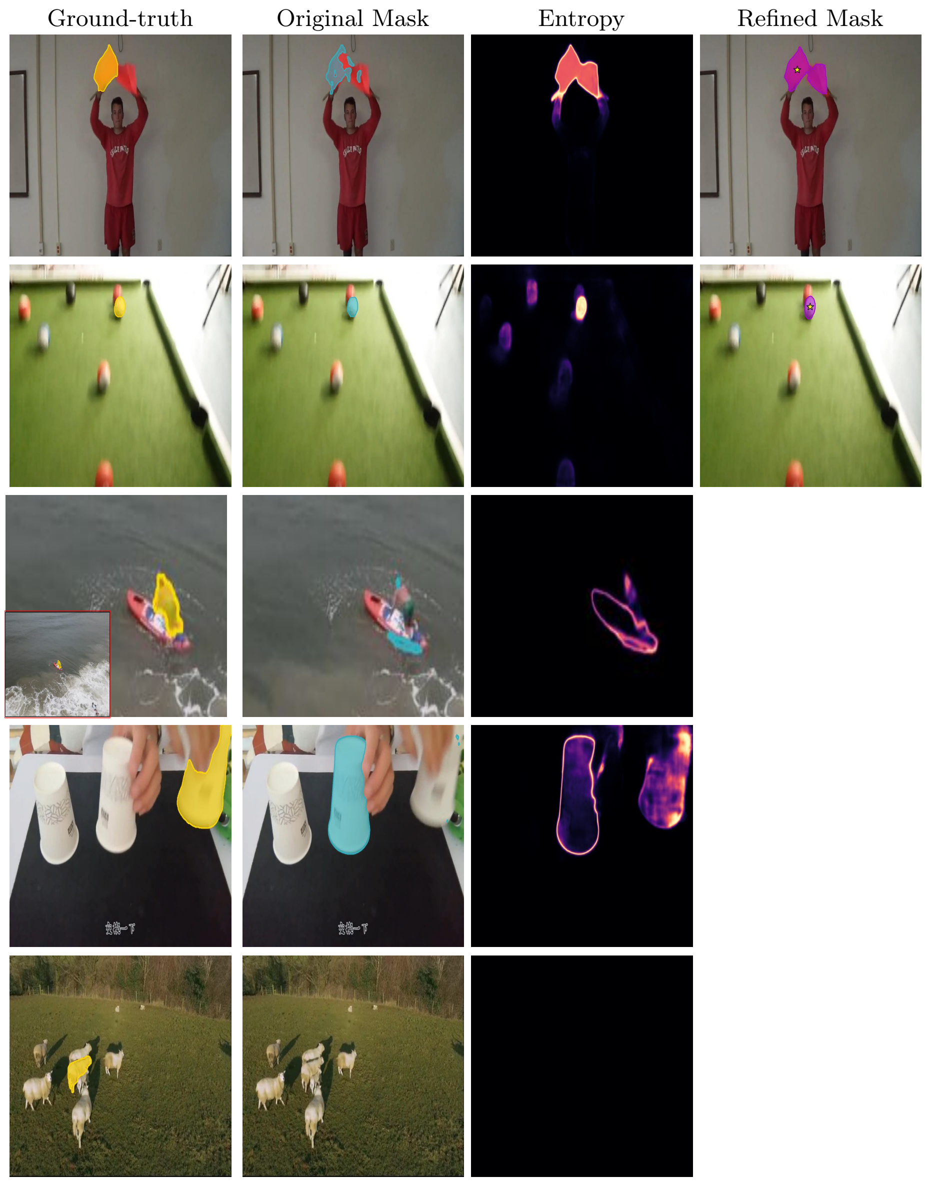
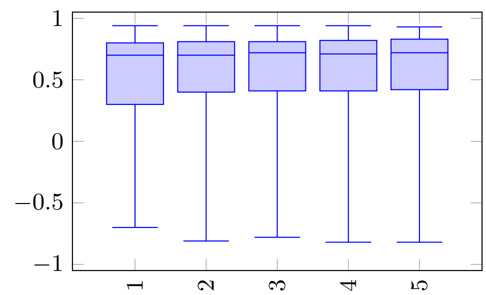


Fig. S4: Qualitative results on the validation set of LVOS [8] when refining the mask through user-corrections (Failure and miss cases). Here we considered a missed opportunity to generate a pseudo- or user-correction whenever the Intersection over Union (IoU) between the original prediction and the ground-truth annotation is below 0.1.



(a) Spearman correlation distribution for different kernel sizes when computing the masked entropy $S_{\mathcal{R}_c}$ on the DAVIS 2017 validation set [10].

Fig. S5: Varying the dilation of the masked entropy

References

1. Cheng, H.K., Oh, S.W., Price, B., Lee, J.Y., Schwing, A.: Putting the object back into video object segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 1
2. Cheng, H.K., Oh, S.W., Price, B., Schwing, A., Lee, J.Y.: Tracking anything with decoupled video segmentation. In: International Conference on Computer Vision (ICCV) (2023) 1
3. Cheng, H.K., Schwing, A.G.: XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: European Conference on Computer Vision (ECCV) (2022) 1, 2
4. Cheng, H.K., Tai, Y.W., Tang, C.K.: Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 2
5. Cheng, H.K., Tai, Y.W., Tang, C.K.: Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In: Neural Information Processing Systems (NeurIPS) (2021) 2
6. Delatolas, T., Kalogeiton, V., Papadopoulos, D.P.: Learning the what and how of annotation in video object segmentation. In: Winter Conference on Applications of Computer Vision (WACV) (2024) 6
7. Ding, H., Liu, C., He, S., Jiang, X., Torr, P.H., Bai, S.: MOSE: A new dataset for video object segmentation in complex scenes. In: International Conference on Computer Vision (ICCV) (2023) 2
8. Hong, L., Chen, W., Liu, Z., Zhang, W., Guo, P., Chen, Z., Zhang, W.: Lvos: A benchmark for long-term video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023) 1, 2, 4, 5, 7, 8
9. Liu, Y., Yu, R., Yin, F., Zhao, X., Zhao, W., Xia, W., Yang, Y.: Learning quality-aware dynamic memory for video object segmentation. In: European Conference on Computer Vision (ECCV) (2022) 1
10. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv:1704.00675 (2017) 2, 6, 9

11. Spearman, C.: The proof and measurement of association between two things. American Journal of Psychology (1904) 6
12. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2015) 2