6 Appendix

6.1 Evaluation environment

Our implementation of CropMAE and SiamMAE are based on the MAE Pytorch open-source implementation[3]. Unless specified otherwise in Table 4, we use the default parameters specified in the original paper [25]. For the evaluation on the downstream tasks, we use the parameters presented in Table 5. Our experiments were performed on 4×4 NVIDIA A100 40GB and on 4× AMD EPYC 7513 32-core. Video decoding on K400 was performed on CPU.

Table 4: Hyperparameters of CropMAE and SiamMAE. Comparison of hyperparameters used for CropMAE, both on ImageNet [12] and K400 [32], and SiamMAE on K400 [32]. The same parameters were used for both methods when possible, and the original parameters of SiamMAE were used.

Config	**CropMAE**	**SiamMAE [23]**
Optimizer	AdamW [35]	AdamW [35]
Optimizer Momentum	$\beta_1, \beta_2 = 0.9, 0.95$ [7]	$\beta_1, \beta_2 = 0.9, 0.95$ [7]
Weight Decay	0.05	0.05
Learning Rate	1.5e-4	1.5e-4
Mask Ratio	0.985	0.95
Learning Rate Schedule	Cosine Decay [34]	Cosine Decay [34]
Warmup Epochs [21]	10	10
Epochs	400	400
Repeated Sampling [28]	1 (IN), 2 (K400)	2
Augmentation V_1	Hflip (p=0.5), Crop $[a, c]$	Hflip (p=0.5), Crop $[0.5, 1]$
Augmentation V_2	Hflip (p=0.5), Crop $[b, d]$	-
Effective Batch Size	2048	2048
Frame Sampling Gap	-	$[4, 48]$
Min Aspect Ratio	3/4 (V_1 & V_2)	3/4
Max Aspect Ratio	4/3 (V_1 & V_2)	4/3
Min Area V_1 (a)	0.10 (IN), 0.50 (K400)	-
Min Area V_2 (b)	0.30	-
Max Area V_1 (c)	1.0	-
Max Area V_2 (d)	0.60	-

Table 5: Parameters used for the downstream tasks.

Config	**DAVIS-2017 [39]**	**VIP [56]**	**JHMDB [30]**
Top-k	7	10	7
Queue Length	20	20	20
Neighborhood Size	20	20	20

[3] https://github.com/facebookresearch/mae

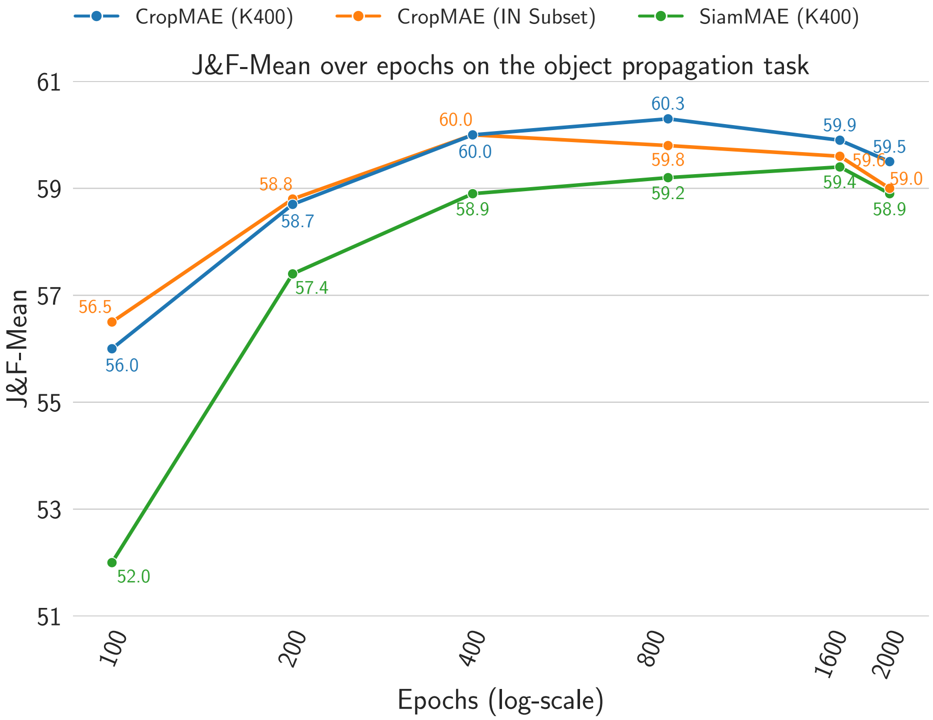

Fig. 7: Performances of CropMAE and SiamMAE on DAVIS during pre-training. For a fixed number of 2,000 epochs, CropMAE trains faster and consistently yields better results than SiamMAE [23], when trained on K400 frames or ImageNet Subset images.

6.2 Longer training

We ran experiments for 2,000 epochs with our different setups: SiamMAE trained on K400, CropMAE trained on K400, and CropMAE trained on ImageNet. The results are presented in Figure 7.

Overall, our approach demonstrates consistent superior performance compared to SiamMAE for both video (K400) and image (ImageNet) training. Our approach demonstrates significantly faster learning than SiamMAE. In particular, our method achieves a $\mathcal{J}\&\mathcal{F}_m$ value of 56.5 after only 100 epochs on our ImageNet Subset, whereas SiamMAE only achieves a value of 52.0 at this stage. At 400 epochs, our method reaches a $\mathcal{J}\&\mathcal{F}_m$ value of 60.0 for both video and image training, while SiamMAE has a value of 58.9. Even though the peak value (59.4) of SiamMAE is achieved later during the pre-training compared to our method, SiamMAE is not able to reach the performance we obtain. We attribute this trend to our pretext task, which does not require any conceptual knowledge to be completely tractable and uses object transformations much more explicitly than SiamMAE, leading to faster propagation comprehension. In contrast, SiamMAE must learn the concept of motion and understand object transformations more implicitly between two frames through more complex perturbations such as occlusions and viewpoint changes. Finally, we can see that none of the three methods is really able to scale well with very long pre-training, which is a behavior already depicted in [31].