Supplementary Material:
MAST: A Memory-Augmented Self-supervised Tracker

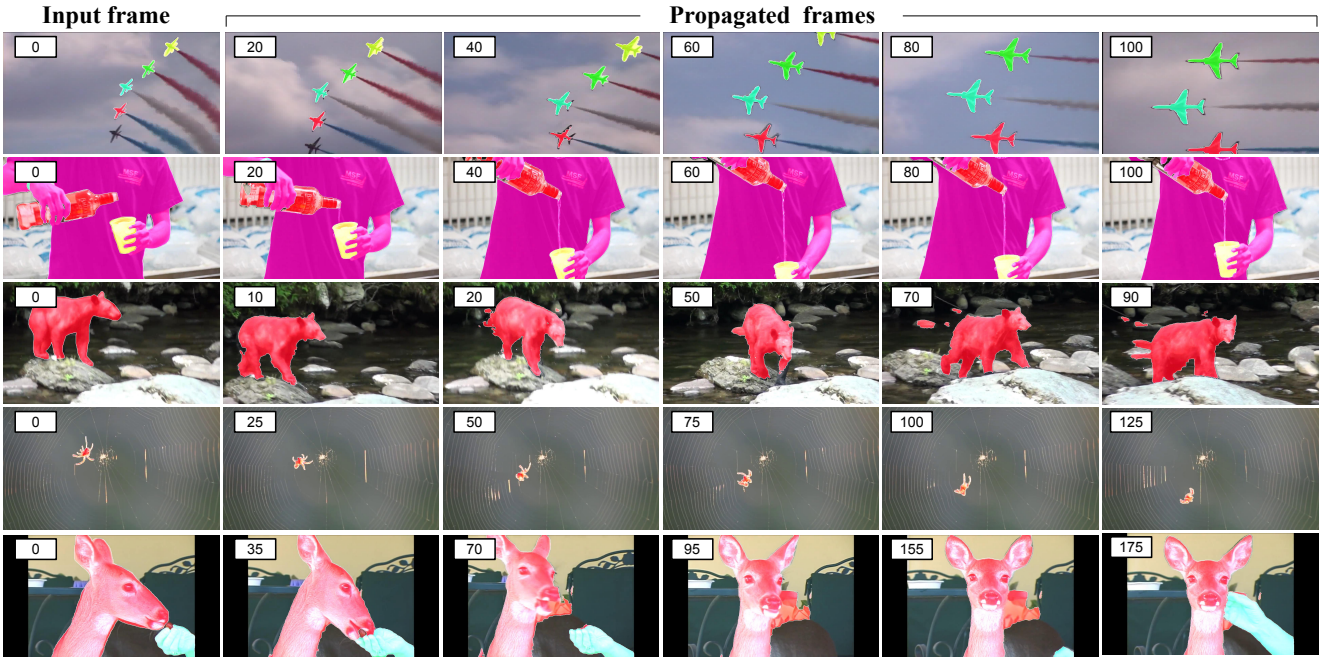

Figure 1: **More qualitative results** from our *self-supervised dense tracking model* on the YouTube-VOS dataset. The number on the top left refers to the frame number in the video. Row 1: Tracking multiple similar objects with scale change. Row 2: Occlusions and out-of-scene objects (hand, bottle, and cup). Row 3: Large camera shake. Row 4: Small object with fine details. Row 5: Inferring unseen pose of the deer; out-of-scene object (hand).

1. Network Architecture

In the same way as CorrFlow[2], we use a modified ResNet-18[1] architecture. Details of the network are illustrated in Table 1.

Stage	Output	Configuration
0	$H \times W$	Input image
conv1	$H/2 \times W/2$	7×7, 64, stride 2
conv2	$H/2 \times W/2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 2$
conv3	$H/4 \times W/4$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 2$
conv4	$H/4 \times W/4$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 2$
conv5	$H/4 \times W/4$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 2$

Table 1: Network architecture. Residual Blocks are shown in brackets (a residually connected sequence of operations). See [1] for details.

2. Accuracy Analysis from Attributes

We provide a more detailed accuracy list broken down by video attributes provided by the DAVIS benchmark[3] (listed in Table 2). The attributes illustrate the difficulties associated with each video sequence. Figure 2 contains the accuracies categorized by attribute. Several trends emerge: *first,* MAST outperforms all other self-supervised and unsupervised models by a large margin in all attributes. This shows that our model is robust to various challenges in dense tracking. *Second,* MAST obtains significant gains on occlusion-related video sequences (*e.g.* OOC, OV), suggesting that memory-augmentation is a key enabler for high-quality tracking: retrieving occluded objects from previous frames is very difficult without memory augmenta-

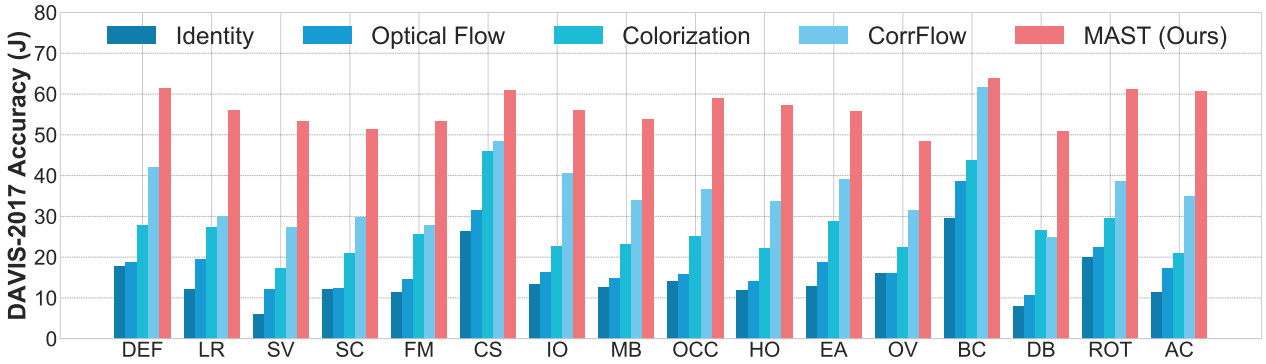

Figure 2: **Accuracy broken down by attribute:** MAST outperforms previous self-supervised methods by a significant margin on all attributes, demonstrating the robustness of our model.

tion. *Third,* in videos involving background clutter, *i.e.* background and foreground share similar colors, MAST obtains a relatively small improvement over previous methods. We conjecture this bottleneck could be caused by a shared photometric loss; thus a different loss type (*e.g.* based on texture consistency) could further improve the result.

ID	Description	ID	Description
AC	Appearance Change	IO	Interacting Objects
BC	Background Clutter	LR	Low Resolution
CS	Camera-Shake	MB	Motion Blur
DB	Dynamic Background	OCC	Occlusion
DEF	Deformation	OV	Out-of-view
EA	Edge Ambiguity	ROT	Rotation
FM	Fast-Motion	SC	Shape Complexity
HO	Heterogeneus Object	SV	Scale-Variation

Table 2: List of video attributes provided in the DAVIS benchmark. We break down the validation accuracy according to the attribute list.

3. Results on YouTube-VOS 2019 dataset

We also evaluate MAST and two other self-supervised methods on YouTube-VOS 2019 validation dataset. The numerical results are reported in Table 3. Augmenting on the 2018 version, the 2019 version contains more videos and object instances. We observe similar trend as reported in the main paper (*i.e.* significant improvement and lower generalization gap).

Method	Sup.	Overall ↑	Seen		Unseen		Gen. Gap ↓
			$\mathcal{J}$ ↑	$\mathcal{F}$ ↑	$\mathcal{J}$ ↑	$\mathcal{F}$ ↑	
Vid. Color.[4]†	✗	39.0	43.3	38.2	36.6	37.5	3.7
CorrFlow[2]	✗	47.0	51.2	46.6	44.5	45.9	3.7
SMAT (Ours)	✗	**64.9**	**64.3**	**65.3**	**61.5**	**68.4**	**0.15**

Table 3: Video segmentation results on Youtube-VOS 2019 dataset. Higher values are better. † indicates results based on our reimplementation.

4. More qualitative results

As shown in Figure 1, we provide more qualitative results exhibiting some of difficulties in the tracking task. These difficulties include tracking multiple similar objects (multi-instance tracking often fails by conflating similar objects), large camera shake (objects may have motion blur), inferring unseen object pose of objects, and so on. As shown in the figure, MAST handles these difficulties well.

5. Supplementary video

In order to better illustrate our results, we provide a supplementary video in our project page (https://github.com/zlai0/MAST). In the video, we give simple description of our method and also qualitative comparison with other self-supervised algorithms. We obtain qualtitative results of Video Colorization[4] from our reimplementation (our reimplemented achieves 34.1 in $\mathcal{J}\&\mathcal{F}$(Mean) while the original paper reports 34.0). The qualitative results for CycleTime[5] and CorrFlow[2] comes from their official codes.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016. 1

[2] Zihang Lai and Weidi Xie. Self-supervised learning for video correspondence flow. In *Proc. BMVC.*, 2019. 1, 2

[3] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proc. CVPR*, 2016. 1

[4] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proc. ECCV*, 2018. 2

[5] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time. In *Proc. CVPR*, 2019. 2