Self-supervised Learning for Video Correspondence Flow

Zihang Lai, Weidi Xie

Contributions

Two main contributions of this work are:

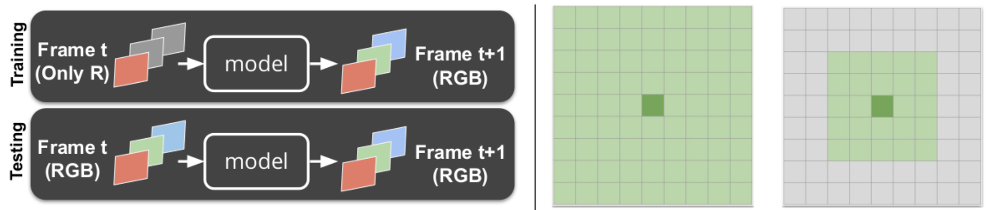
- Simple information bottleneck (frame reconstruction by pixelwise matching) that forces the model to learn robust features for correspondence matching.
- The model is trained recursively on videos over long temporal windows to alleviate the tracker drifting effects with scheduled sampling and cycle consistency.

Figure 2: An overview of the proposed self-supervised learning for correspondence flow. A recursive model is used to compute the dense correspondence matching over a long temporal window with forward-backward cycle consistency.

Method

Feature Embedding: Used ResNet as a feature encoder. Zero out 0, 1 or 2 channels in each RGB frame and apply data augmentation (brightness, contrast, saturation). Data augmentation prevents models from co-adaptation of low-level colors or illumination changes.
Restricted Attention It helps decrease in computation and memory consumption compared to full attention. Maximum disparity of M pixels is imposed in reference frame t to search for locally in square patch size $(2M+1)x(2M+1)$ centered at target pixel.

$$A^{ijkl} = \frac{\exp\left\langle f_t^{(i+k-M)(j+l-M)}, f_{t+1}^{ij}\right\rangle}{\sum_p \sum_q \exp\left\langle f_t^{(i+q)(j+p)}, f_{t+1}^{ij}\right\rangle} \tag{1}$$



(a) Colour Dropout **(b) Full Attention** **(c) Restricted Attention**

Figure 3: Restricted attention and colour dropout. See text for details.

where *(i,j,k,l)* is the entry of tensor denotes similarity between *(i,j)* of target frame, and pixel *(i+k-M, j+l-M)* of the reference frame

$$\hat{I}_{t+1} = \psi\left(A_{(t,t+1)}, I_t\right) = \sum_p \sum_q A^{ij(p+M)(q+M)} I_t \tag{2}$$

Long-Term Correspondence Flow Sampling training frames is difficult. Too close will result in no change and farther apart will result in way too much complex change.

- Scheduled Sampling: Replace Ground-truth tokens by model's prediction. Shared embedding network is used to get feature embeddings ($f_i = \Phi\left(g\left(I_i\right); \theta\right)$ where $i = 1, \ldots, n$). The reconstruction is a recursive process where nth frame $(\tilde{I_n})$ may have access to previous frame's ground truth (I_{n-1}) or model prediction $(\tilde{I_{n-1}})$.

$$\hat{I}_n = \begin{cases} \psi\left(A_{(n-1,n)}, I_{n-1}\right) \\ \psi\left(A_{(n-1,n)}, \hat{I}_{n-1}\right) \end{cases} \tag{3}$$

- Cycle-Consistency: It is used as a regularizer. Apply n frames forward and backward in future.

The final objective function is defined as:

$$L = \alpha_1 \cdot \sum_{i=1}^{n} \mathcal{L}_1\left(I_i, \hat{I}_i\right) + \alpha_2 \cdot \sum_{j=n}^{1} \mathcal{L}_2\left(I_j, \hat{I}_j\right) \tag{4}$$

where $\mathcal{L}_1, \mathcal{L}_2$ are pixelwise cross entropy loss between groundtruth and reconstructed frames in the forward and backward path.

Results

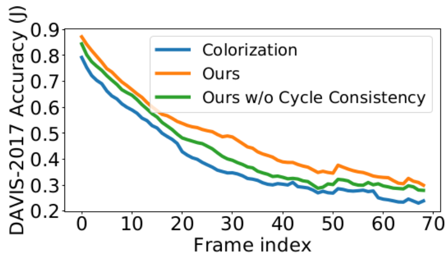


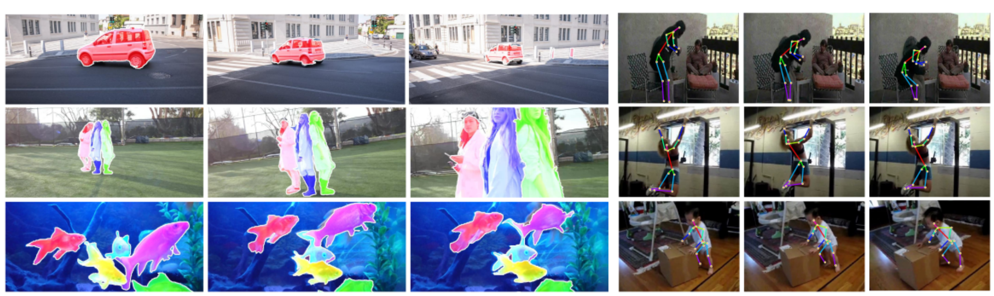
Figure 4: Model comparison on the problem of tracker drifting. The proposed model with cycle consistency has shown to be most robust as masks propagate.

Method	$\mathcal{J}$(Mean)	$\mathcal{F}$(Mean)
Ours (Full Model)	47.7	51.3
Ours w/o Colour Dropout	40.5	39.5
Ours w/o Restricted Attention	40.8	39.7
Ours w/o Scheduled Sampling	40.2	39.2
Ours w/o Cycle Consistency	41.0	40.4

Table 1: Ablation Studies on DAVIS-2017. $\mathcal{J}$: region overlapping, $\mathcal{F}$: contour accuracy respectively.

Method	Supervised	Dataset	$\mathcal{J}$&$\mathcal{F}$(Mean)	$\mathcal{J}$(Mean)	$\mathcal{J}$(Recall)	$\mathcal{F}$(Mean)	$\mathcal{F}$(Recall)
Identity	✗	-	22.9	22.1	15.9	23.6	11.7
Optical Flow (FlowNet2) [15]	✗	-	26.0	26.7	-	25.2	-
SIFT Flow [28]	✗	-	34.0	33.0	-	35.0	-
Transitive Inv. [42]	✗	-	29.4	32.0	-	26.8	-
DeepCluster [50]	✗	YFCC100M	35.4	37.5	-	33.2	-
Video Colorization [40]	✗	Kinetics	34.0	34.6	34.1	32.7	26.8
CycleTime (ResNet-50) [43]	✗	VLOG	40.7	41.9	40.9	39.4	33.6
Ours (Full Model ResNet-18)	✗	Kinetics [23]	**49.5**	**47.7**	**53.2**	**51.3**	**56.5**
Ours (Full Model ResNet-18)	✗	OxUvA [39]	**50.3**	**48.4**	**53.2**	**52.2**	**56.0**
ImageNet (ResNet-50) [13]	✓	ImageNet	49.7	50.3	-	49.0	-
SiamMask [41]	✓	YouTube-VOS	53.1	51.1	60.5	55.0	64.3
OSVOS[6]	✓	DAVIS	60.3	56.6	63.8	63.9	73.8

Table 2: Video segmentation results on DAVIS-2017 dataset. Higher values are better.



(a) DAVIS 2017 Video Segmentation (b) Keypoint Tracking

Method	Supervised	Dataset	PCK$_{instance}$		PCK$_{max}$	
			@.1	@.2	@.1	@.2
SIFT Flow[28]	✗	-	49.0	68.6	-	-
Video Colorization [40]	✗	Kinetics	45.2	69.6	-	-
CycleTime (ResNet-50) [43]	✗	VLOG	57.7	78.5	-	-
Ours (Full Model ResNet-18)	✗	Kinetics	**58.5**	**78.8**	**71.9**	**88.3**
ImageNet (ResNet-50) [13]	✓	ImageNet	58.4	78.4	-	-
Fully Supervised [38]	✓	JHMDB	-	-	68.7	81.6

Table 4: Keypoint tracking on JHMDB dataset (validation split 1). Higher values are better.