Learning Correspondence from the Cycle-consistency of Time

Xiaolong Wang, Allan Jabri, Alexei A. Efros

Contributions

The main aim of this work is to use cycle consistency in time as free supervisory signal for learning visual representations. The key idea is that we can use unlimited supervision by template matching along a cycle of time. Trivial solutions are avoided by forcing the tracker to localize patch in each successive frame.

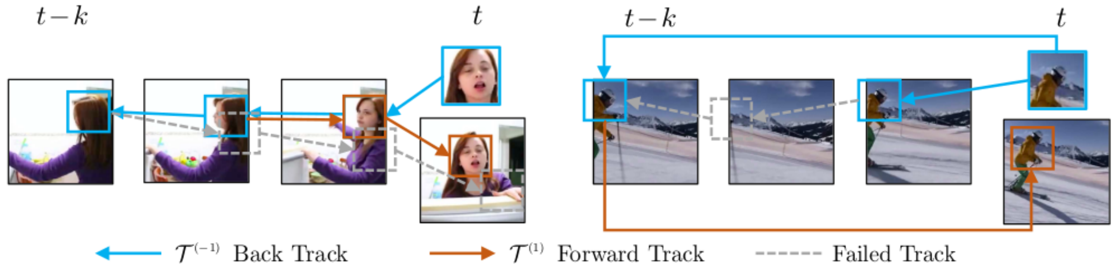

$t-k$
t
$t-k$
t
$\mathcal{T}^{(-1)}$ Back Track
$\mathcal{T}^{(1)}$ Forward Track
Failed Track

Figure 3: **Multiple Cycles and Skip Cycles.** Cycle-consistency may not be achievable due to sudden changes in object pose or occlusions. Our solution is to optimize multiple cycles of different lengths simultaneously. This allows learning from shorter cycles when the full cycle is too difficult (left). This also allows cycles that skip frames, which can deal with momentary occlusions (right).

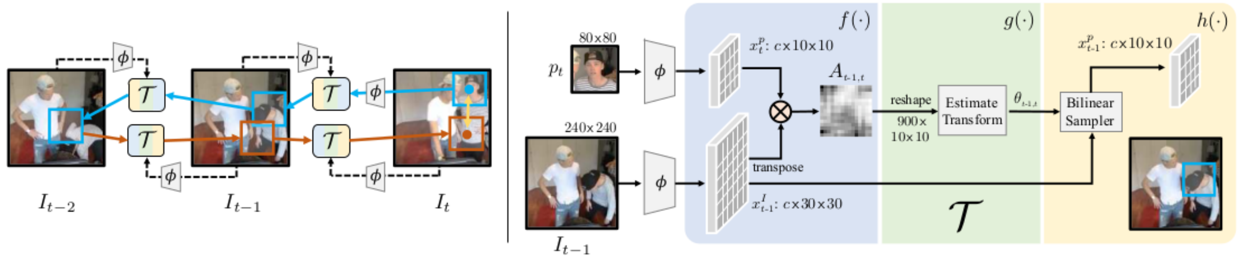
Method

Stage 1: Tracking operator (T) takes input the features of current patch and target image and return the image feature region with maximum similarity. It can iteratively be applied in both direction of time. Cycle Consistency Loss is the euclidean distance between the original and predicted patch.

Stage 2: Input patch and sequence of video frames are mapped to a feature space by an encoder. Tracker (T) helps to localize patch feature x_s^p in image feature x_s^I by applying the tracker backwards as follows:

$$\mathcal{T}^{(-i)}\left(x_{t-1}^I, x^p\right) = \mathcal{T}\left(x_{t-i}^I, \mathcal{T}\left(x_{t-i+1}^I, \ldots \mathcal{T}\left(x_{t-1}^I, x^p\right)\right)\right) \tag{1}$$

The learning objectives include three losses namely:



(a) Training ϕ by End-to-end Cycle-consistent Tracking

(b) Differentiable Tracking Operation $\mathcal{T}$

- Tracking: Tracker attempts to follow features backward and then forward.

$$\mathcal{L}_{\text{long}}^i = l_\theta \left(x_t^p, \mathcal{T}^{(i)} \left(x_{t-i+1}^I, \mathcal{T}^{(-i)} \left(x_{t-1}^I, x_t^p \right) \right) \right) \tag{2}$$

- Skip Cycle: Alongwith consecutive frames, they skip through frames.

$$\mathcal{L}_{skip}^i = l_\theta \left(x_t^p, \mathcal{T} \left(x_t^I, \mathcal{T} \left(x_{t-i}^I, x_t^p \right) \right) \right) \tag{3}$$

- Feature Similarity: negative Frobenius products between query patch and localized patch.

$$\mathcal{L}_{\text{sim}}^i = - \left\langle x_t^p, \mathcal{T} \left(x_{t-i}^I, x_t^p \right) \right\rangle \tag{4}$$

Total Loss(L) is defined as :

$$\mathcal{L} = \sum_{i=1}^{k} \mathcal{L}_{\text{sim}}^i + \lambda \mathcal{L}_{\text{skip}}^i + \lambda \mathcal{L}_{\text{long}}^i \tag{5}$$

Stage3: Spatial features are encoded by ResNet-50 network. Affinity function provides a measure of similarity between coordinates of spatial features of image and patch is calculated using below equation

$$A(j,i) = \frac{\exp \left(x^I(j)^\top x^p(i) \right)}{\sum_j \exp \left(x^I(j)^\top x^p(i) \right)} \tag{6}$$

Stage4: Localizer takes affinity matrix A and estimates localization parameters θ correspond to the patch in image feature which best matches to patch feature. Finally, alignment objective is applied on cycle-consistency losses L_{long}^i and L_{skip}^i measuring the error in alignment between two patch regions:

$$l_\theta \left(x_*^p, \hat{x}_t^p \right) = \frac{1}{n} \sum_{i=1}^{n} \left\| M \left(\theta_{x_*^p} \right)_i - M \left(\theta_{\hat{x}_t^p} \right)_i \right\|_2^2 \tag{7}$$

Results

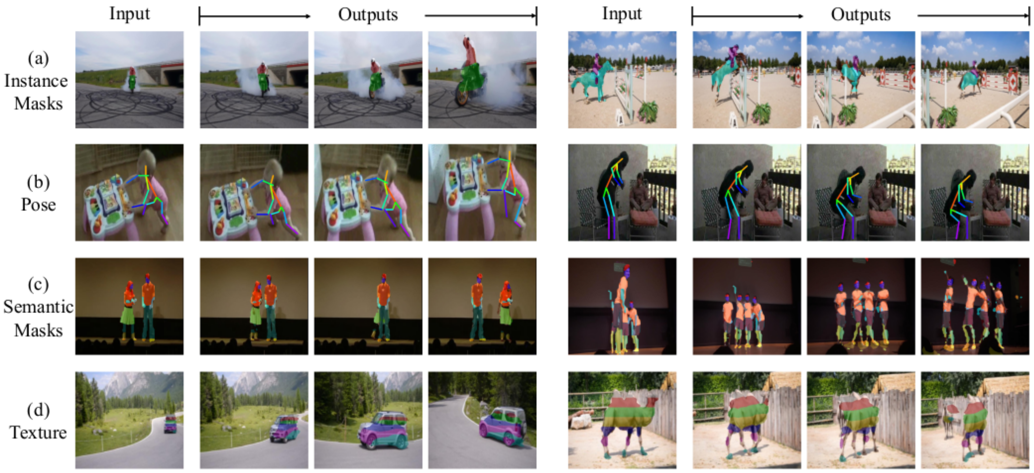


Figure 5: Visualizations of our propagation results. Given the labels as input in the first frame, our feature can propagate them to the rest of frames, without further fine-tuning. The labels include (a) instance masks in DAVIS-2017 [48], (b) pose keypoints in JHMDB [26], (c) semantic masks in VIP [85] and even (d) texture map.

model	Supervised	$\mathcal{J}$(Mean)	$\mathcal{F}$(Mean)
Identity		22.1	23.6
Random Weights (ResNet-50)		12.4	12.5
Optical Flow (FlowNet2) [22]		26.7	25.2
SIFT Flow [39]		33.0	35.0
Transitive Inv. [74]		32.0	26.8
DeepCluster [8]		37.5	33.2
Video Colorization [69]		34.6	32.7
Ours (ResNet-18)		40.1	38.3
Ours (ResNet-50)		**41.9**	**39.4**
ImageNet (ResNet-50) [18]	✓	50.3	49.0
Fully Supervised [81, 7]	✓	55.1	62.1

Table 1: Evaluation on instance mask propagation on DAVIS-2017 [48]. We follow the standard metric on region similarity $\mathcal{J}$ and contour-based accuracy $\mathcal{F}$.

model	Supervised	PCK@.1	PCK@.2
Identity		43.1	64.5
Optical Flow (FlowNet2) [22]		45.2	62.9
SIFT Flow [39]		49.0	68.6
Transitive Inv. [74]		43.9	67.0
DeepCluster [8]		43.2	66.9
Video Colorization [69]		45.2	69.6
Ours (ResNet-18)		57.3	78.1
Ours (ResNet-50)		**57.7**	**78.5**
ImageNet (ResNet-50) [18]	✓	58.4	78.4
Fully Supervised [59]	✓	68.7	92.1

Table 2: Evaluation on pose propagation on JHMDB [26]. We report the PCK in different thresholds.