Learning Correspondence from the Cycle-consistency of Time

Xiaolong Wang[*]
Carnegie Mellon University
xiaolonw@cs.cmu.edu

Allan Jabri[*]
UC Berkeley
ajabri@eecs.berkeley.edu

Alexei A. Efros
UC Berkeley
efros@eecs.berkeley.edu

Appendix A. Ablations

A.1. Removing Skip-Cycles

Removing the skip-cycle loss – i.e. keeping only the long tracking cycle loss and dense similarity loss – results in worse performance when applying the representations to the DAVIS-2017 dataset. This suggests the skip-cycle loss is useful in cases of occlusion or drift, and provides supplementary training data (c.f. Table 1).

Experiment	$\mathcal{J}$(Mean)	$\mathcal{F}$(Mean)
Ours	41.9	39.4
Ours without Skip Cycles	39.5	37.9

Table 1: Removing Skip Cycles, test on DAVIS.

A.2. Effect of k in k-NN Label Propagation

We vary the number of nearest neighbors used in voting for label propagation, finding that aggregating fewer nearest neighbors improves performance (c.f. Table 2).

Experiment	$\mathcal{J}$(Mean)	$\mathcal{F}$(Mean)
Ours (5-NN)	41.9	39.4
Ours (20-NN)	40.8	38.5
Ours (10-NN)	41.5	39.1
Ours (1-NN)	41.0	38.9

Table 2: Effect of k in k-NN Label Propagation, test on DAVIS.

A.3. Training with the Kinetics Dataset

Besides the VLOG dataset, we have also trained our model on the Kinetics Dataset, which contains around 230K training videos (with 10s per video). Compared to the VLOG dataset, the Kinetics dataset contains more videos under less environment constraints: There are videos with both indoor and outdoor scenes; some videos also have large camera motion. After applying the learned representation for label propagation on DAVIS, we observe similar performance by training with VLOG and Kinetics datasets (c.f. Table 3).

Experiment	$\mathcal{J}$(Mean)	$\mathcal{F}$(Mean)
Ours (VLOG)	41.9	39.4
Ours (Kinetics)	42.5	39.2

Table 3: Train with VLOG / Kinetics, test on DAVIS.

A.4. Fine-tuning on the Test Domain

We emphasize that our method learns features that generalize even *without* fine-tuning. Here we study the effect of fine-tuning on the DAVIS training set before testing. We find this does not improve test set performance significantly (c.f. Table 4). There is a risk of overfitting since datasets like DAVIS are so small; this is part of the reason why unsupervised methods are desirable.

Experiment	$\mathcal{J}$(Mean)	$\mathcal{F}$(Mean)
Ours (ResNet-50)	41.9	39.4
Fine-tune	42.0	39.1

Table 4: Finetuning on DAVIS train before test.

Appendix B. Capacity of $\mathcal{T}$

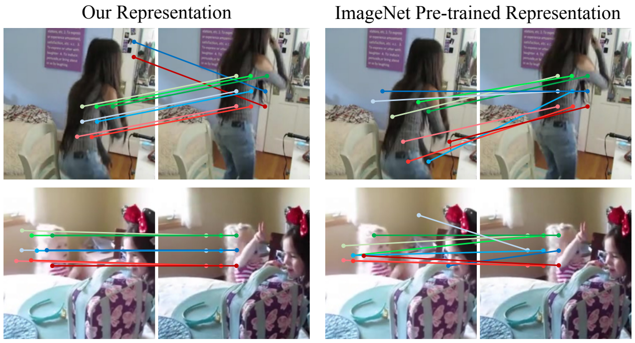
As mentioned in Section 3.2.2, the tracking operation T is deliberately constrained in capacity in order to maximize the representational responsibility of ϕ. In our implementation, the only parameters learned by $\mathcal{T}$ are those of the localizer g, which processes the affinity tensor A to estimate the localization parameters. The affinity A ($\mathbb{R}^{900\times100}$) is first reshaped to a tensor with dimension $\mathbb{R}^{900\times10\times10}$ as the input for g. The localizer g is a small ConvNet with two convolutional layers (3×3 kernels with 512 channels) and one fully connected layer. The output of the ConvNet is a 3-dimension vector corresponding to 2D translation and rotation.

Appendix C. Correspondence Visualization

In Fig. 1 we visualize the correspondences (top-1 nearest neighbor) between regions with large movement in consecutive frames, comparing our features to ImageNet pretrained features. Our method produces more detailed correspondence. However, for certain object-level tasks (e.g. DAVIS), high-level semantics (captured by ImageNet) are

[*] Equal contribution.

more useful than good correspondences, which explains the difference in performance.


Our Representation

ImageNet Pre-trained Representation

Figure 1: Visualizations of correspondence.