香港中文大學計算機科學與工程學系
Department of Computer Science and Engineering
The Chinese University of Hong Kong

Contact

Modelling Neighbor Relation in Joint Space-Time Graph for Video Correspondence Learning

Zixu Zhao, Yueming Jin, Pheng-Ann Heng
The Chinese University of Hong Kong

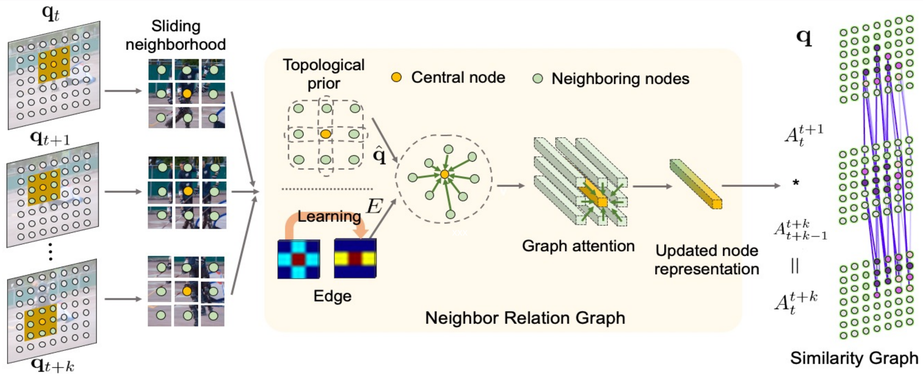
Project

2021 ICCV OCTOBER 11-17 VIRTUAL

Summary

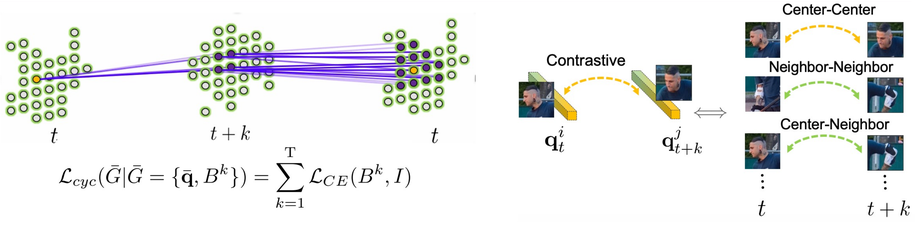
- Learning representations for space-time correspondence is a fundamental task to describe the dynamics of natural scenes from large-scale unlabeled videos.
- Three contributions:
  - A joint graph that models neighbor relations in space and similarity relations in time.
  - Formulate contrastive learning as an attentive walk on the graph with node dropout and cycle-consistency constraints.
  - State-of-the-art results on three visual tasks, *i.e.*, object, part propagation and pose tracking.

Modelling Neighbor Relation in Space-Time Graph



- Neighbor Relation Graph: it connects the central node with local neighbors, with edges initialized with topological prior and learned to determine the aggregation strength from neighbors.
- Similarity Graph: it links inter-frame nodes with pair-wise similarity affinities, to form multi-step association on a long-range sequence.

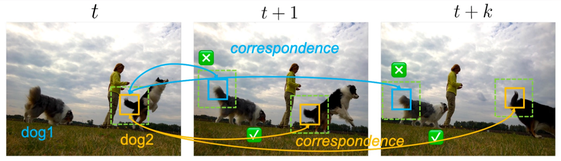
Correspondence as an Attentive Walk



$$\mathcal{L}_{cyc}(\bar{G}|\bar{G}=\{\bar{\mathbf{q}},B^k\})=\sum_{k=1}^{\mathrm{T}}\mathcal{L}_{CE}(B^k,I)$$

- A palindrome graph for training with node dropout to filter out “common-fate” nodes.
- An optimal correspondence is the path that can walk back to its initial position.
- A chain of contrastive learning problem with *extra* positive pairs, such as center-neighbor pairs and neighbor-neighbor pairs, which encourage the model to learn general neighbor relations.

Main Idea



- How to find the correspondence of a small object such as the dog tail in a video?
- We identify that both longer views (temporal dynamics) and broader views (neighbor relations) are crucial to distinguish similar instances.
- We capture these two cues in a joint graph to learn correspondence, such that the model can see longer and broader when performing query-target matching.

Quantitative Results

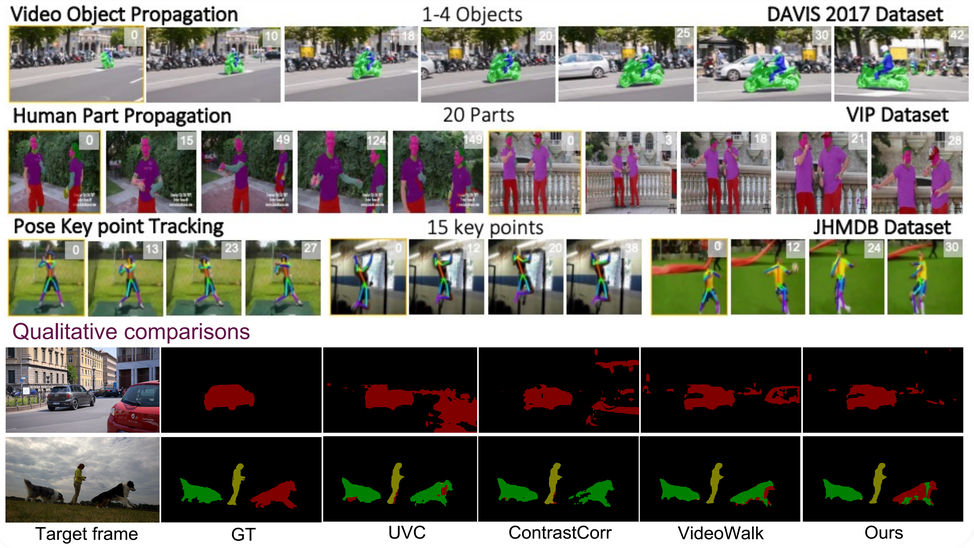
Video Object Segmentation on DAVIS 2017

Method	Supervised	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{J}_r$	$\mathcal{F}_m$	$\mathcal{F}_r$
MoCo		60.8	58.6	68.7	63.1	72.7
VINCE		60.4	57.9	66.2	62.8	71.5
CorrFlow		50.3	48.4	53.2	52.2	56.0
MAST		63.7	61.2	73.2	66.3	78.3
MAST		65.5	63.3	73.2	67.6	77.7
TimeCycle		48.7	46.4	50.0	50.0	48.0
UVC		60.9	59.3	68.8	62.7	70.9
ContrastCorr		63.0	60.5	70.6	65.5	73.0
VideoWalk		67.6	64.8	76.1	70.2	82.1
Ours		**68.7**	**65.8**	**77.7**	**71.6**	**84.3**
ImageNet	✓	62.9	60.6	69.9	65.2	73.8
SiamMask	✓	56.4	54.3	62.8	58.5	67.5
OSVOS	✓	60.3	56.6	63.8	63.9	73.8
OnAVOS	✓	65.4	61.6	67.4	69.1	75.4
OSVOS-S	✓	68.0	64.7	74.2	71.3	80.7

Part Segmentation on VIP & Pose Tracking on JHMDB

Method	Supervised	Pose		Part
		PCK@0.1	PCK@0.2	mIoU
TimeCycle		57.3	78.1	28.9
UVC		58.6	79.6	34.1
ContrastCorr		61.1	80.8	37.4
VideoWalk		59.3	84.9	38.6
Ours		**61.4**	**85.3**	**40.2**
ImageNet	✓	53.8	74.6	31.9
ATEN	✓	-	-	37.9
Thin-Slicing Net	✓	68.7	92.1	-

Qualitative Results

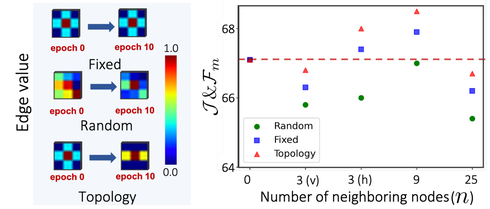


Ablation Studies

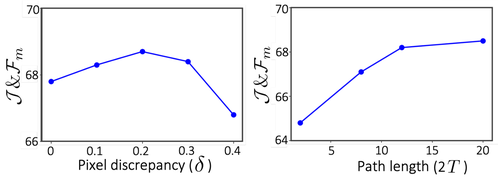
Component analysis

G_s	G_r	Node Dropout	$\mathcal{J}\&\mathcal{F}_m$
✓			65.6
✓	✓		67.8 (+2.2%)
✓	✓	✓	68.7 (+3.1%)

- 9 neighboring nodes peak the performance. Larger neighborhood size induces noise.
- Encoding topological prior in edges is essential for modelling neighbor relation of nodes.
- Moderate node dropout (0.1-0.3) boost the performance on DAVIS benchmark.
- Longer path length improves results as model can see longer views of instances for contrastive learning.



Neighborhood size & Edge value



Node dropout | Path length

Acknowledgement

This work was supported by Hong Kong Research Grants Council with Project No. CUHK 14201620.