Rethinking Self-Supervised Correspondence Learning: A Video Frame-level Similarity Perspective

Jiarui Xu, Xiaolong Wang

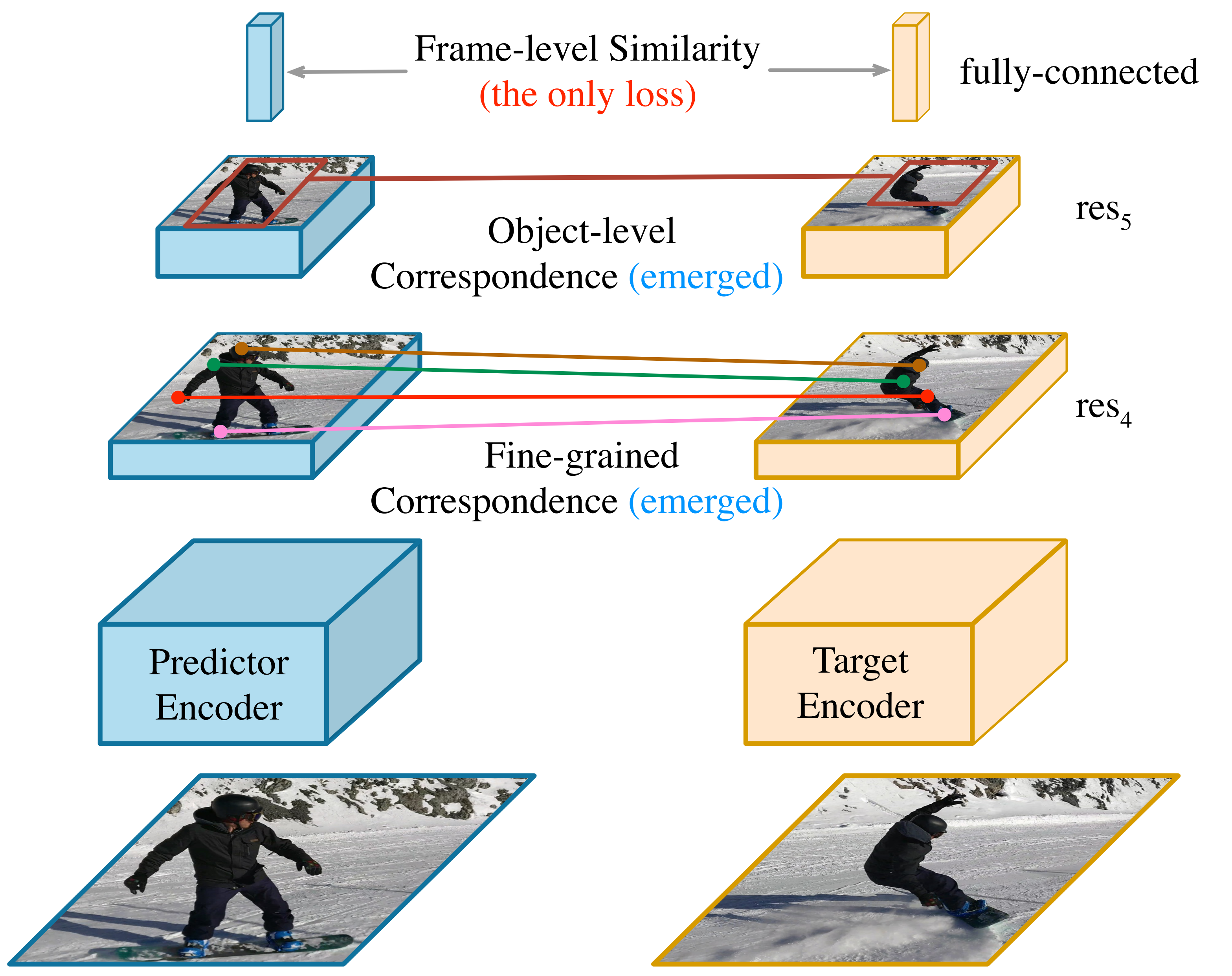
UC San Diego

Goal: Learn a representation for space-time correspondence by learning frame-level similarity. No tracking-based pretext task is required.

Overview

Compare the fully-connected layer embeddings of frames from the same video for learning.

By minimizing the frame-level feature, the fine-grained and object-level correspondence emerges in res_4 and res_5

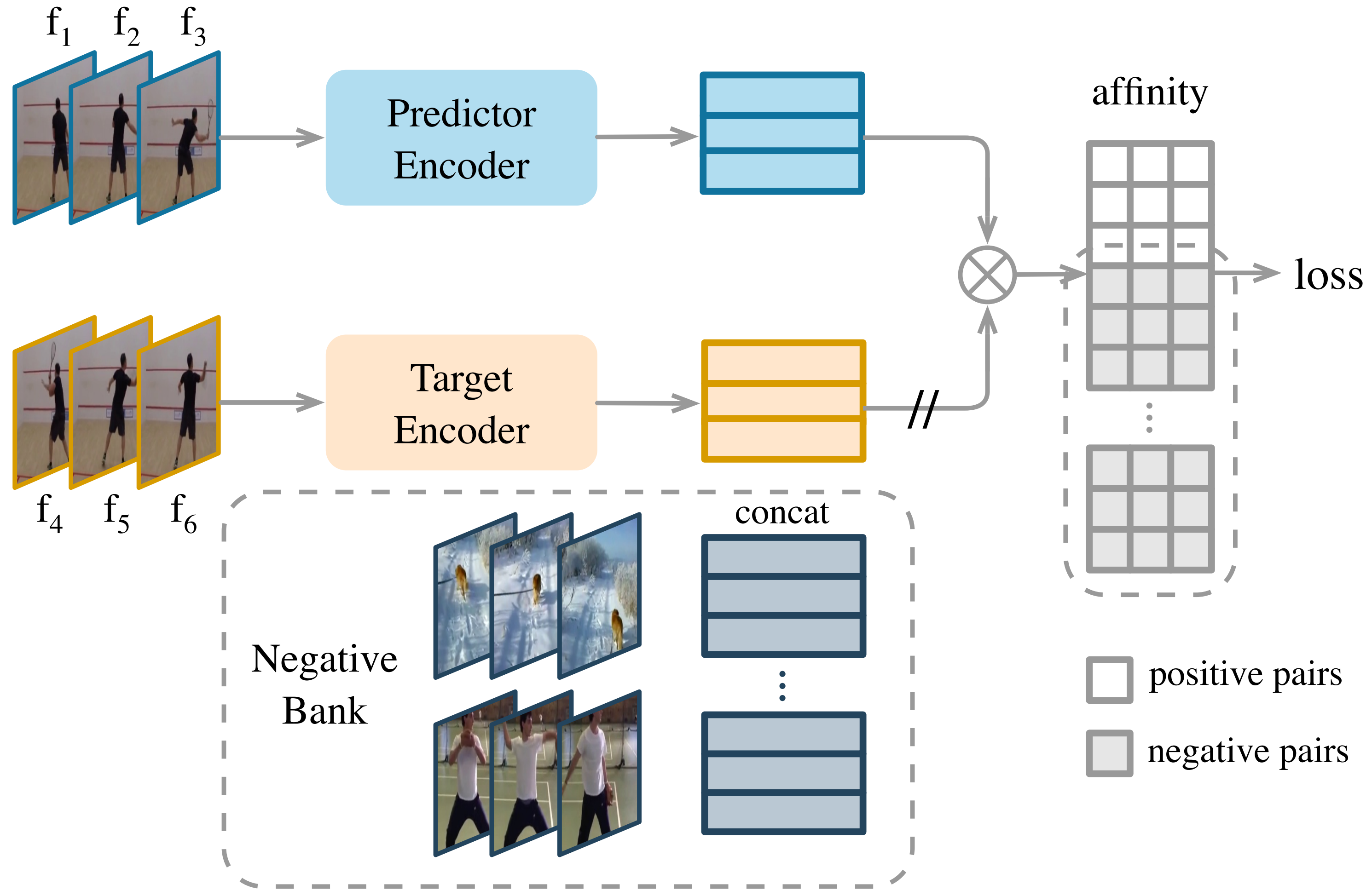


Contributions:

- Large frame gaps and multiple frame pairs helps
- Color augmentation is harmful for fine-grained correspondence, but beneficial for object-level one
- Deep networks significantly improves

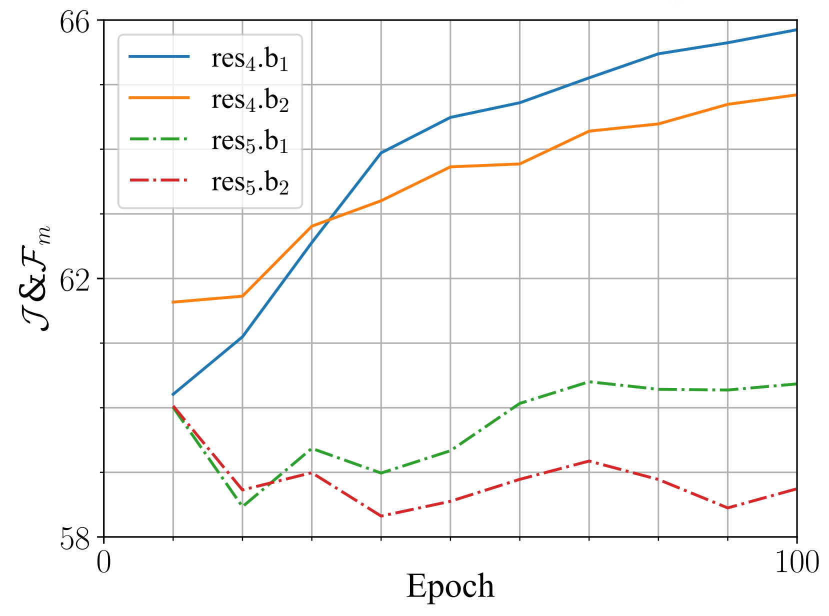
Video Frame-level Similarity Pipeline

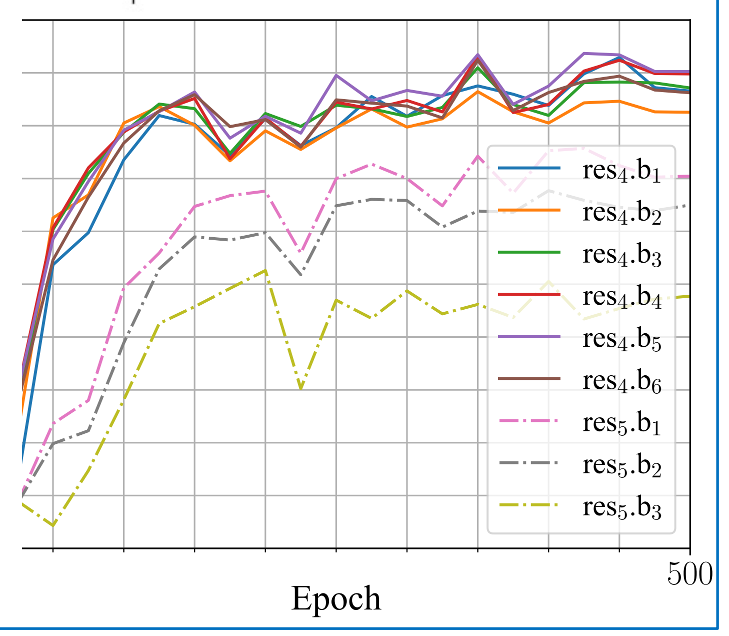
Minimize the affinity of negative, and maximize the affinity of positive



Insights

different frame	color aug	spatial aug	DAVIS $\mathcal{J}\&\mathcal{F}_m$	DAVIS $\mathcal{J}_m$	DAVIS $\mathcal{F}_m$	OTB Precision	OTB Success
			38.6	37.3	39.9	5.4	4.7
	✓		50.9	49.3	52.6	0.4	0.3
		✓	62.2	60.8	63.6	30.6	26.1
	✓	✓	61.2	59.3	63.1	53.0	39.3
✓			63.4	61.1	65.7	37.4	28.8
✓	✓		58.4	56.4	60.4	46.2	34.6
✓		✓	65.0	62.6	67.4	48.1	37.9
✓	✓	✓	61.9	59.5	64.3	57.3	43.0





Fine-grained correspondence on DAVIS Object-level correspondence on OTB

Method	Backbone	J&F	J	F	Prec.	Succ.
Supervise	ResNet-18	62.9	60.6	65.2	61.4	43.0
SimSiam	ResNet-18	62.0	60.0	64.0	58.8	42.9
MoCo	ResNet-18	60.8	58.6	63.1	62.0	47.0
VINCE	ResNet-18	60.4	57.9	62.8	62.9	46.5
CRW	ResNet-18	67.6	64.8	70.2	52.6	40.1
VFS	**ResNet-18**	**66.7**	**64.0**	**69.4**	**68.9**	**52.2**
Supervise	ResNet-50	66.0	63.7	68.4	65.8	45.5
SimSiam	ResNet-50	66.3	64.5	68.2	61.0	43.2
MoCo	ResNet-50	65.4	63.2	67.6	63.7	46.5
VFS	**ResNet-50**	**68.9**	**66.5**	**71.3**	**68.9**	**52.2**