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

Jiarui Xu

Xiaolong Wang

UC San Diego

2021 ICCV OCTOBER 11-17 VIRTUAL

Spatial Temporal Correspondence

Spatial Temporal Correspondence

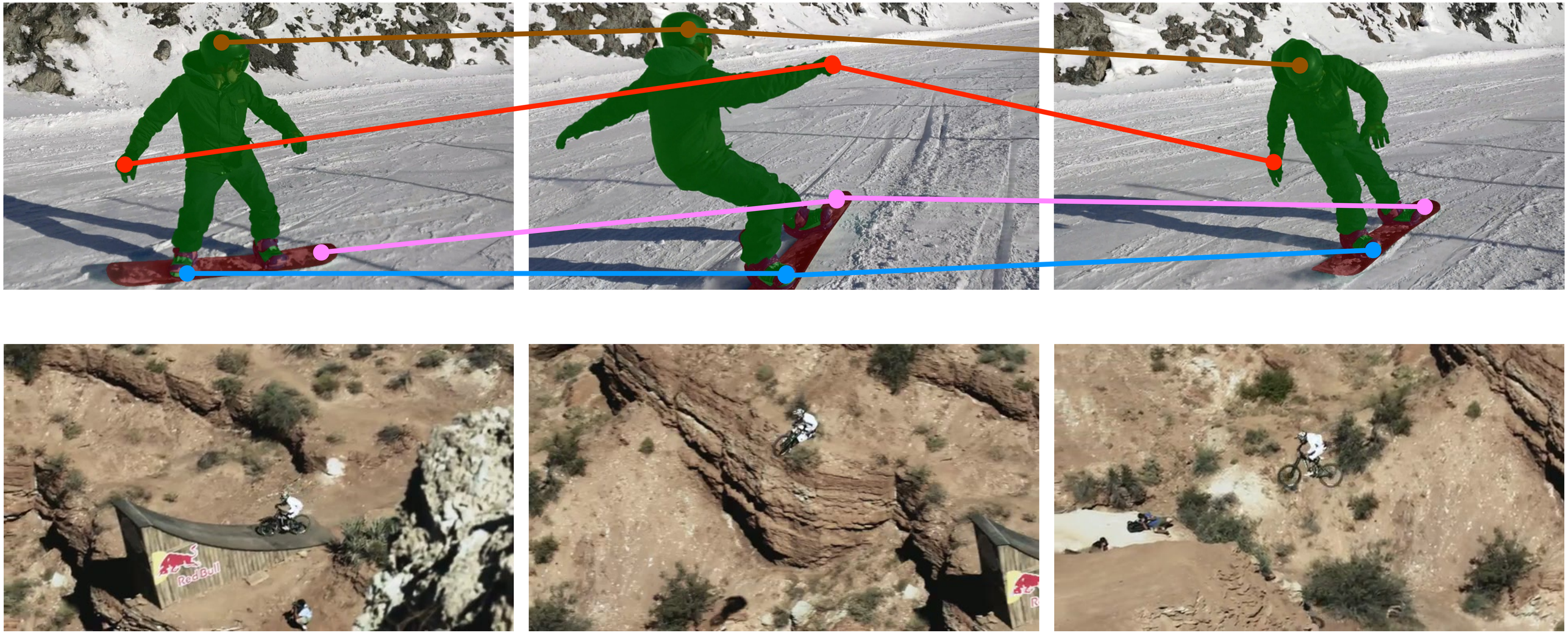
Spatial Temporal Correspondence





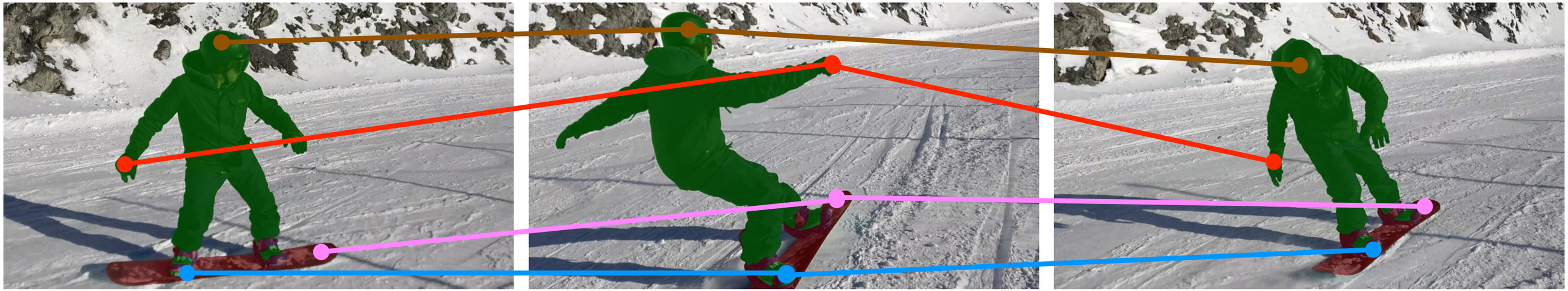
Spatial Temporal Correspondence

Fine-grained Correspondence



Spatial Temporal Correspondence

Fine-grained Correspondence



Object-level Correspondence

Supervision of Learning Correspondence

Human Supervision from Annotated Datasets

Supervision of Learning Correspondence

Human Supervision from Annotated Datasets

- Manually Labeled datasets:
  - DAVIS VOS
  - YouTube-VOS
  - OTB
  - GOT-10K
  - …

Supervision of Learning Correspondence

Human Supervision from Annotated Datasets

- Manually Labeled datasets:
  - DAVIS VOS
  - YouTube-VOS
  - OTB
  - GOT-10K
  - …

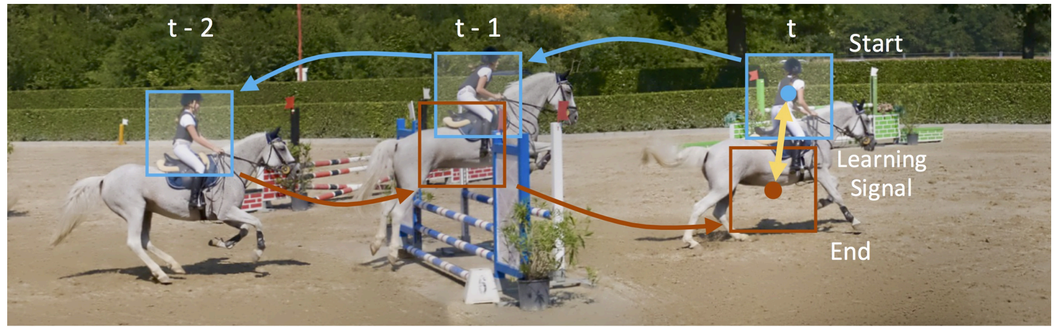
Supervision of Learning Correspondence

Self-Supervision from Temporal Signals

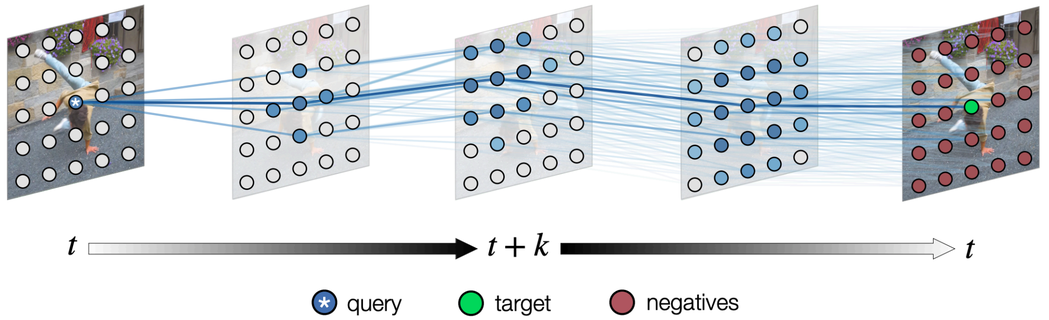
Supervision of Learning Correspondence

Self-Supervision from Temporal Signals

Forward-backward tracking as self-supervision



Wang & Jabri et al. 2019



Jabri et al. 2020

Is explicit tracking the only way to learn correspondence?

Tracking or not?

Tracking or not?

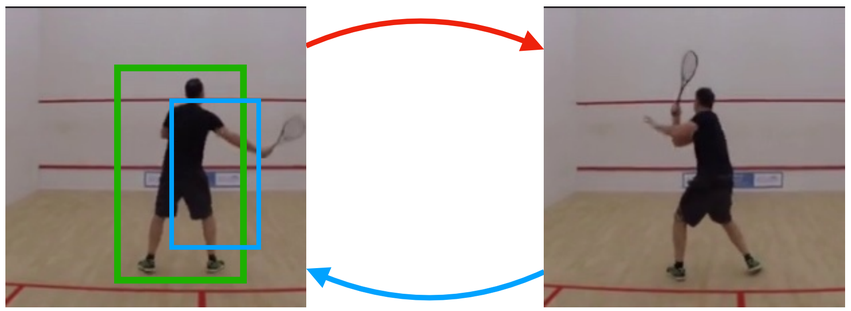
Tracking based pretext task

Tracking or not?

Tracking based pretext task

Tracking forward and backward

Maximize tracking consistence

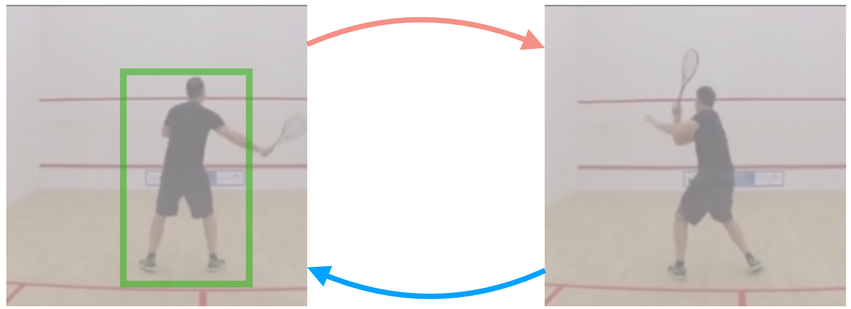


Tracking or not?

Tracking based pretext task

Tracking forward and backward

Maximize tracking consistence



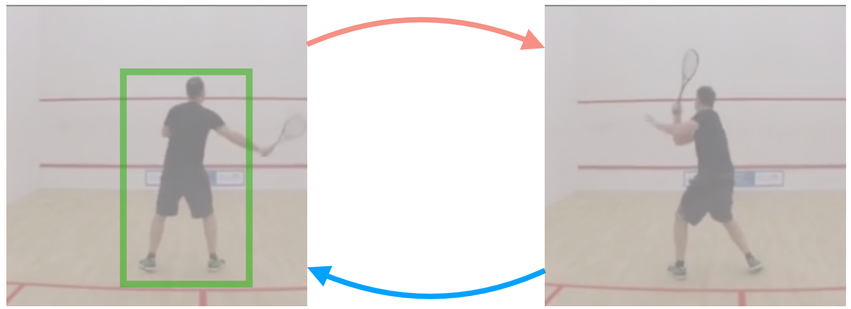
Ours

Tracking or not?

Tracking based pretext task

Tracking forward and backward

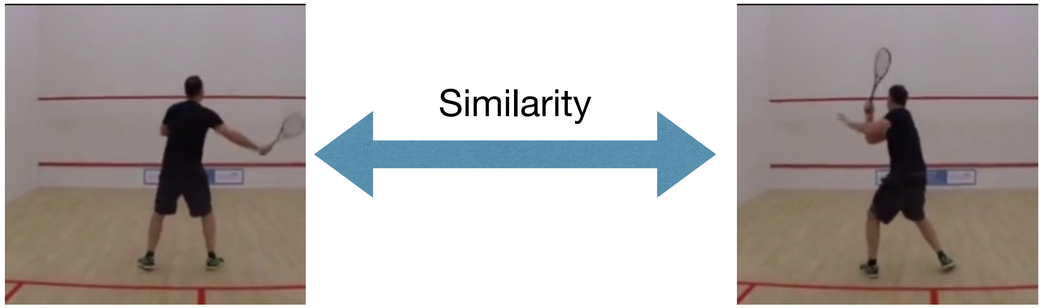
Maximize tracking consistence



Ours

Without explicit tracking

Maximize Video Frame-level Similarity

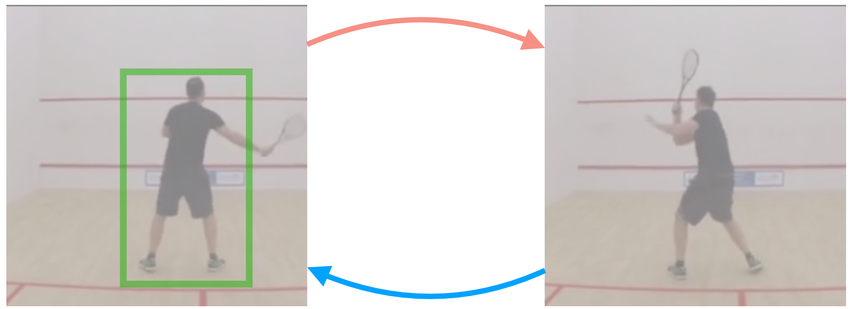


Tracking or not?

Tracking based pretext task

Tracking forward and backward

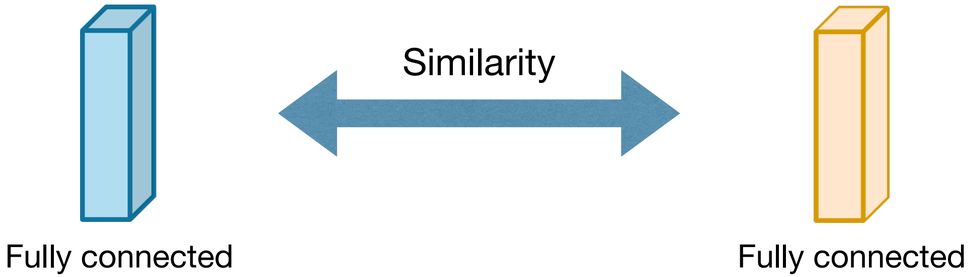
Maximize tracking consistence



Ours

Without explicit tracking

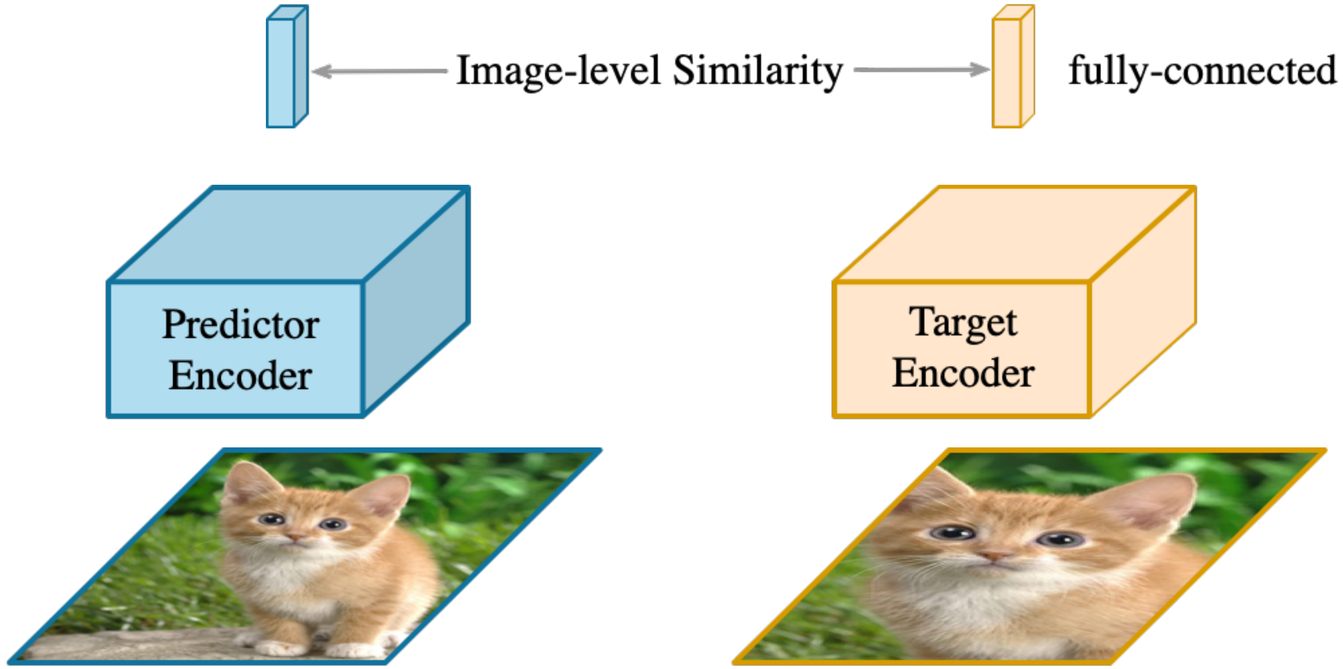
Maximize Video Frame-level Similarity



Similarity Learning

Image-level Similarity Learning

Enforce two views of the same image to have similar features in high-level fully-connected layer



He et al. 2020
Chen et al. 2020
Chen et al. 2020
Grill et al. 2020

Image-level Similarity Learning

Enforce two views of the same image to have similar features in high-level fully-connected layer

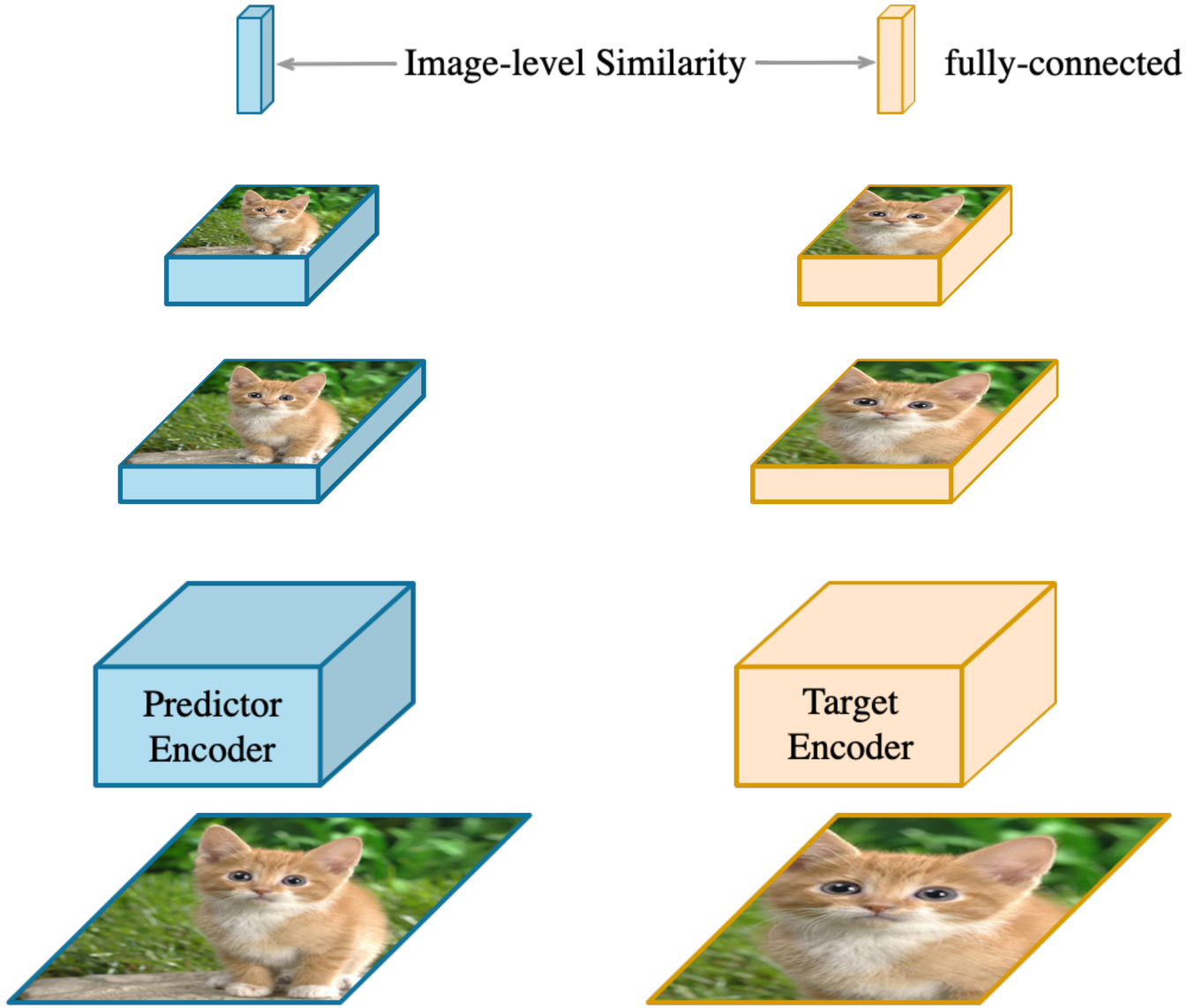
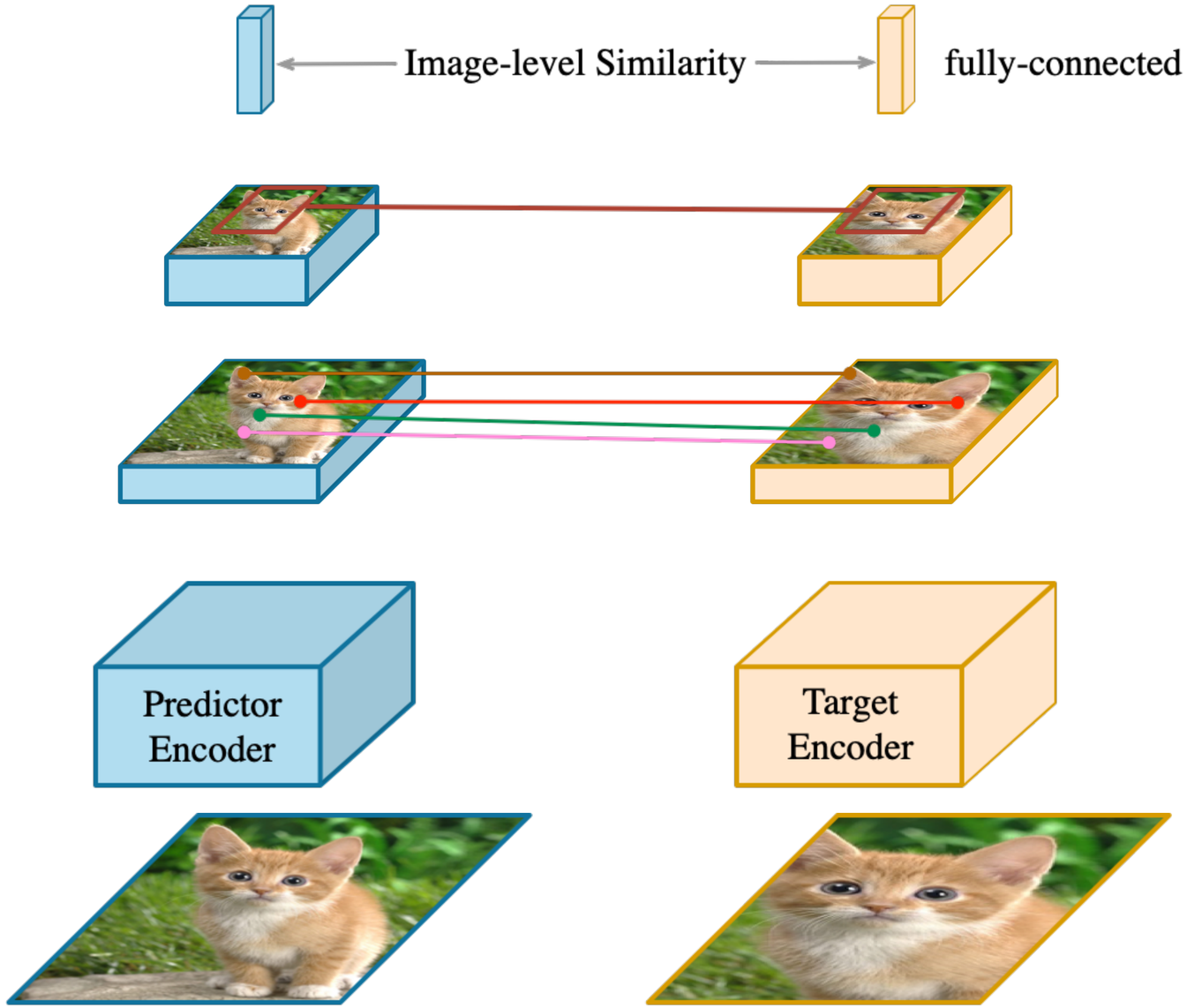


Image-level Similarity Learning

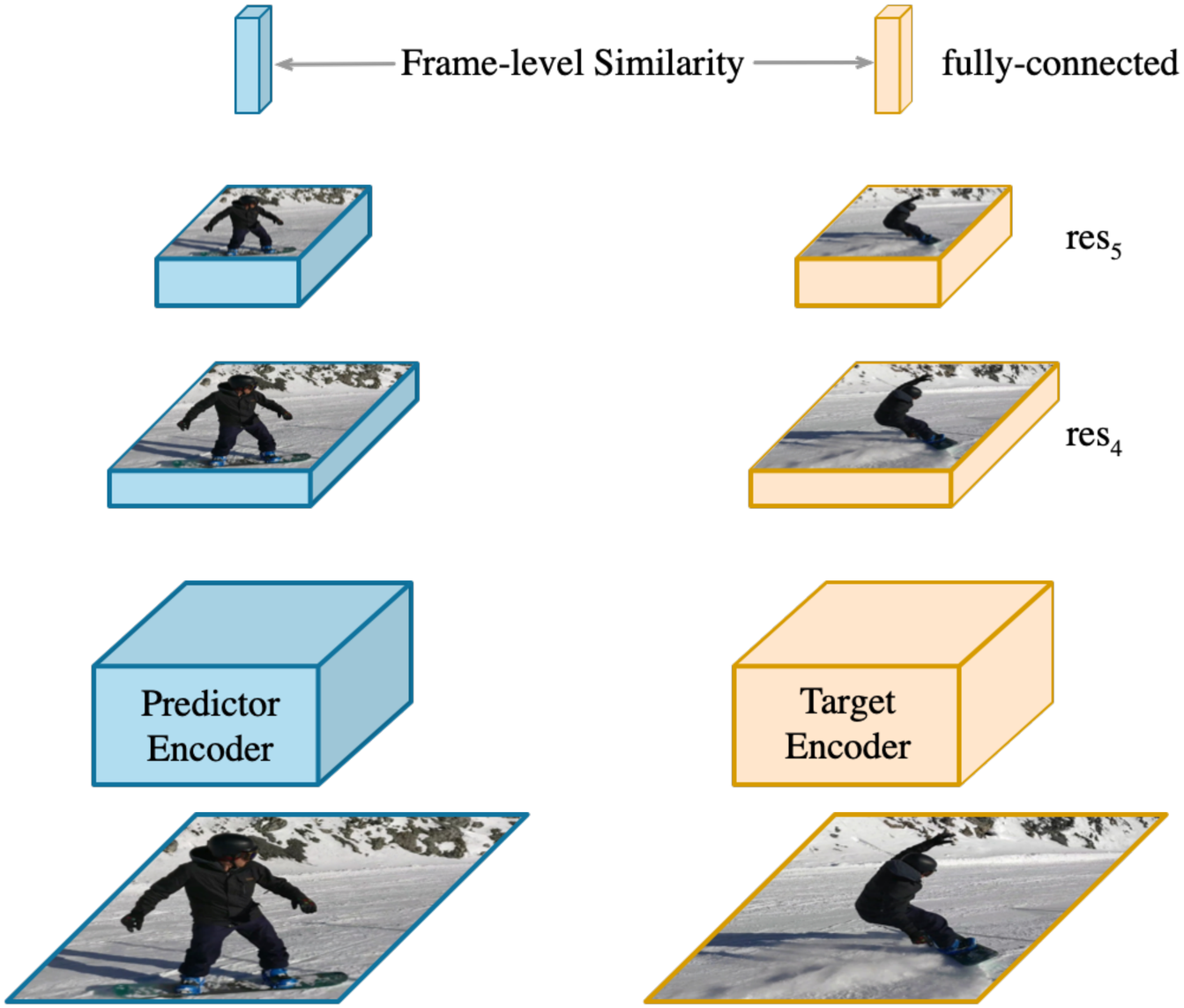
Enforce two views of the same image to have similar features in high-level fully-connected layer

The mid-level features may learn correspondence implicitly.



Video Frame-level Similarity (VFS) Learning

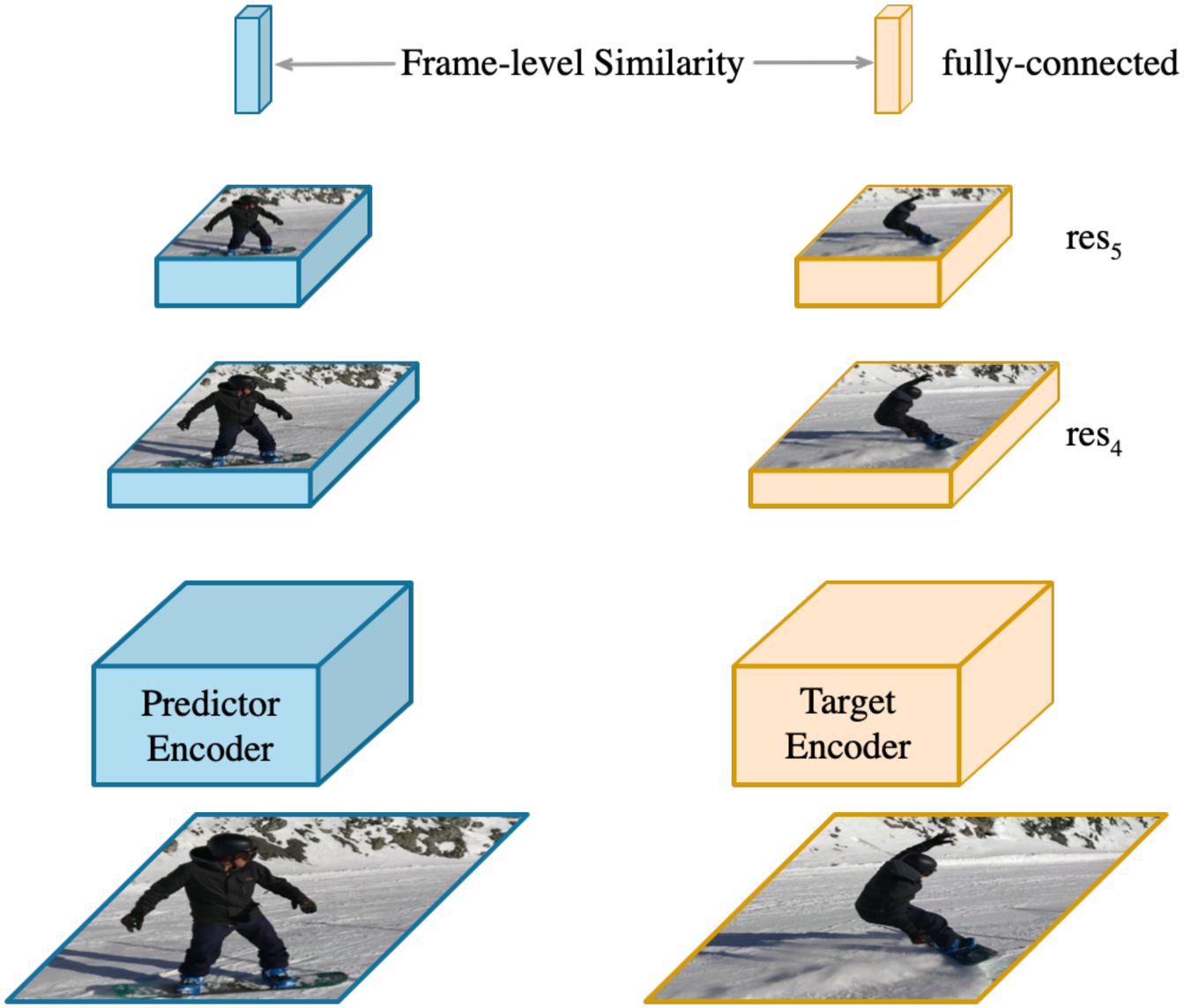
Image -> Video frame



Video Frame-level Similarity (VFS) Learning

Image -> Video frame

Enforce two frames from the same video to have similar features in high-level fully-connected layer

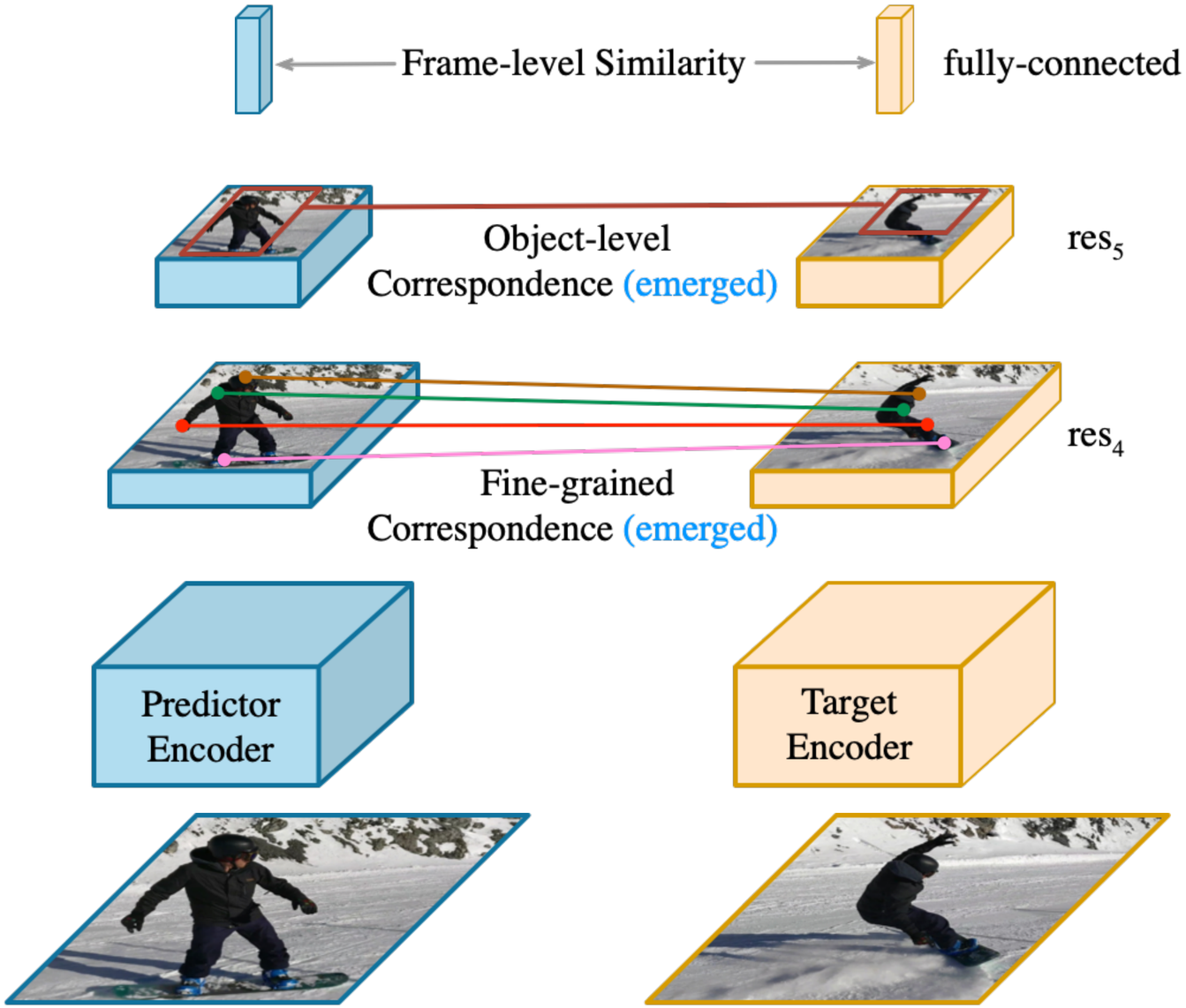


Video Frame-level Similarity (VFS) Learning

Image -> Video frame

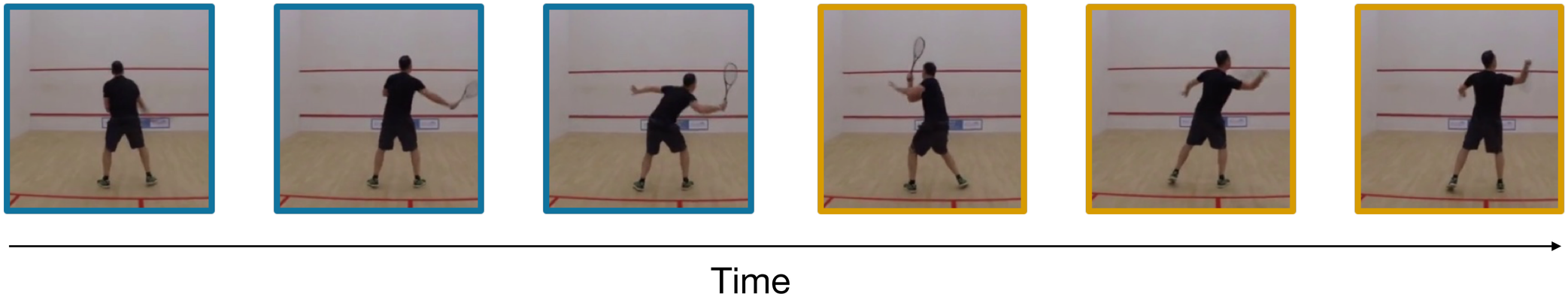
Enforce two frames from the same video to have similar features in high-level fully-connected layer

Correspondence emerges in res_4/res_5 by maximizing the frame-level similarity only



VFS Pipeline

VFS Pipeline



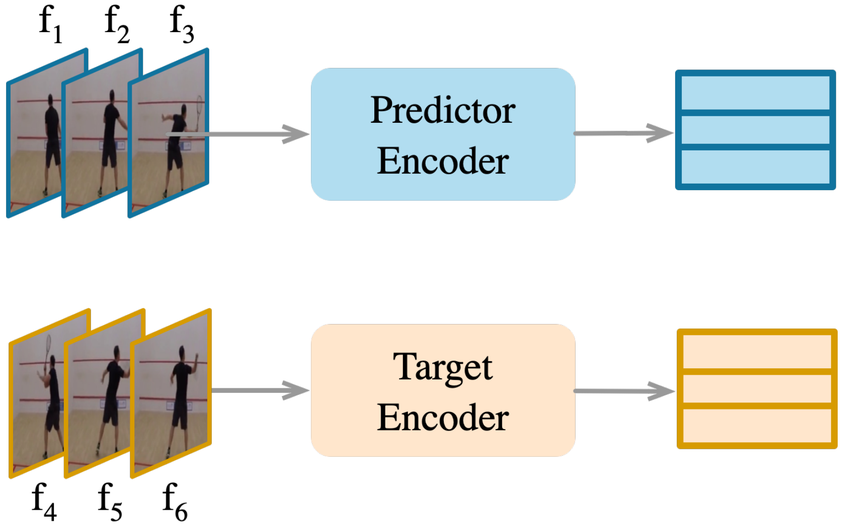
VFS Pipeline

Without negative pairs

VFS Pipeline

Without negative pairs

Encode frames with Predictor/Target Encoder

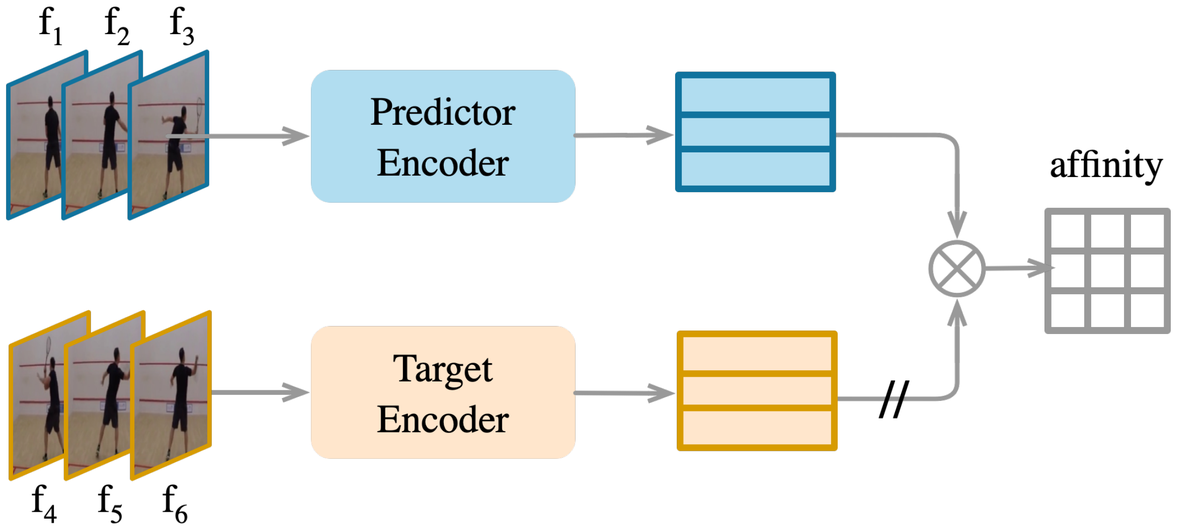


VFS Pipeline

Without negative pairs

Encode frames with Predictor/Target Encoder

Compute affinity between two branches



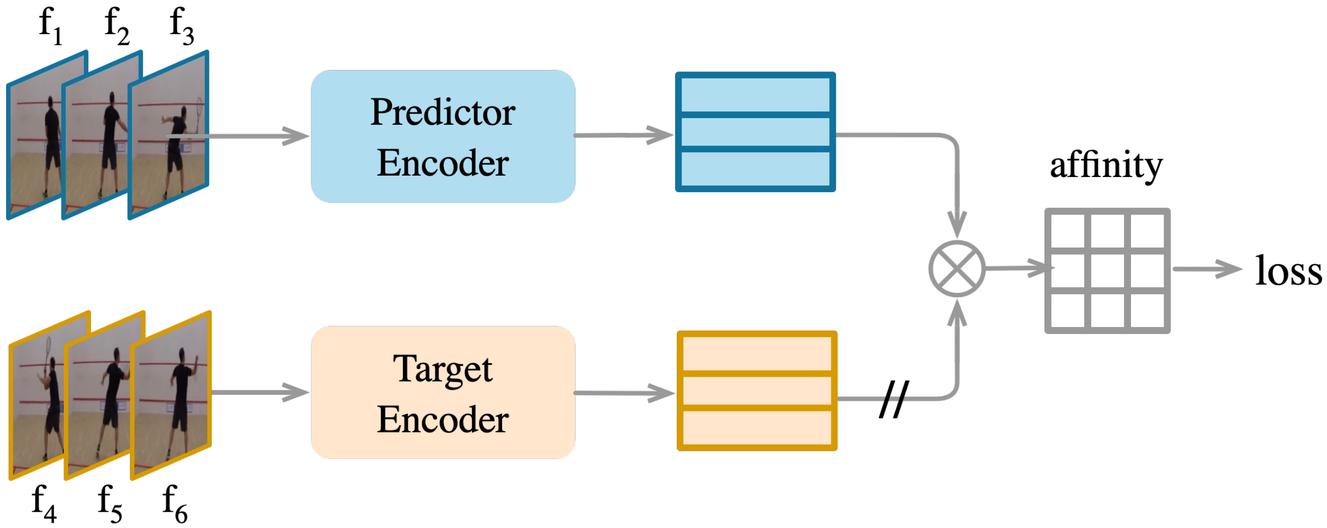
VFS Pipeline

Without negative pairs

Encode frames with Predictor/Target Encoder

Compute affinity between two branches

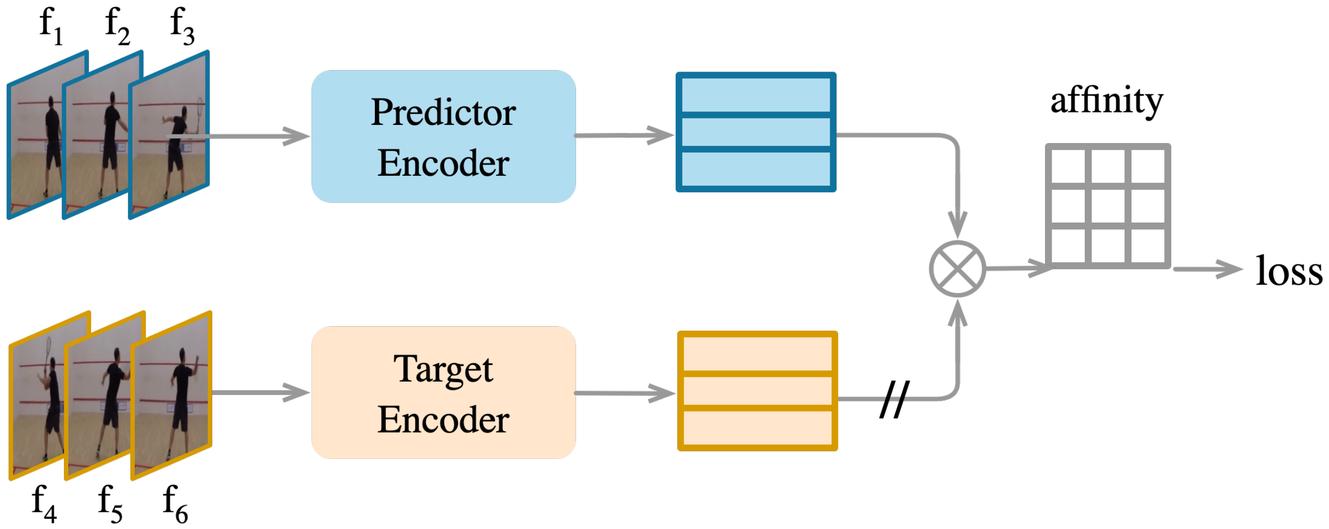
Maximize the affinity



VFS Pipeline

With negative pairs

Encode frames with Predictor/Target Encoder

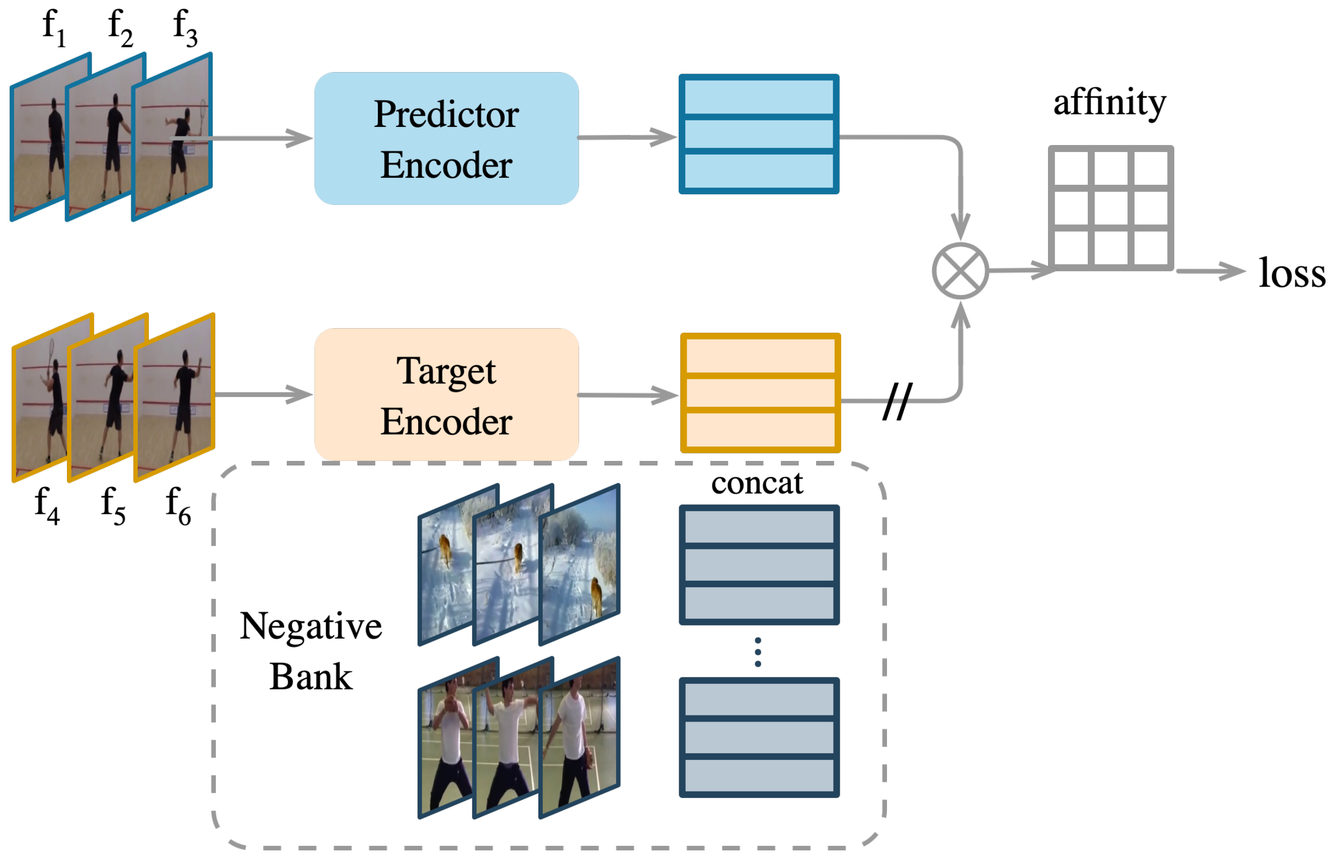


VFS Pipeline

With negative pairs

Encode frames with Predictor/Target Encoder

Concatenate features from negative bank



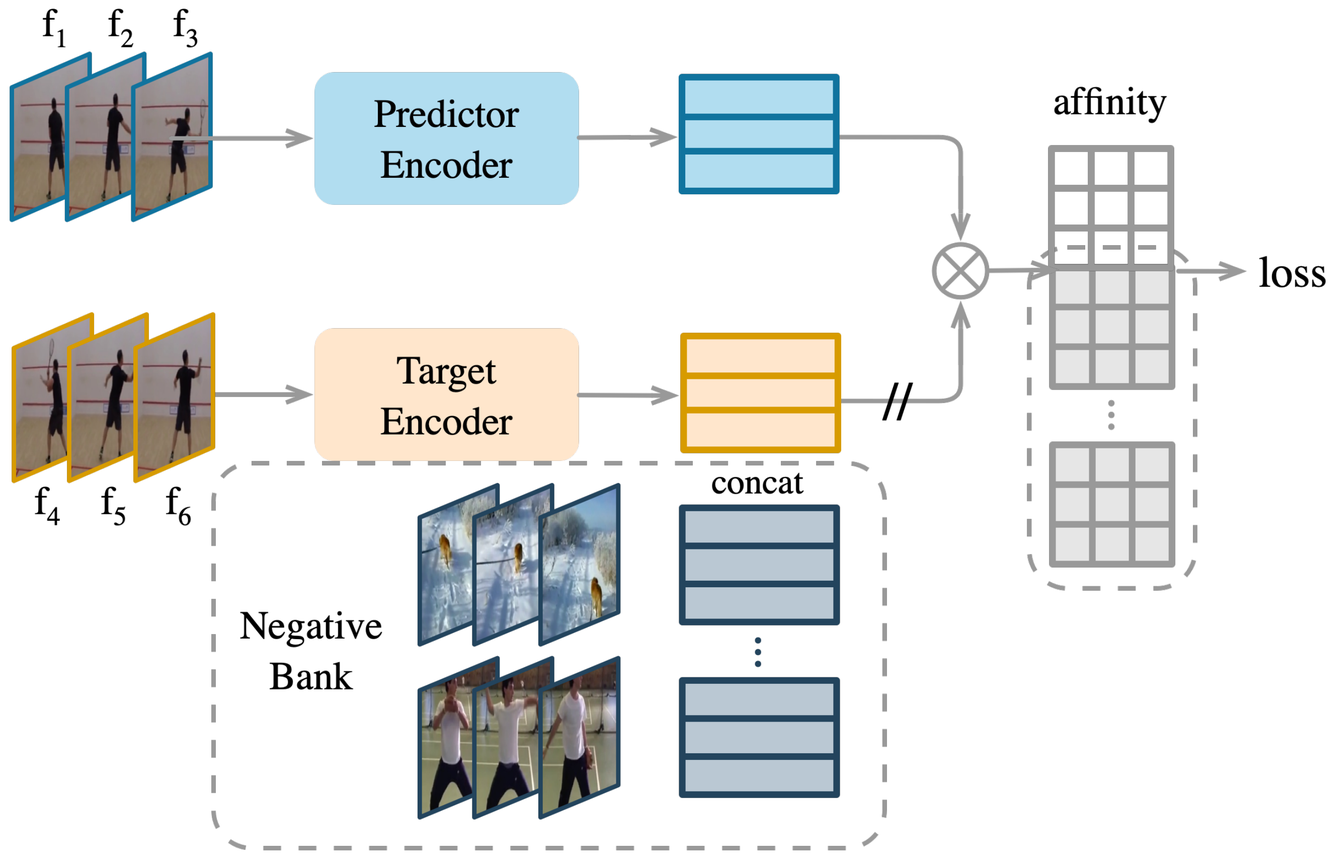
VFS Pipeline

With negative pairs

Encode frames with Predictor/Target Encoder

Concatenate features from negative bank

Compute affinity between two branches



VFS Pipeline

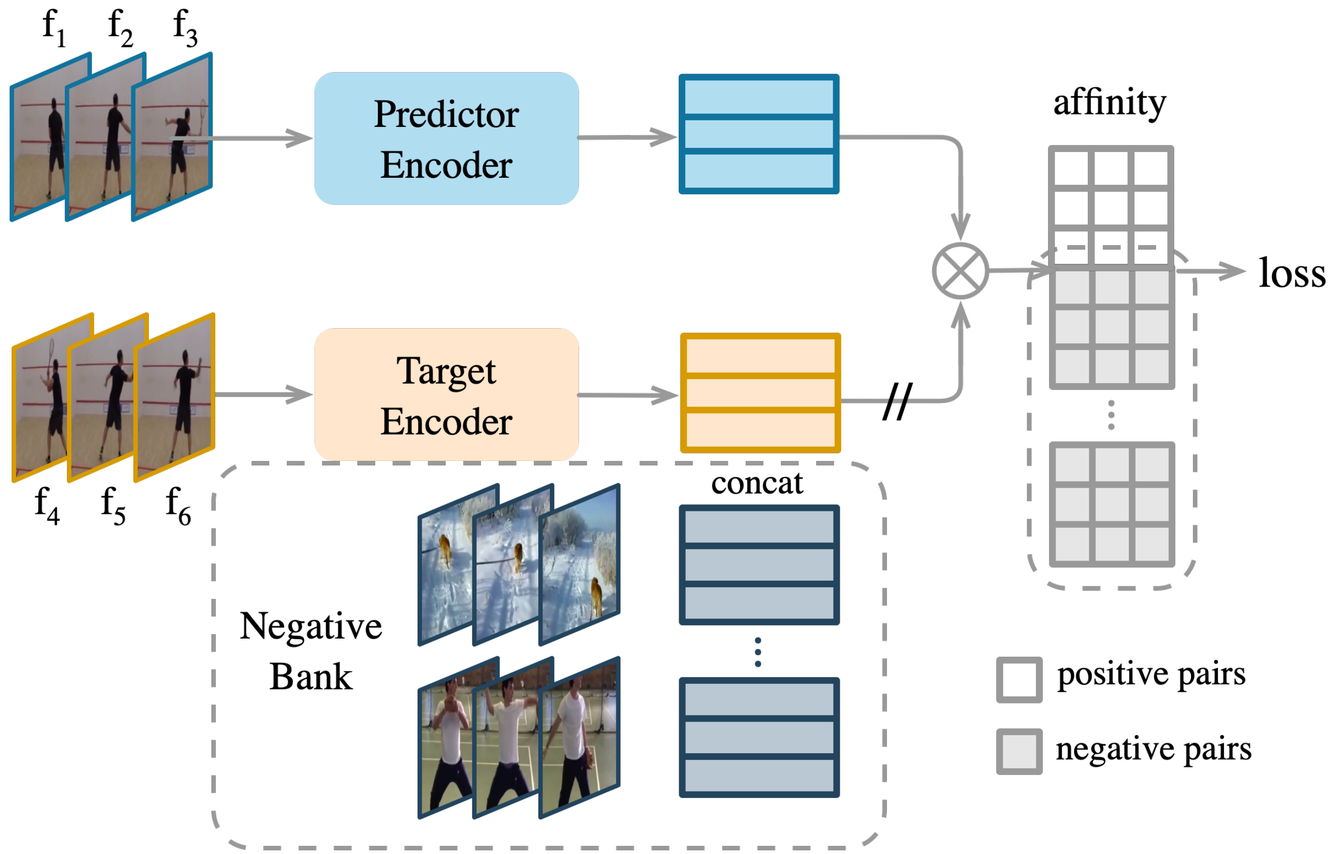
With negative pairs

Encode frames with Predictor/Target Encoder

Concatenate features from negative bank

Compute affinity between two branches

Maximize the affinity of positive pairs
Minimize the affinity of negative pairs

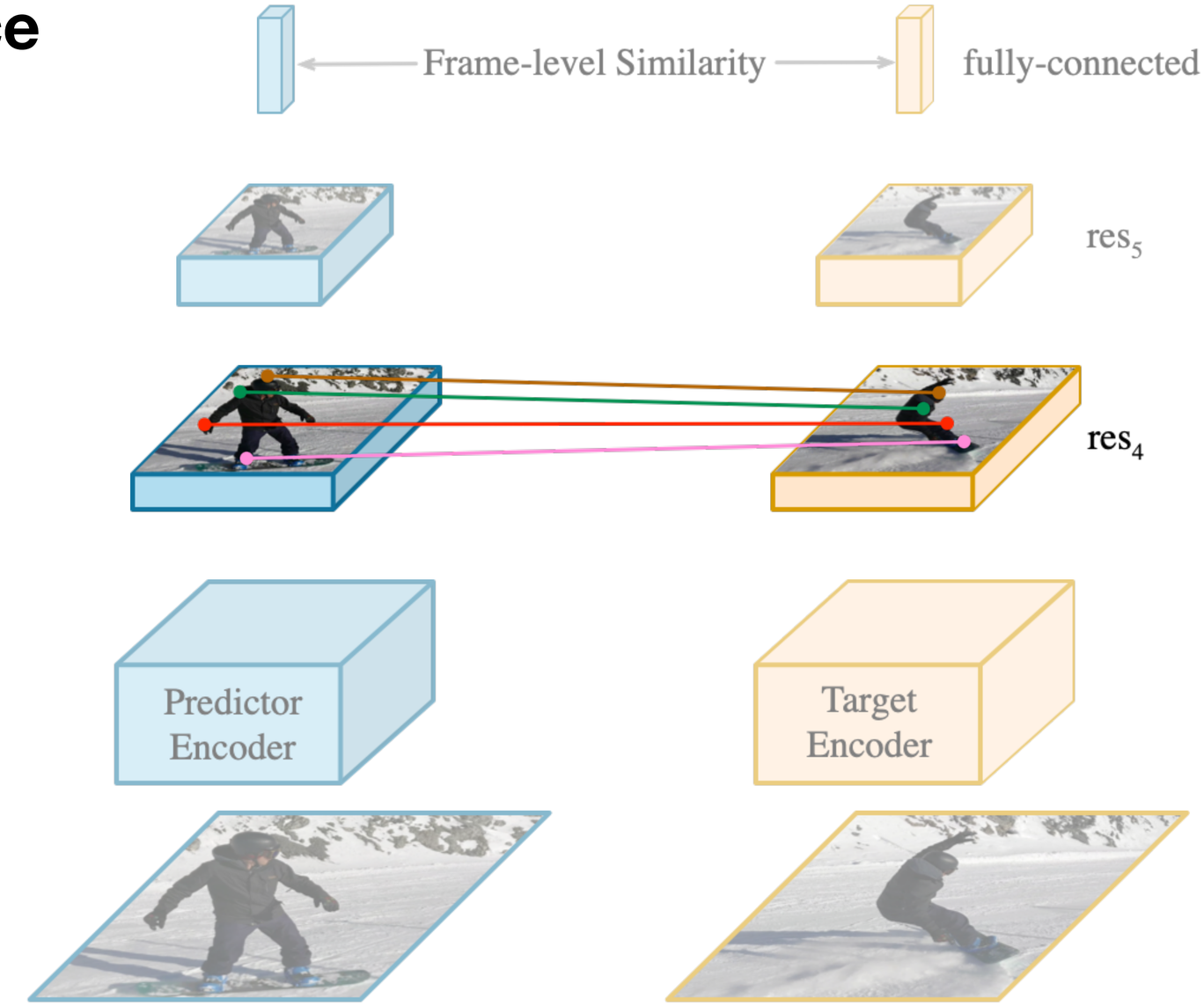


Evaluation

Fine-grained Correspondence

Evaluation

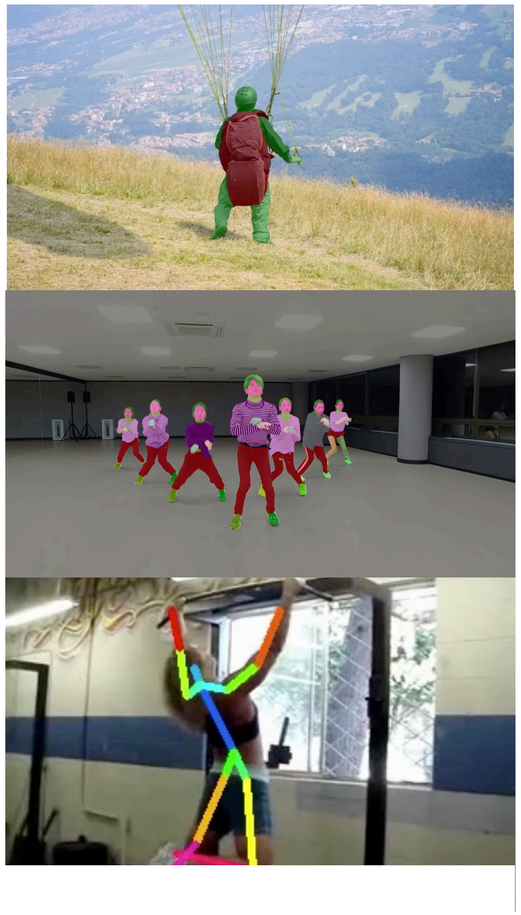
Fine-grained Correspondence

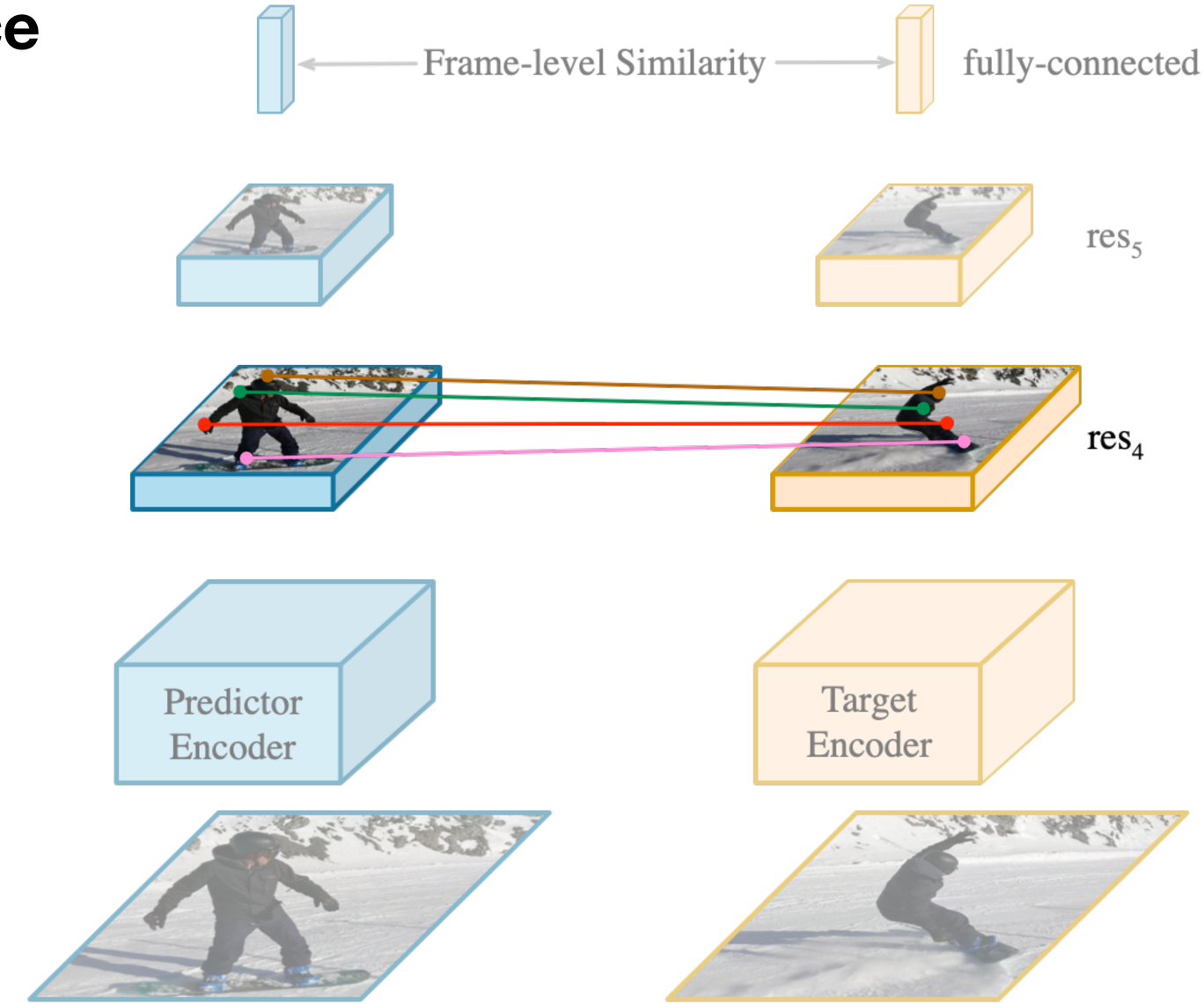


Evaluation

Fine-grained Correspondence

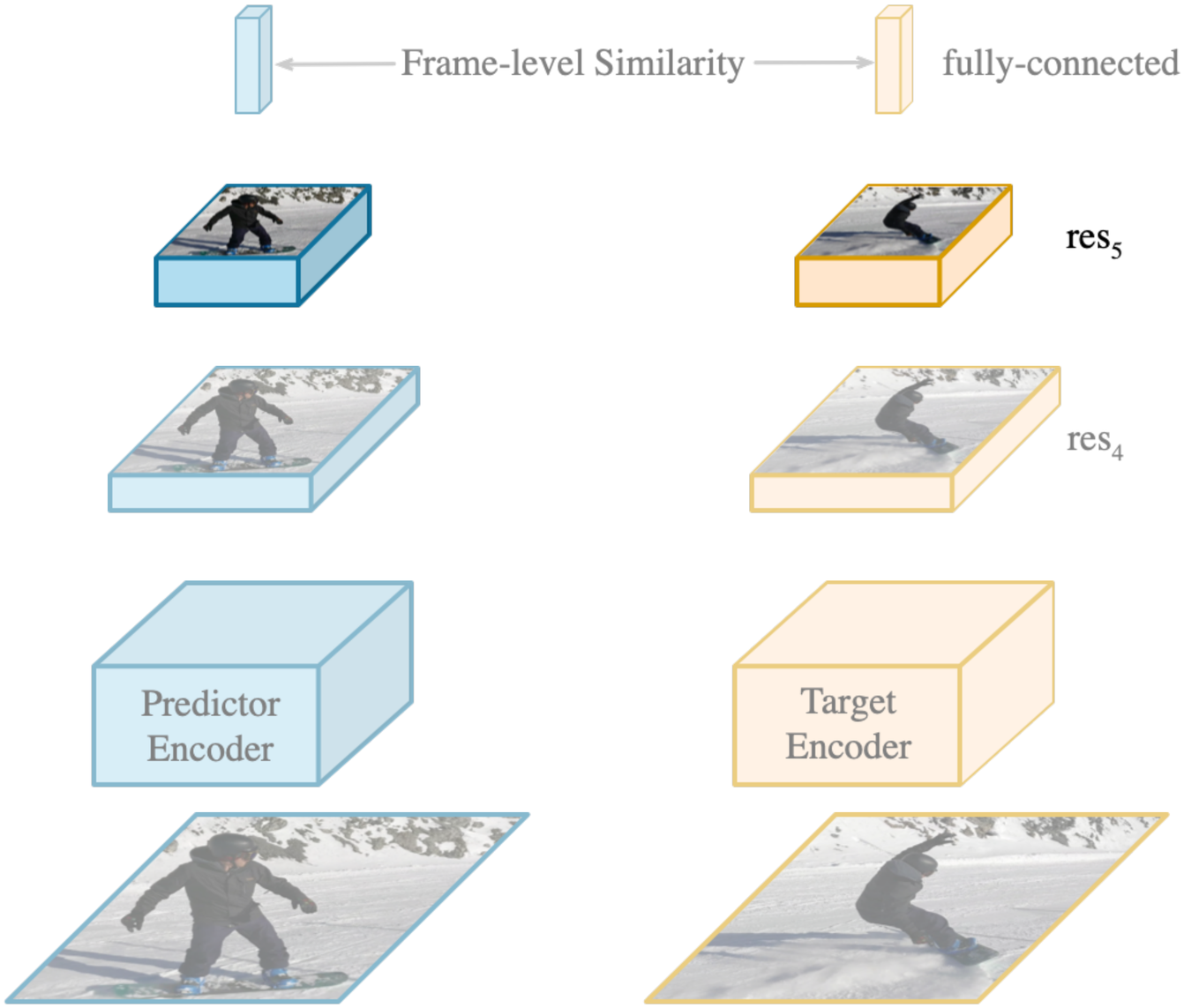
Label propagation





Evaluation

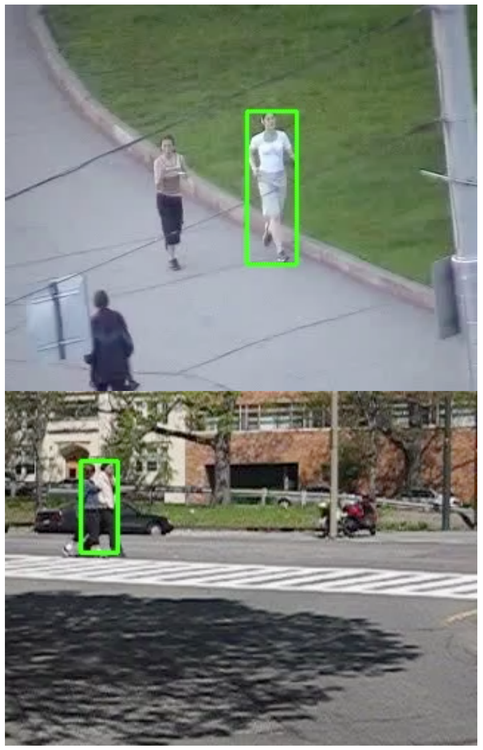
Object-level Correspondence

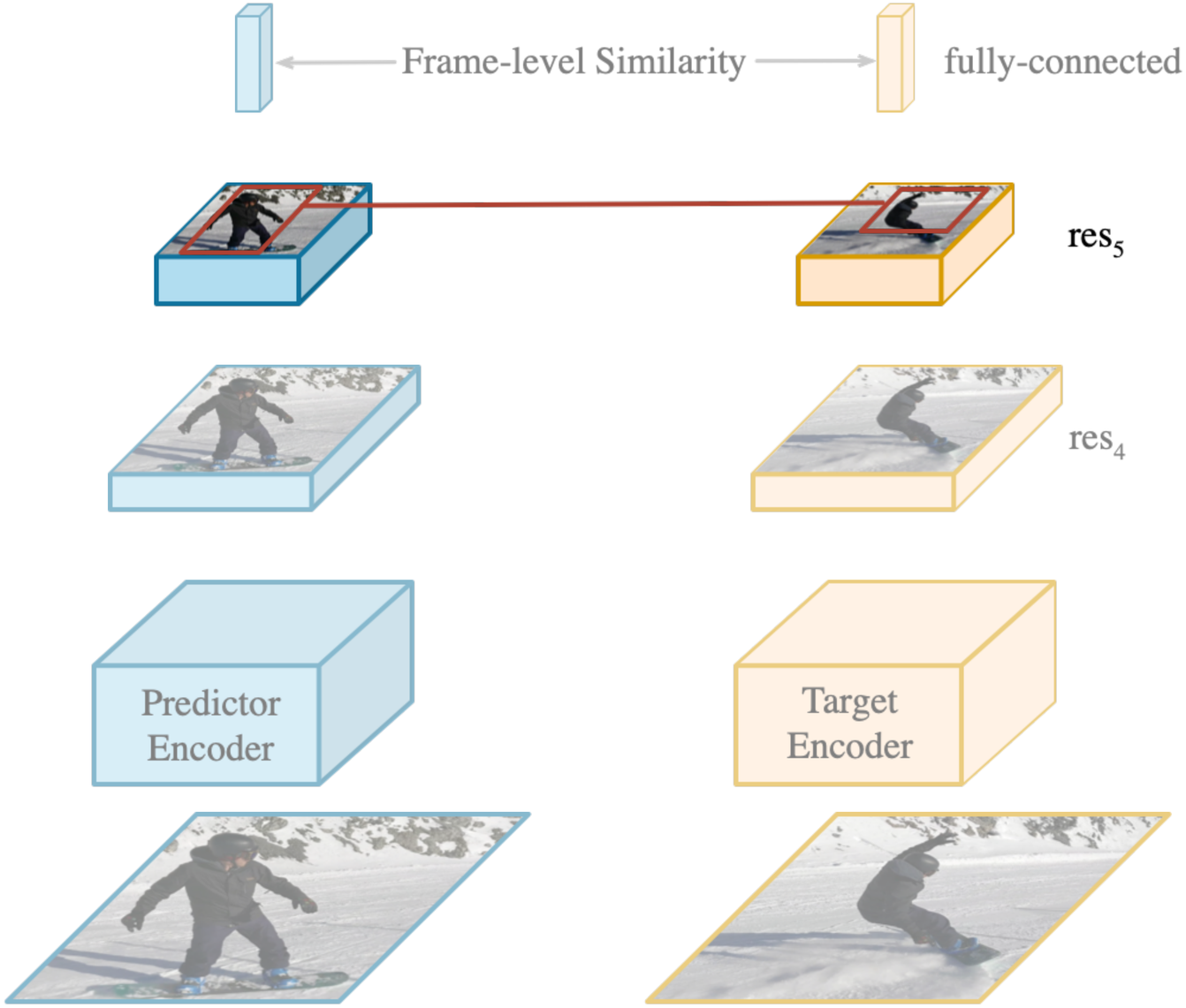


Evaluation

Object-level Correspondence

SiamFC Tracker



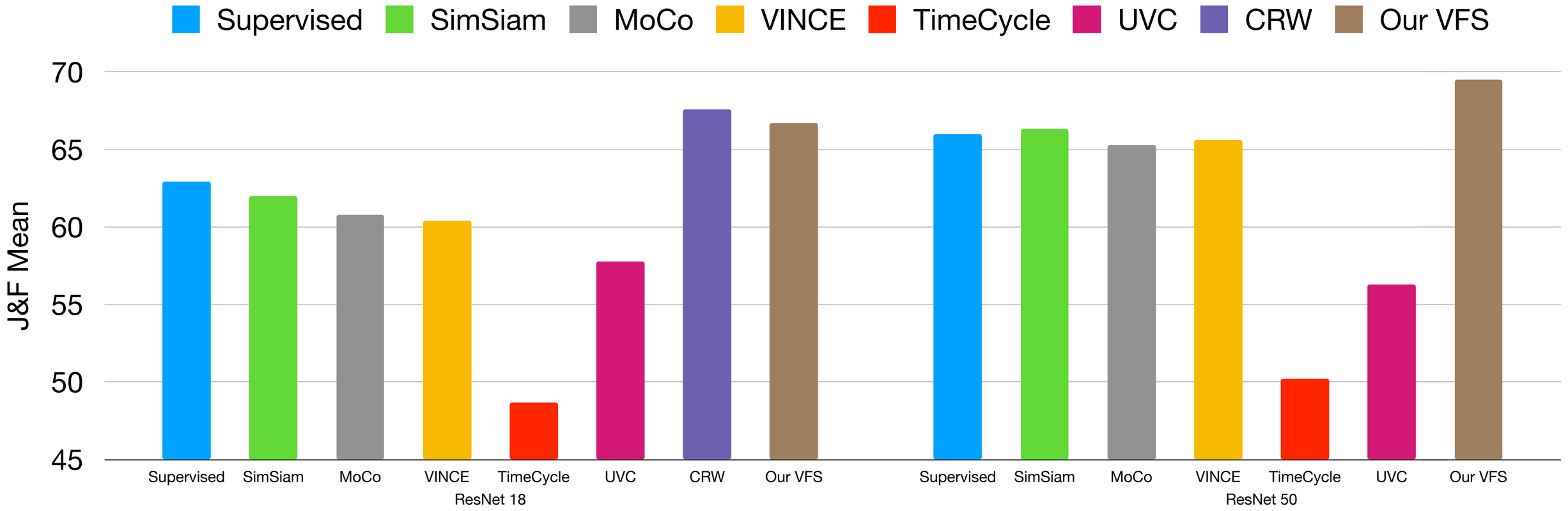


Compare with State-Of-The-Art

DAVIS 2017 Video Object Segmentation

Compare with State-Of-The-Art

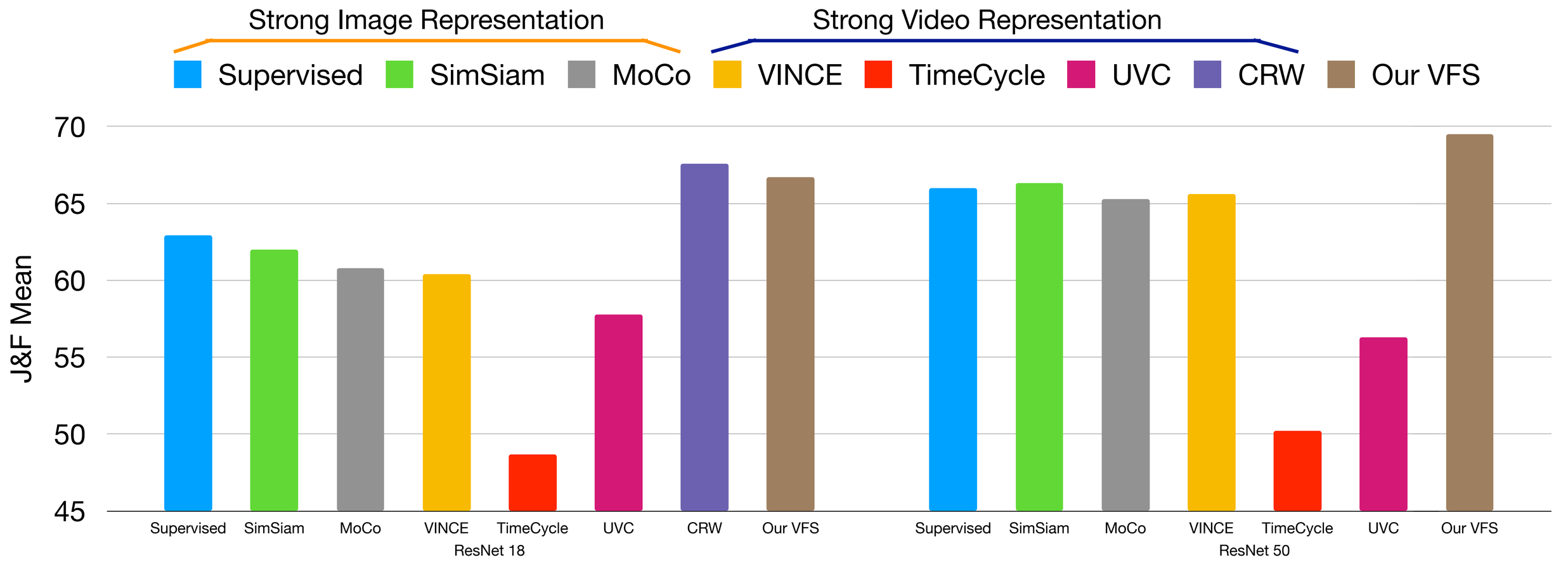
DAVIS 2017 Video Object Segmentation



SimSiam: Chen et al. (2020)
MoCo: He et al. (2020)
VINCE: Gordon et al. (2020)
TimeCycle: Wang & Jabri et al. (2019)
UVC: Li et al. (2019)
CRW: Jabri et al. (2020)

Compare with State-Of-The-Art

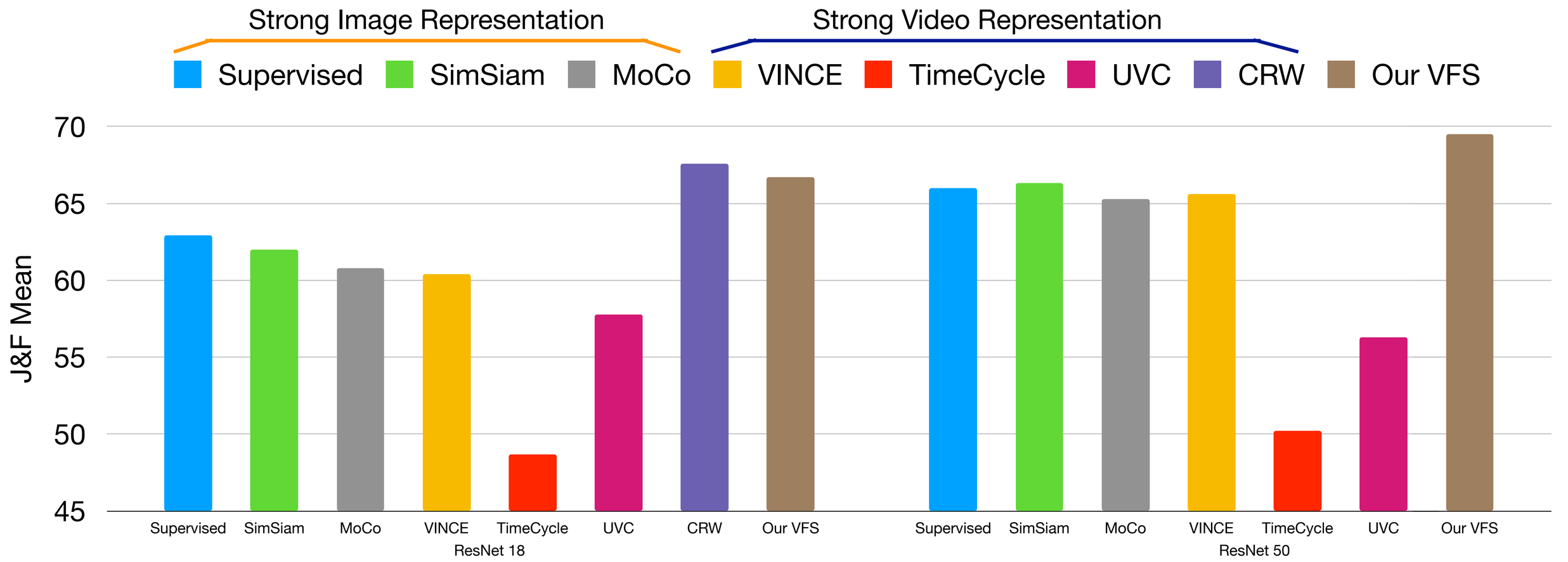
DAVIS 2017 Video Object Segmentation



SimSiam: Chen et al. (2020)
MoCo: He et al. (2020)
VINCE: Gordon et al. (2020)
TimeCycle: Wang & Jabri et al. (2019)
UVC: Li et al. (2019)
CRW: Jabri et al. (2020)

Compare with State-Of-The-Art

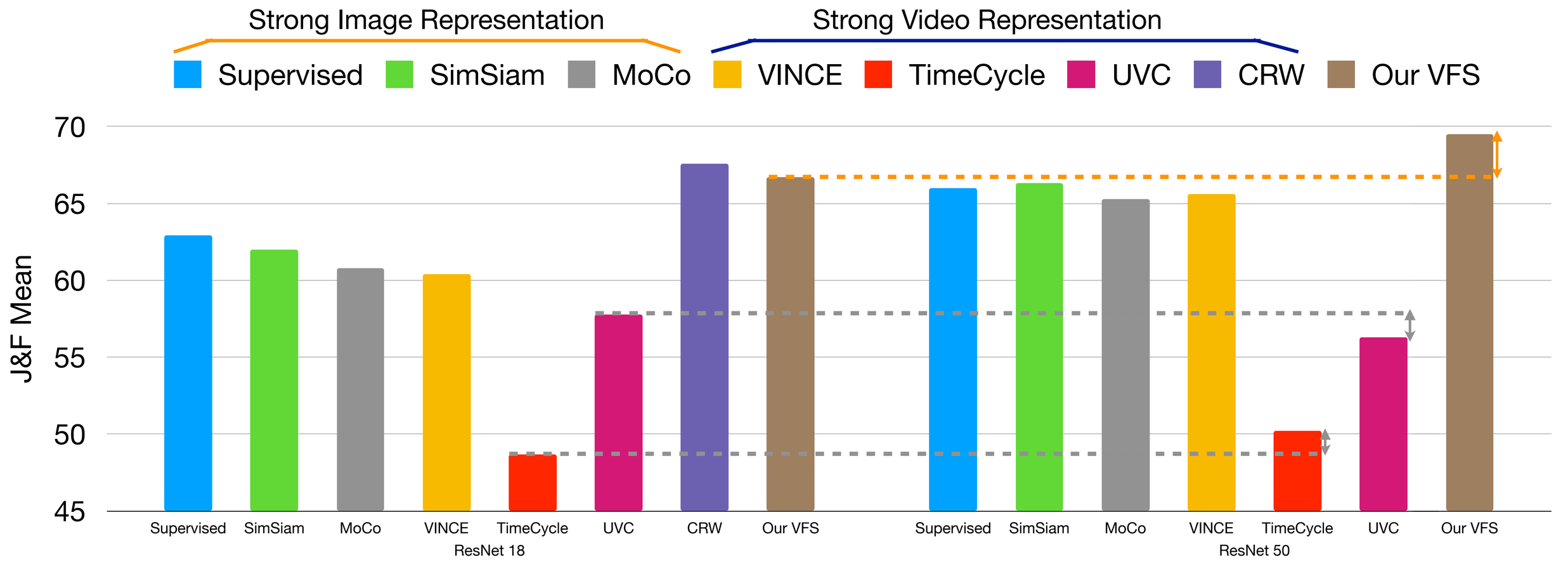
DAVIS 2017 Video Object Segmentation



SimSiam: Chen et al. (2020)
MoCo: He et al. (2020)
VINCE: Gordon et al. (2020)
TimeCycle: Wang & Jabri et al. (2019)
UVC: Li et al. (2019)
CRW: Jabri et al. (2020)

Compare with State-Of-The-Art

DAVIS 2017 Video Object Segmentation



SimSiam: Chen et al. (2020)
MoCo: He et al. (2020)
VINCE: Gordon et al. (2020)
TimeCycle: Wang & Jabri et al. (2019)
UVC: Li et al. (2019)
CRW: Jabri et al. (2020)

Qualitative Results

DAVIS 2017 Video Object Segmentation

CRW

Jabri et al. (2020)

Ours

Qualitative Results

DAVIS 2017 Video Object Segmentation

CRW

Jabri et al. (2020)

Ours

Qualitative Results

DAVIS 2017 Video Object Segmentation

CRW

Jabri et al. (2020)

Ours

Qualitative Results

DAVIS 2017 Video Object Segmentation

CRW

Jabri et al. (2020)

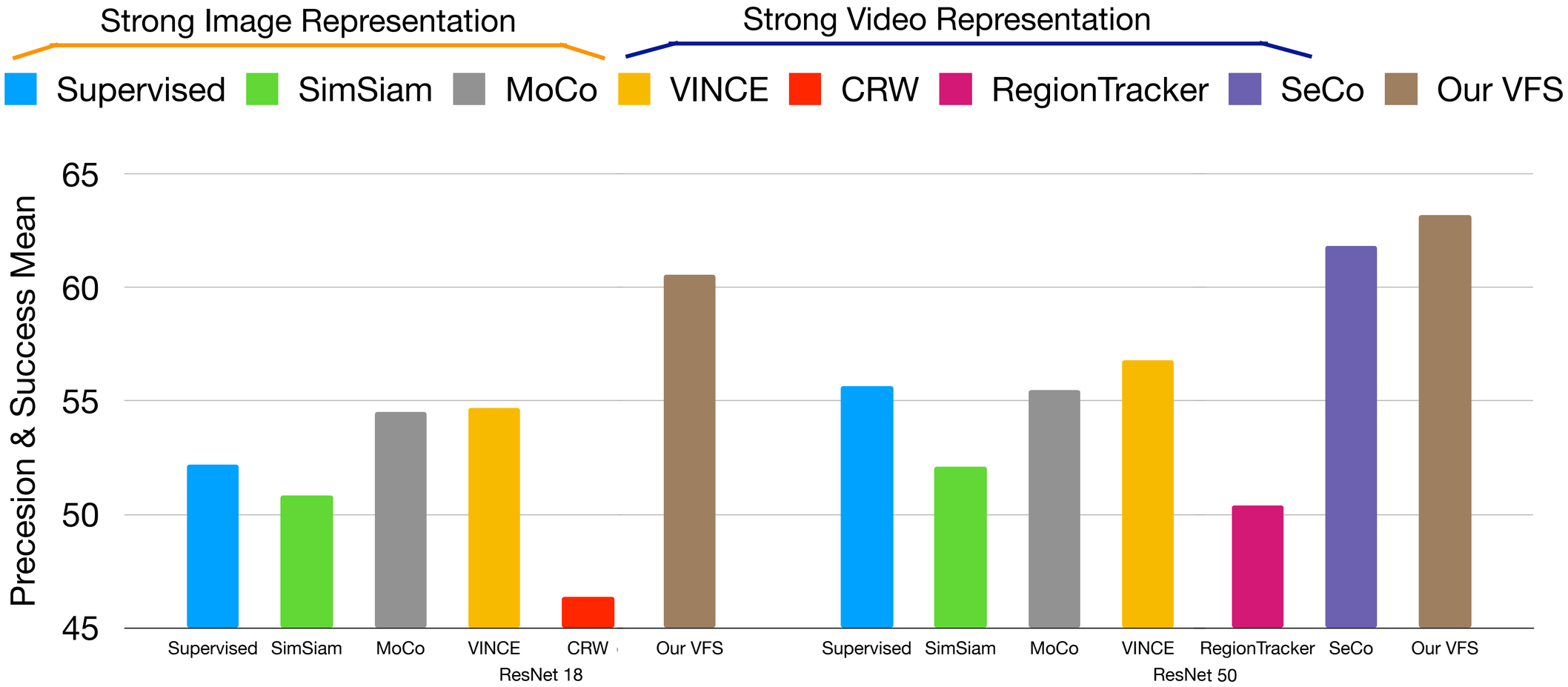
Ours

Compare with State-Of-The-Art

OTB-100 Visual Object Tracking

Compare with State-Of-The-Art

OTB-100 Visual Object Tracking



SimSiam: Chen et al. (2020)
MoCo: He et al. (2020)
VINCE: Gordon et al. (2020)
Supervised: He et al. (2015)
RegionTracker: Purushwalkam et al. (2020)
CRW: Jabri et al. (2020)

Qualitative Results

OTB-100 Visual Object Tracking

CRW

Jabri et al. (2020)

Ours

Qualitative Results

OTB-100 Visual Object Tracking

CRW

Jabri et al. (2020)

Ours

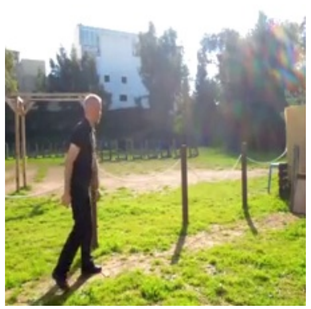
Findings and insights

Findings and insights

Color Augmentation

Findings and insights

Color Augmentation





Findings and insights

Color Augmentation

Findings and insights

Color Augmentation

No aug

Findings and insights

Color Augmentation

No aug

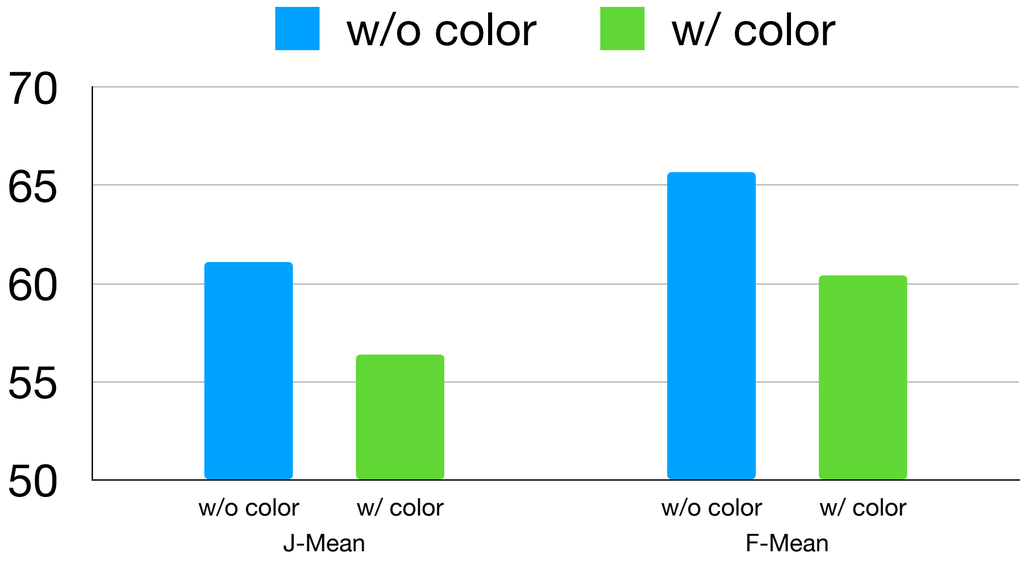
Color aug

Findings and insights

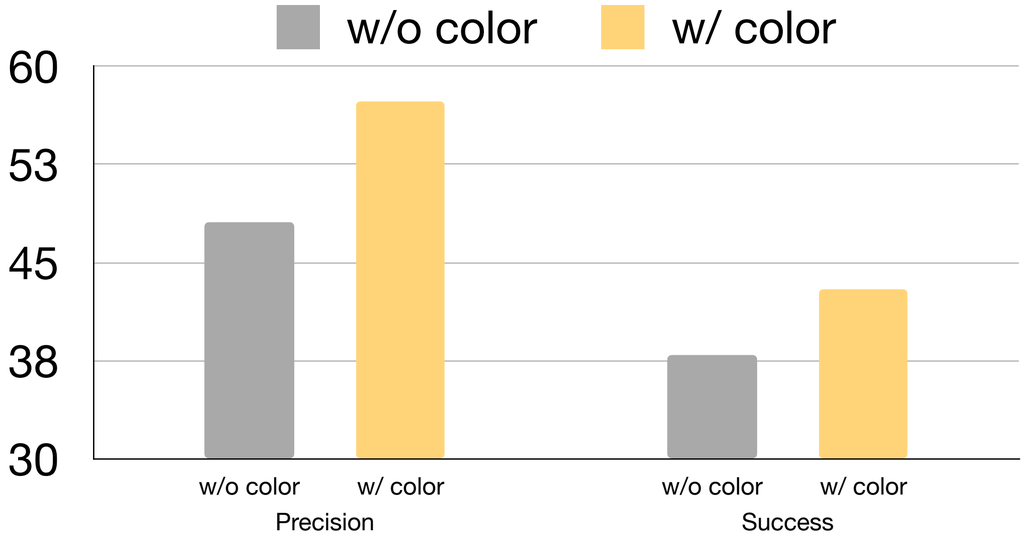
Color Augmentation

Findings and insights

Color Augmentation



Fine-grained Correspondence on DAVIS 2017



Object-level Correspondence on OTB-100

Findings and insights

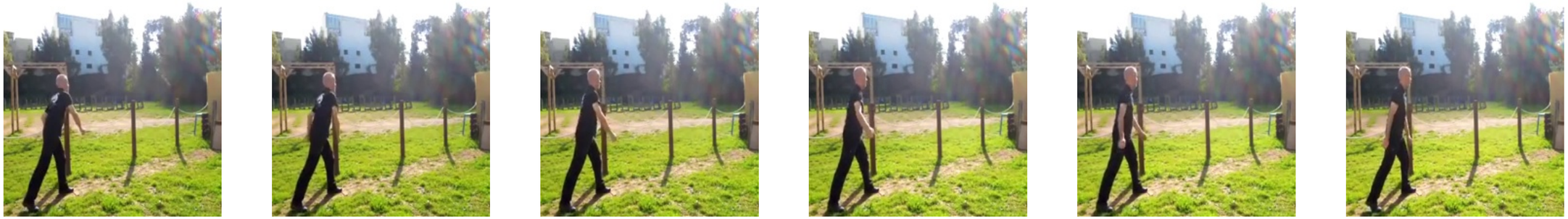
Temporal Sampling

Findings and insights

Temporal Sampling

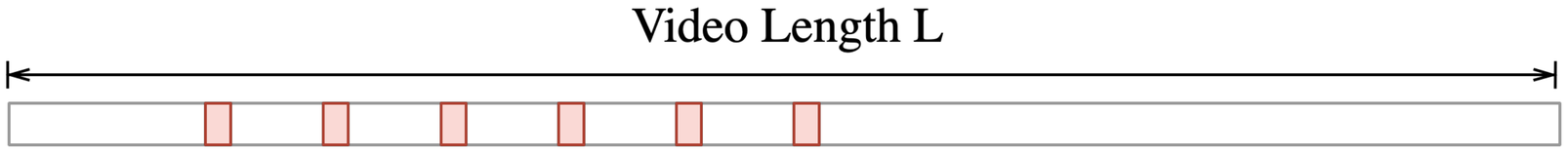
Findings and insights

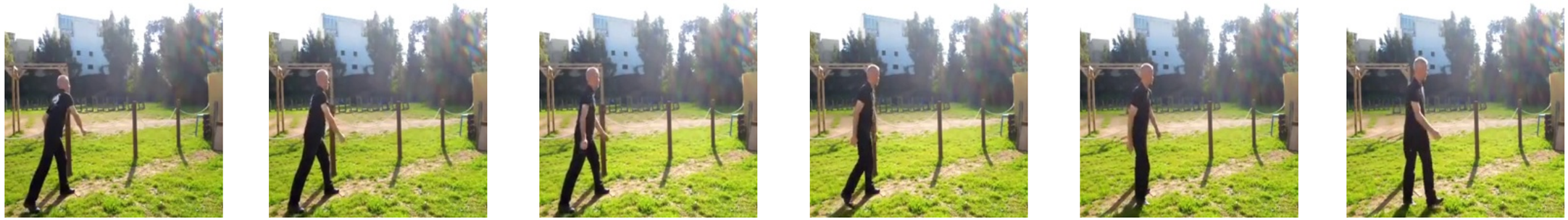
Temporal Sampling



Findings and insights

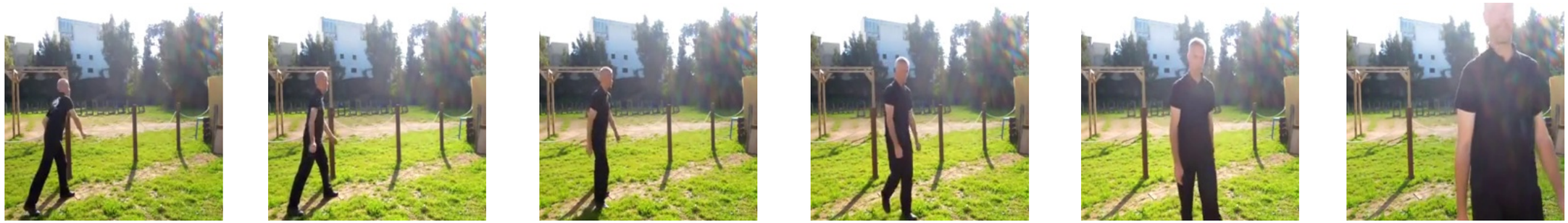
Temporal Sampling





Findings and insights

Temporal Sampling

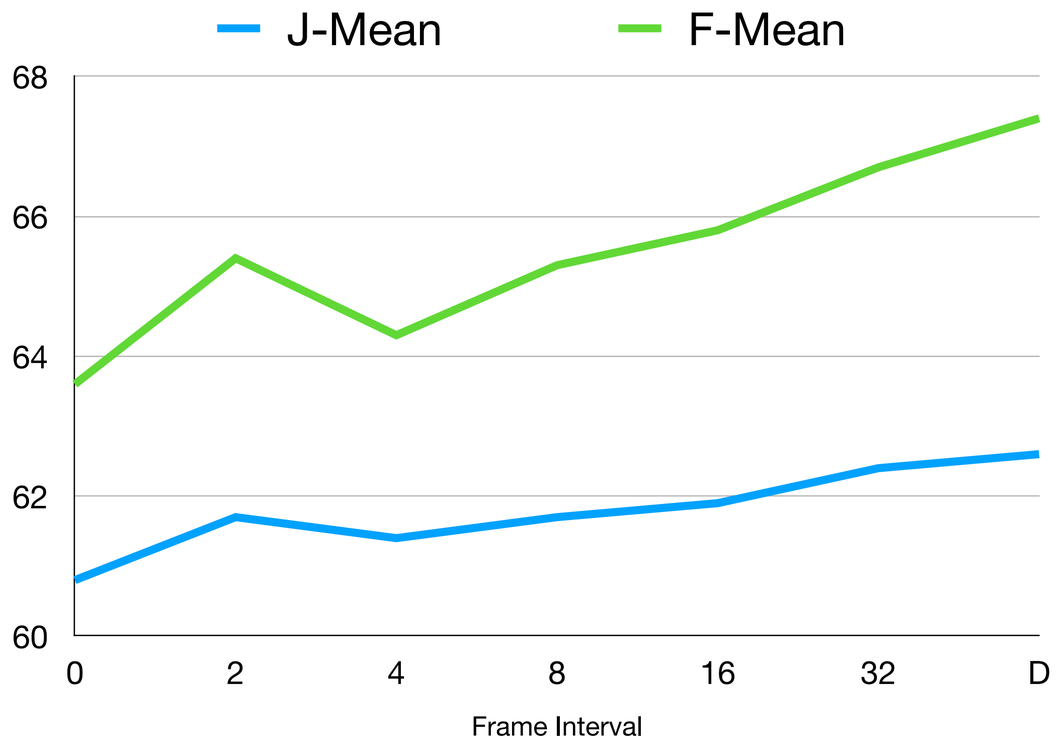


Findings and insights

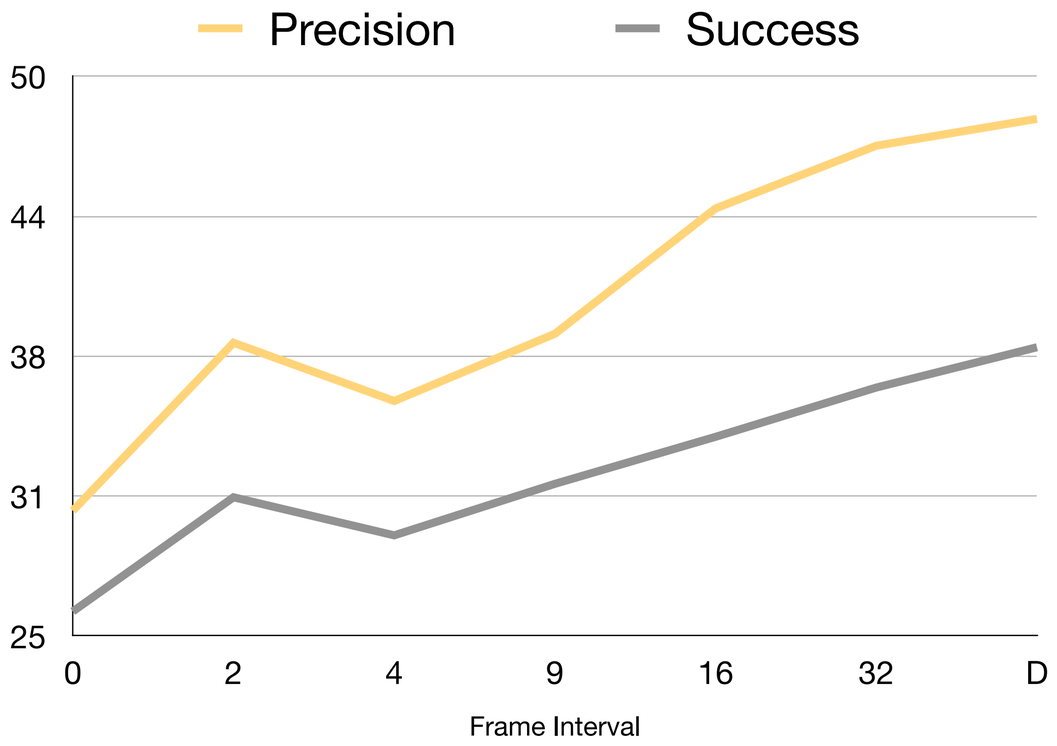
Temporal Sampling

Findings and insights

Temporal Sampling



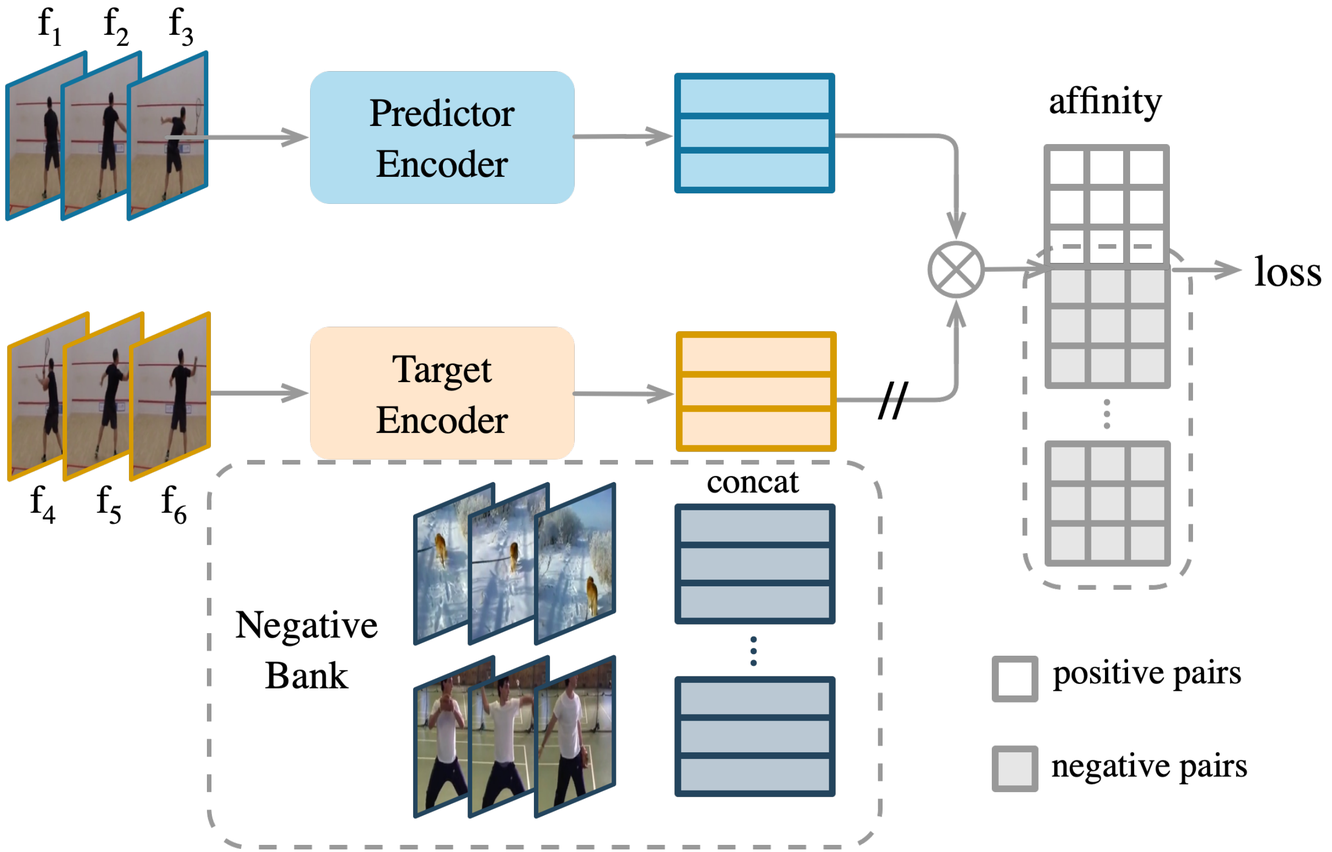
Fine-grained Correspondence on DAVIS 2017



Object-level Correspondence on OTB-100

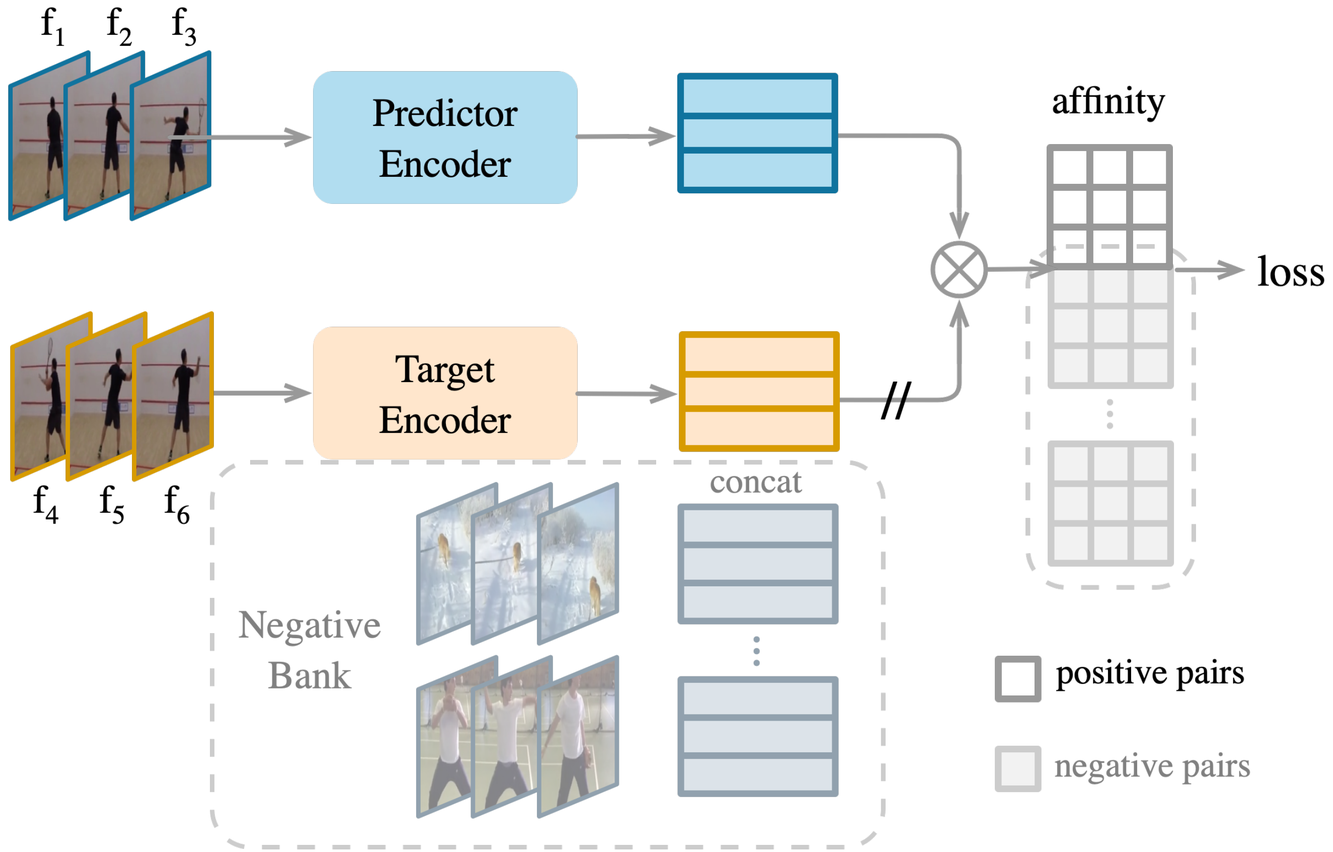
Findings and insights

Negative pairs



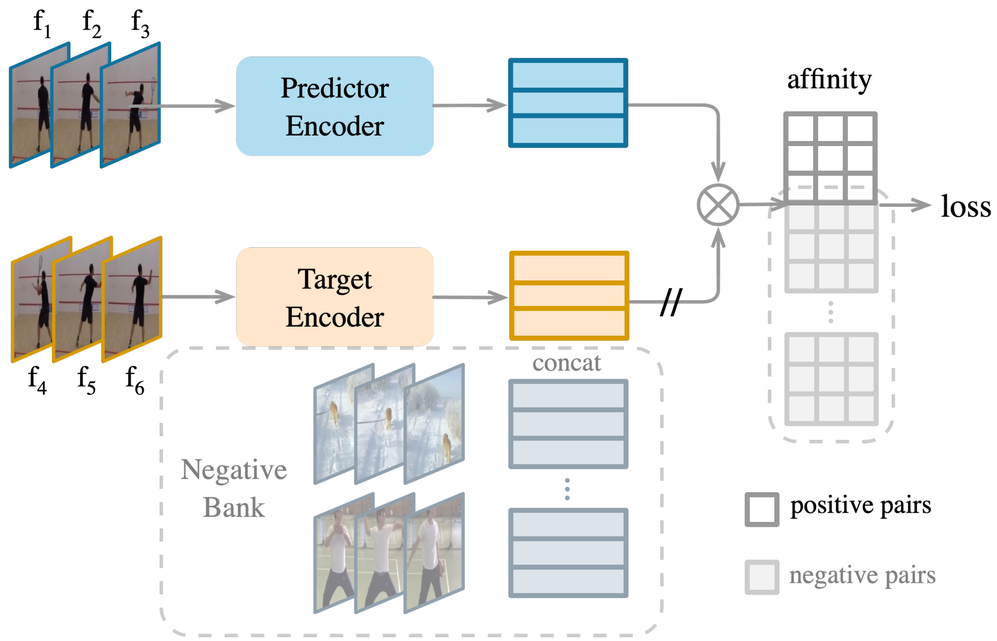
Findings and insights

Negative pairs



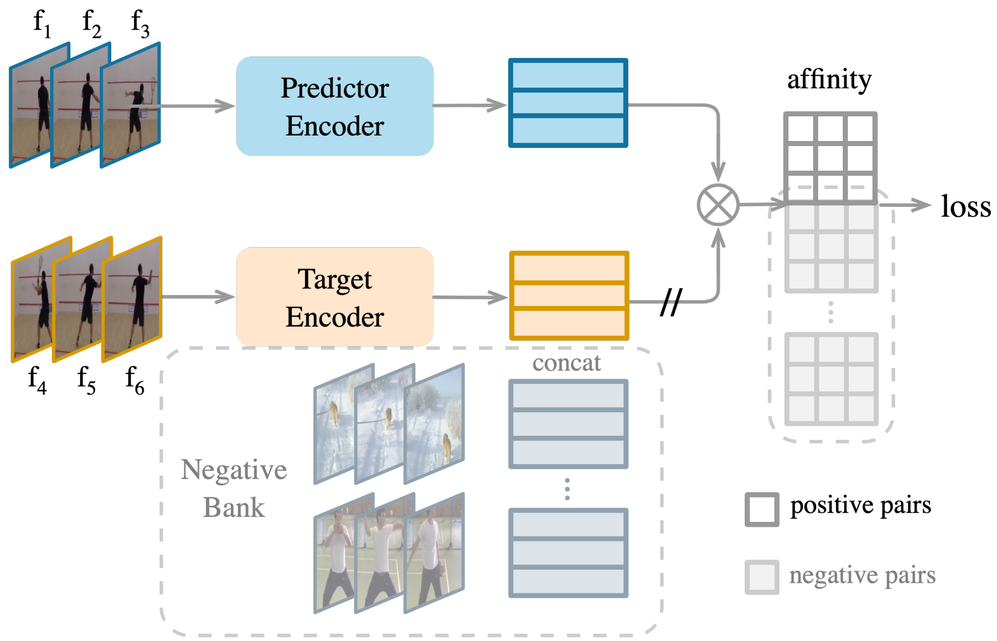
Findings and insights

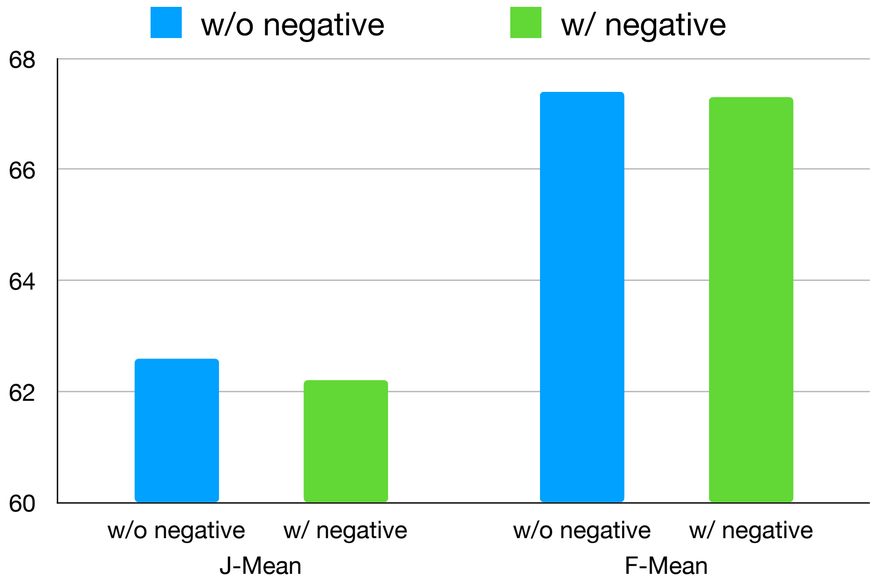
Negative pairs



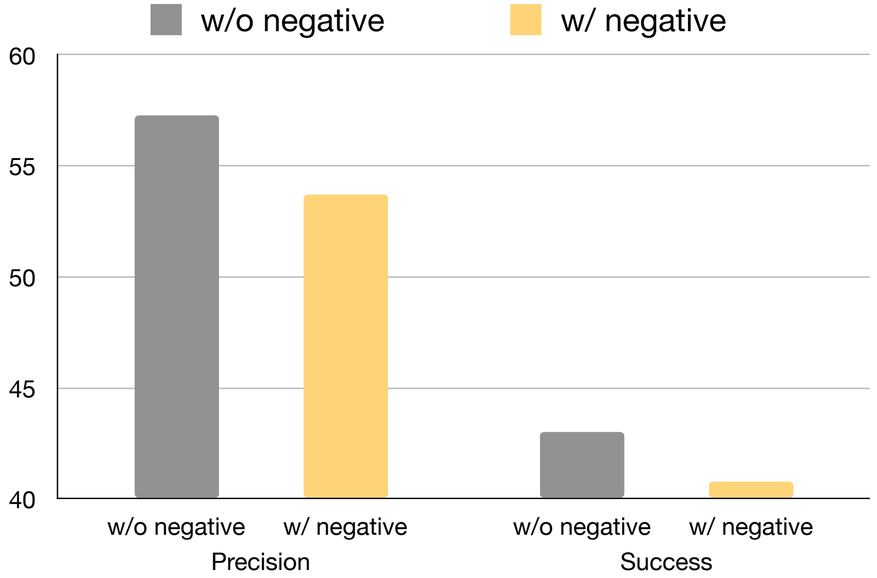
Findings and insights

Negative pairs





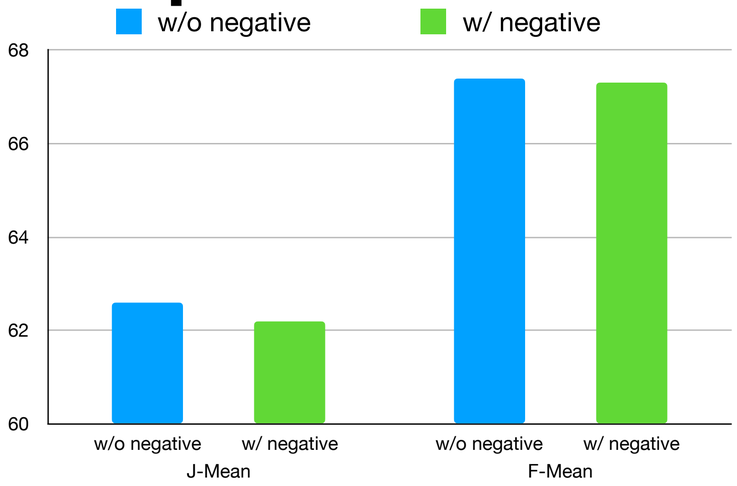
Fine-grained Correspondence on DAVIS 2017



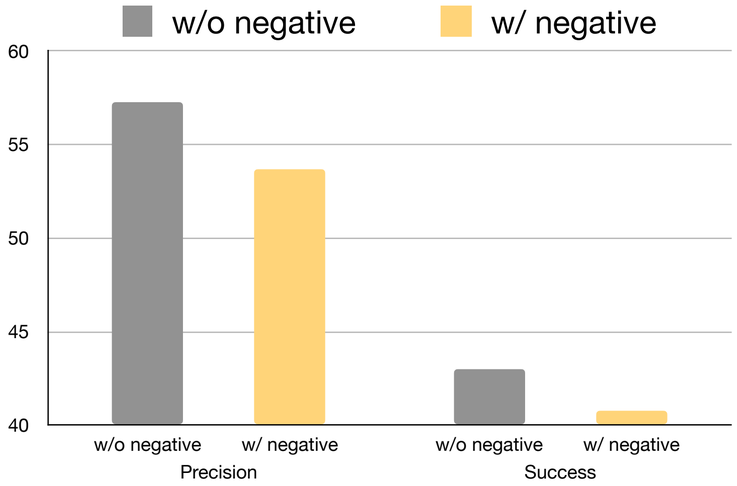
Object-level Correspondence on OTB-100

Findings and insights

Negative pairs



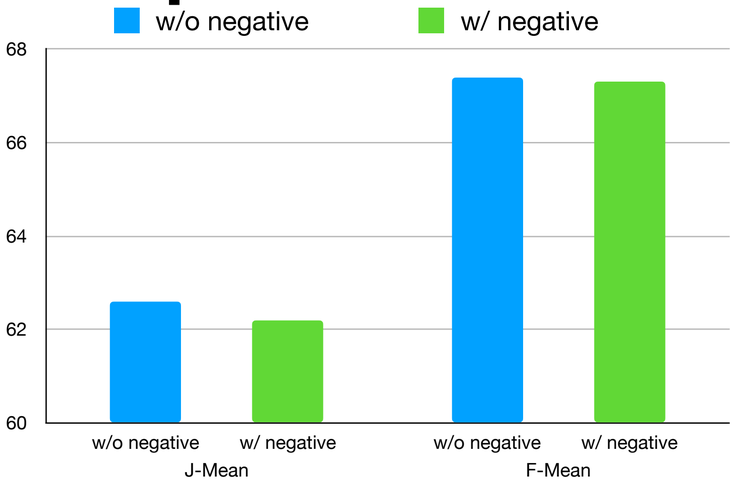
Fine-grained Correspondence on DAVIS 2017



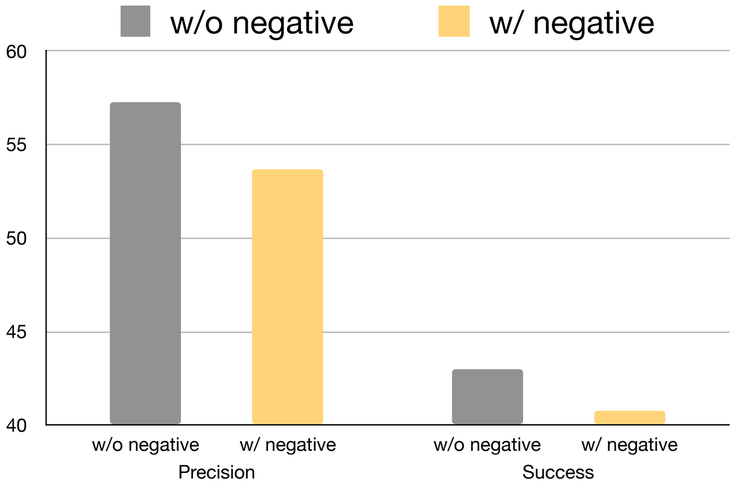
Object-level Correspondence on OTB-100

Findings and insights

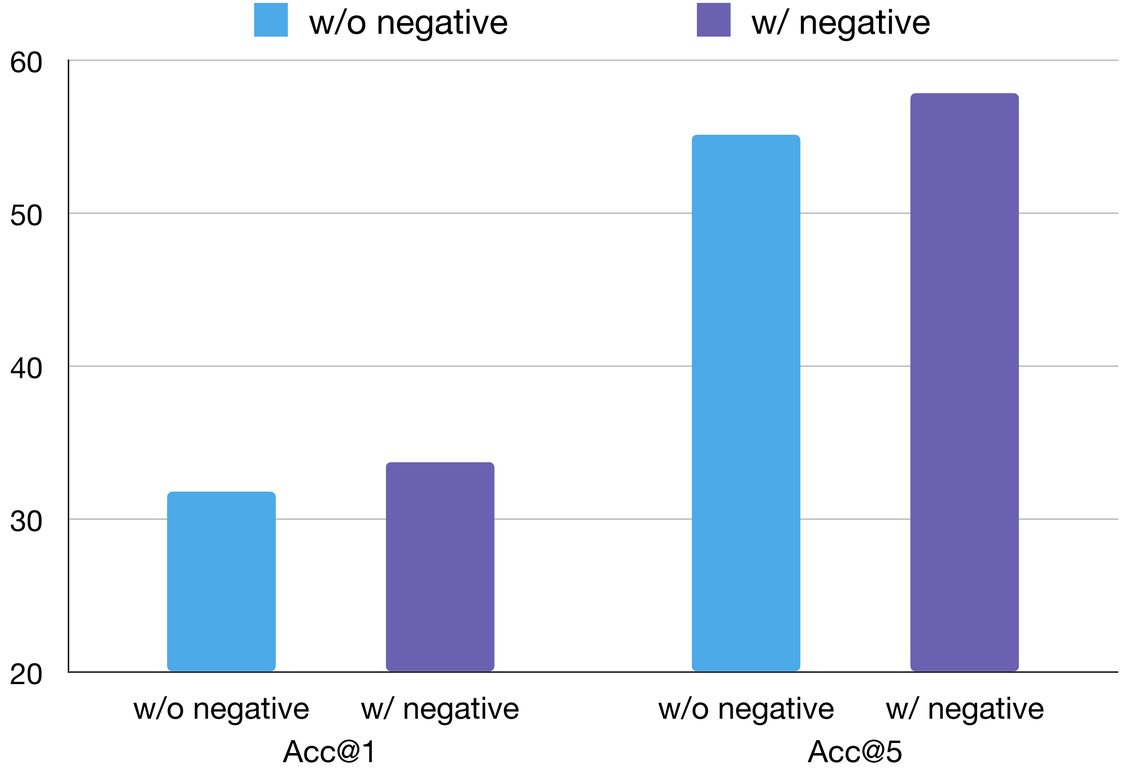
Negative pairs



Fine-grained Correspondence on DAVIS 2017



Object-level Correspondence on OTB-100



ImageNet Linear Classification

Summary

jerryxu.net/VFS

github.com/xvjiarui/VFS

Summary

- The simple VFS achieves state-of-the-art performance for self-supervised correspondence learning.

jerryxu.net/VFS

github.com/xvjiarui/VFS

Summary

- The simple VFS achieves state-of-the-art performance for self-supervised correspondence learning.
- Tracking based pretext task may not be necessary for self-supervised correspondence learning.

jerryxu.net/VFS

github.com/xvjiarui/VFS

Summary

- The simple VFS achieves state-of-the-art performance for self-supervised correspondence learning.
- Tracking based pretext task may not be necessary for self-supervised correspondence learning.
- Color augmentation is beneficial in object-level but jeopardize the fine-grained correspondence.

jerryxu.net/VFS

github.com/xvjiarui/VFS

Summary

- The simple VFS achieves state-of-the-art performance for self-supervised correspondence learning.
- Tracking based pretext task may not be necessary for self-supervised correspondence learning.
- Color augmentation is beneficial in object-level but jeopardize the fine-grained correspondence.
- Learning without negative improves correspondence learning.

jerryxu.net/VFS

github.com/xvjiarui/VFS