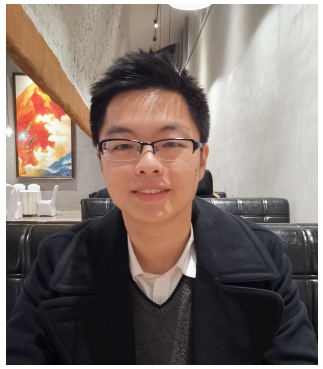
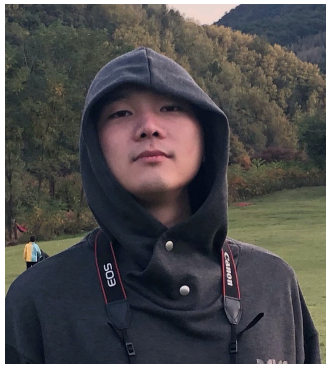
Breaking Shortcut: Exploring Fully Convolutional Cycle-Consistency for Video Correspondence Learning







Yansong Tang[1,4,*] Zhenyu Jiang[2,4,*] Zhenda Xie[3,4,*] Yue Cao[4] Zheng Zhang[4] Philip H. S. Torr[1] Han Hu[4]

[1]University of Oxford, [2]University of Texas, [3]Tsinghua University, [4]Microsoft Research Asia

* indicates equal contribution

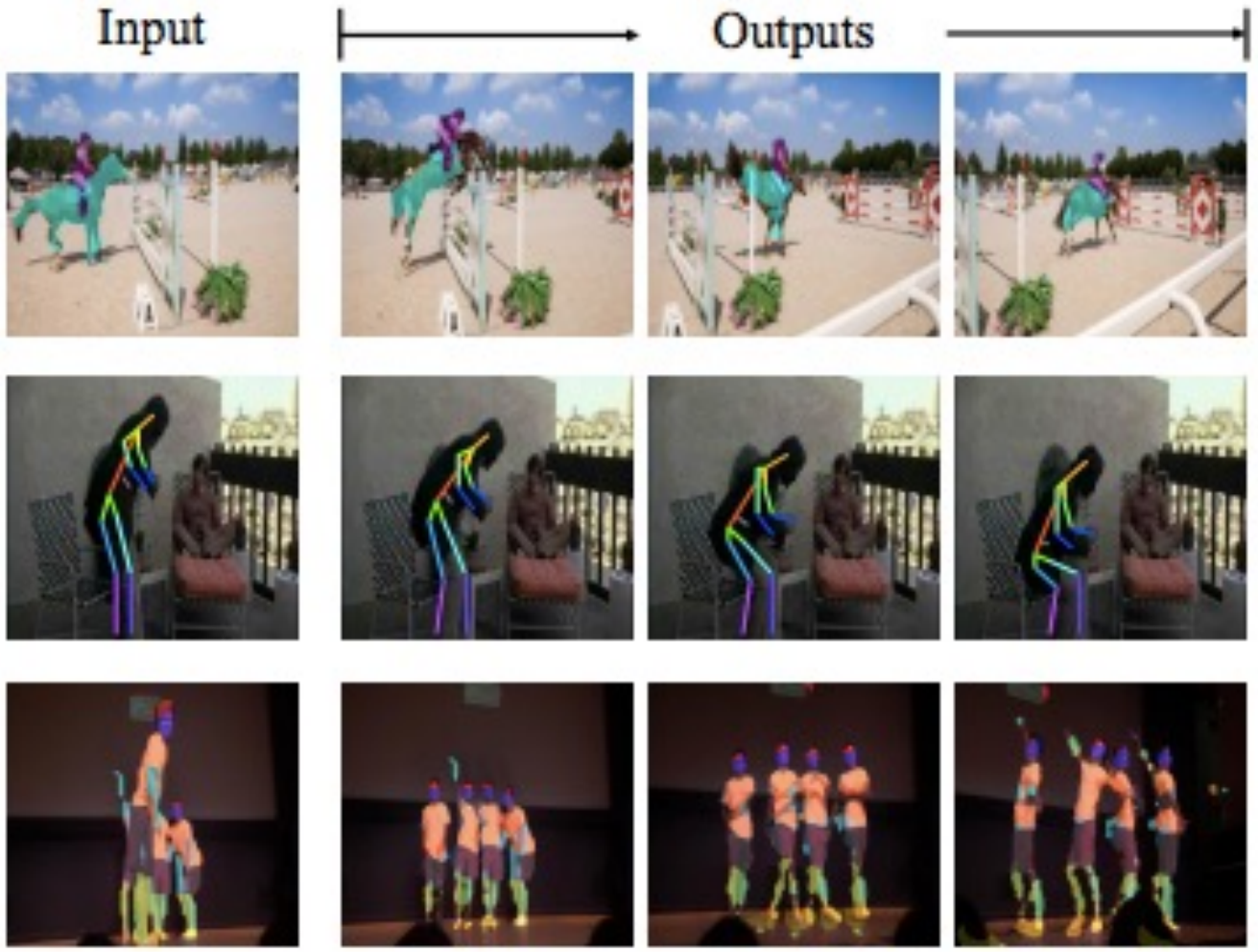
Code: https://github.com/Steve-Tod/STFC3

Video Correspondence Learning

Video Object Segmentation

Pose Tracking

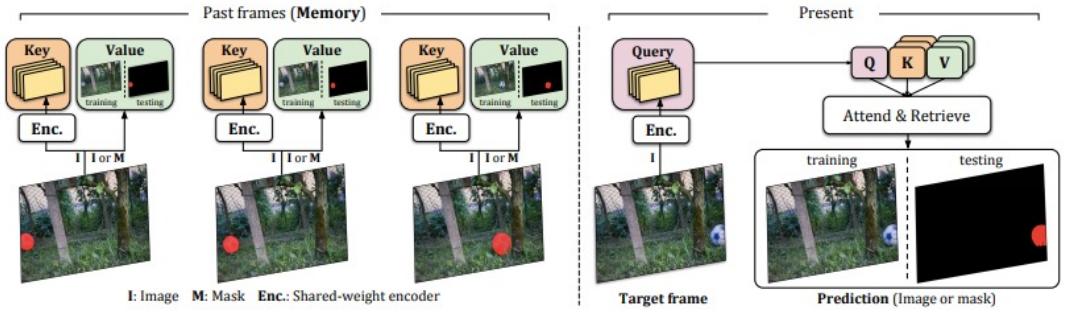
Human Body Part Tracking



Wang et al. CVPR 2019

Self-supervised Learning: A New Trend

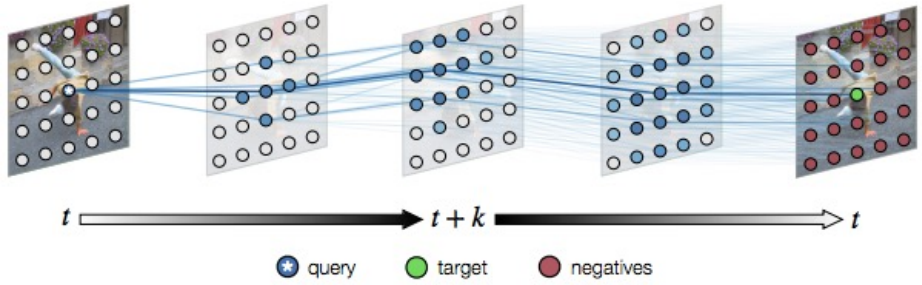
Self-supervision: Color reconstruction



MAST
Lai et al. CVPR 2020

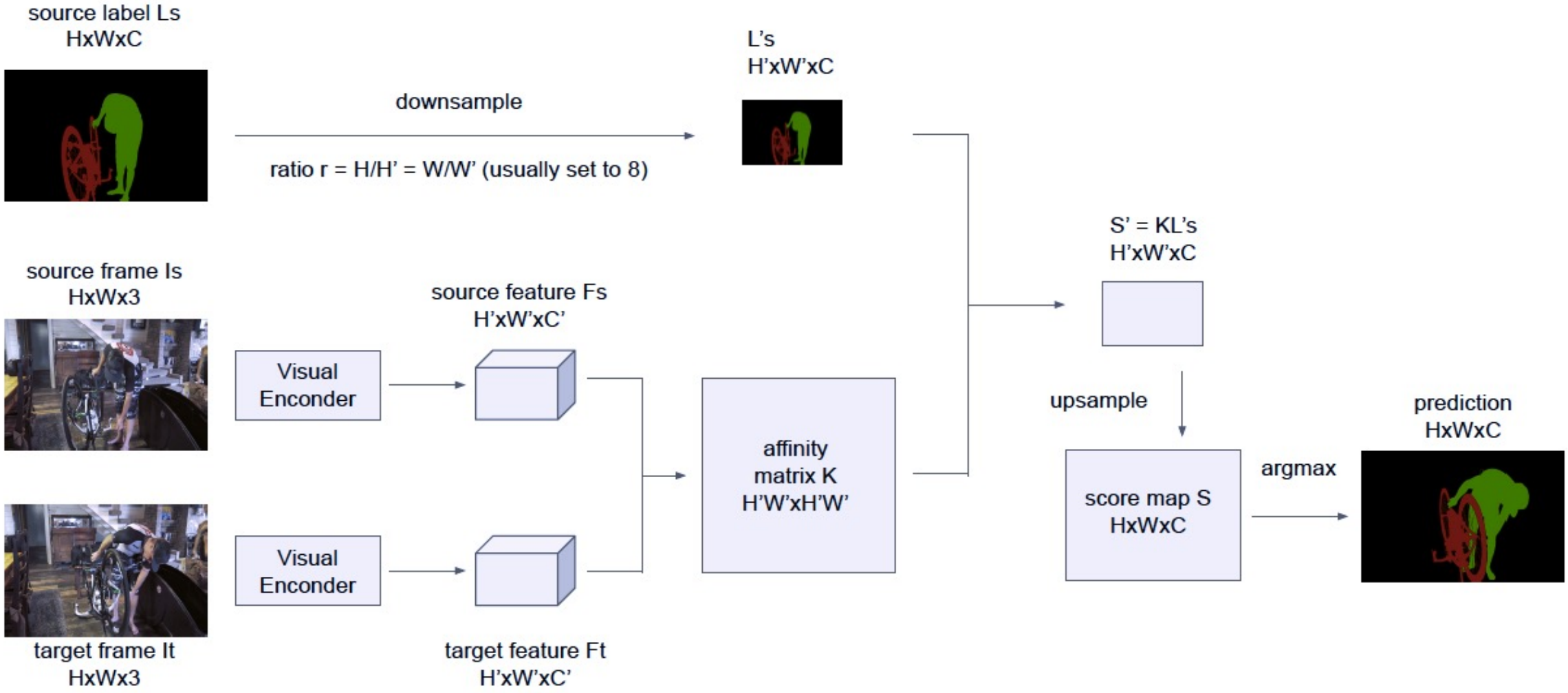
Self-Supervision: Cycle-consistency in time

Time-cycle
Wang et al. CVPR 2019

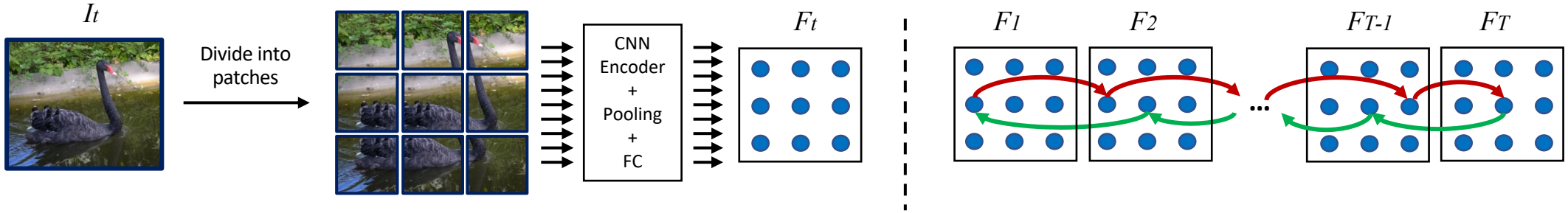


Contrastive random walk (CRW)
Allan et al. NeurIPS 2020

Testing pipeline of self-supervised video object segmentation



Motivation



CRW [1] Uses **patch-based** method (left) to perform cycle-consistency learning (right)

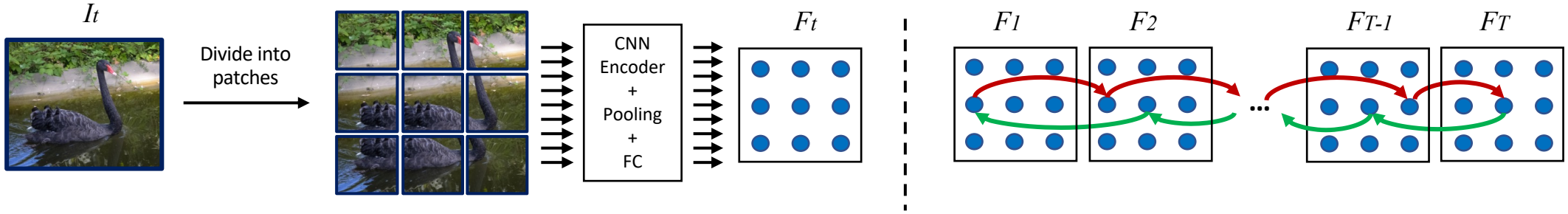
Can we apply a fully convolutional network (FCN) method rather than the patch-based one during the training process?

Bridge the inconsistency between training and testing

FCN has a larger receptive field

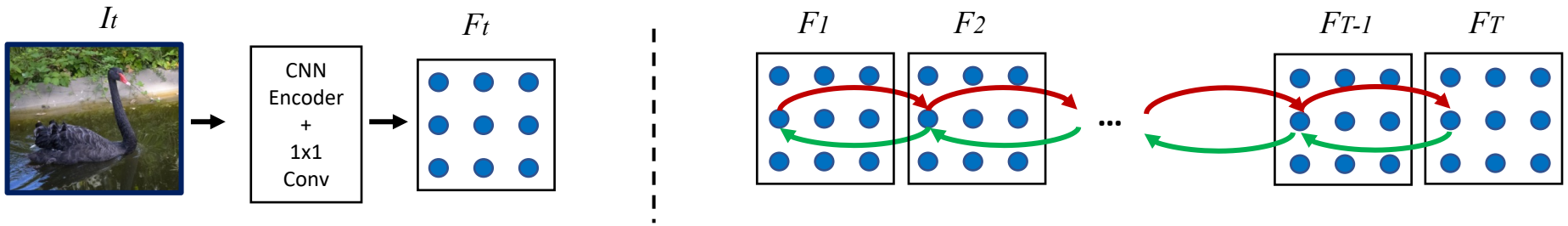
[1] CRW, Allan et al. NeurIPS 2020

Motivation



CRW [1] Uses patch-based method (left) to perform cycle-consistency learning (right)

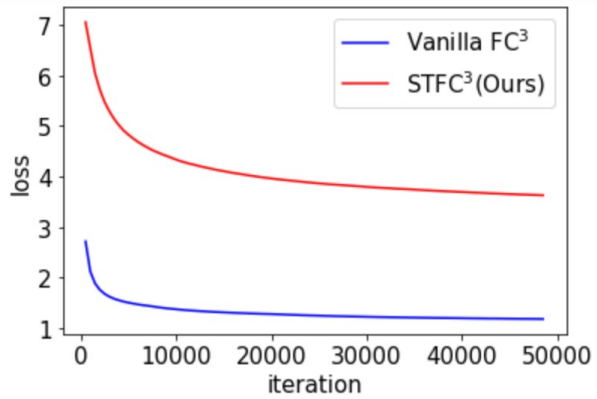
Can we apply a fully convolutional network (FCN) method rather than the patch-based one during the training process?



Vanilla fully convolutional cycle-consistency (FC3) learning method

[1] CRW, Allan et al. NeurIPS 2020

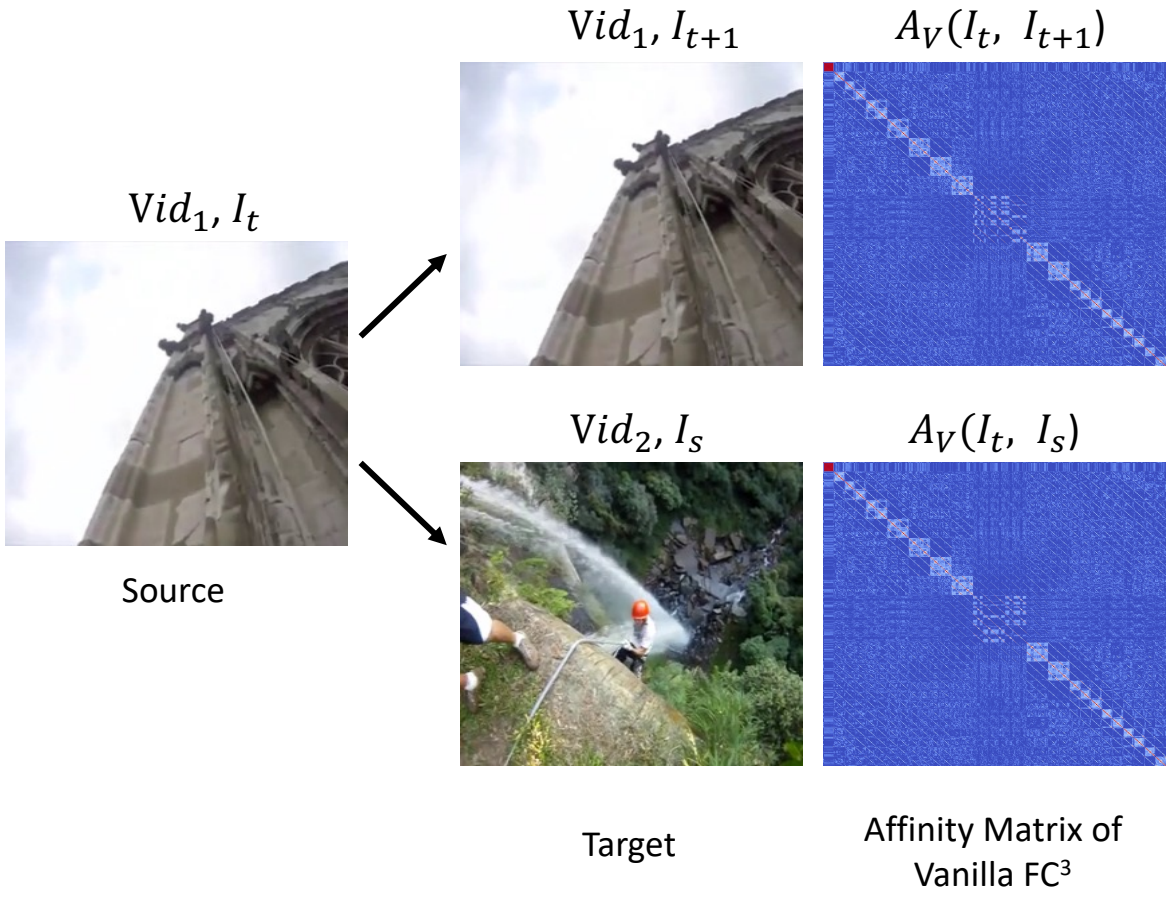
Vanilla FC^3 and its Shortcut



Training loss curve

Input Image Vanilla FC^3

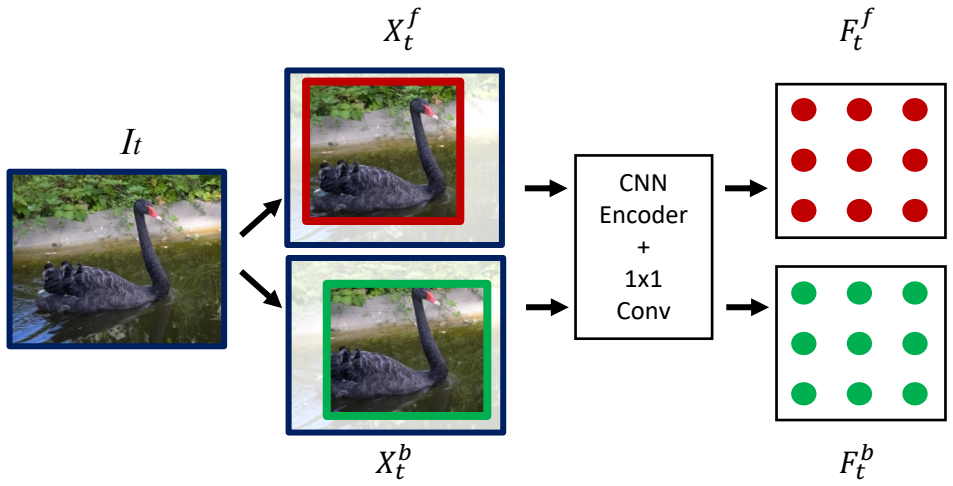
Visualization of feature map

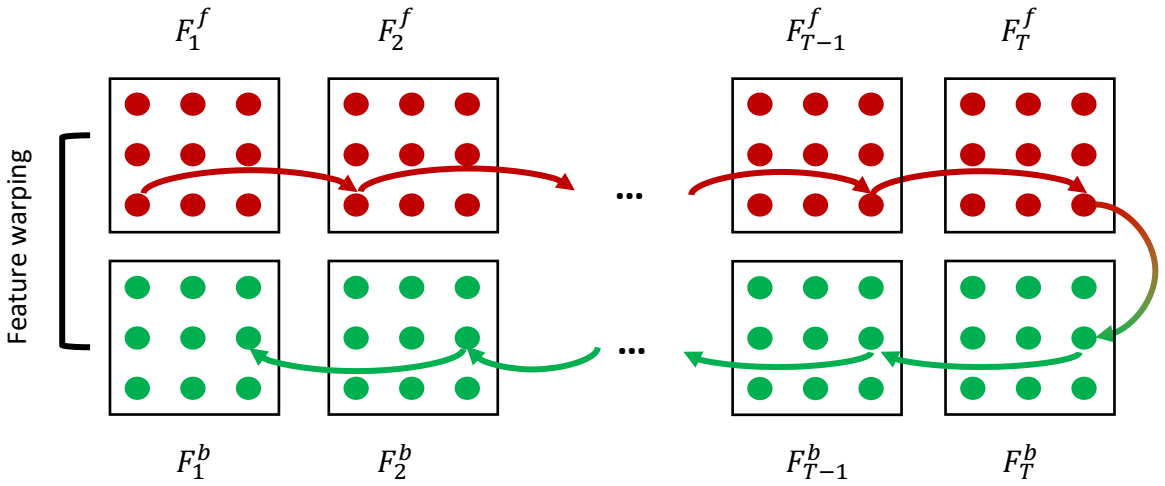


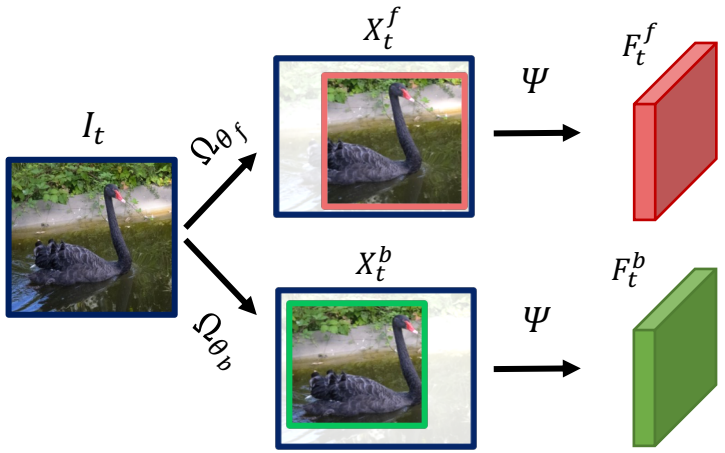
Source Target Affinity Matrix of Vanilla FC^3

Visualization of affinity matrices

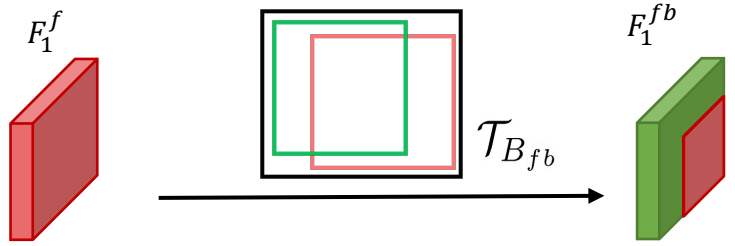
Spatial transformation fully convolutional cycle-consistency learning (STFC3)







(a) Spatial transformation and feature extraction



(b) Feature warping [2]

[2] Spatial transformer networks, Jaderberg et al. NeurIPS 2015

Experiment

Pretrained on the unlabeled Kinetics dataset

Downstream Task	Dataset	Number of Videos	Evaluation Metrics
Pose Tracking	J-HMDB	268	PCK@α
Face Landmark Tracking	300VW	31	RMSE (the lower is the better)
Video Object Segmentation	DAVIS2017	30	J&F

Experiment

Table 1. Experiment results on different methods to avoid the shortcut solution. Our method achieves significant improvement over the vanilla FC^3 method on three label propagation tasks. Experiments are conducted on the J-HMDB [19], 300VW [36] and DAVIS-2017 [33] for pose tracking, face landmark tracking and video object segmentation respectively.

Task	Pose Tracking		Face Landmark Tracking	Video Object Segmentation		
Metric	PCK@0.1 ↑	PCK@0.2 ↑	RMSE ↓	$\mathcal{J}\&\mathcal{F}_m$ ↑	$\mathcal{J}_m$ ↑	$\mathcal{F}_m$ ↑
Vanilla FC^3 (Zero Padding)	32.4	48.4	56.7	18.0	15.7	20.2
FC^3 (Replicate Padding)	49.7	67.6	28.2	31.5	29.8	33.3
FC^3 (Reflect Padding)	45.9	63.7	26.9	28.8	26.3	32.0
FC^3 (No Padding)	35.1	52.7	50.2	38.8	35.6	41.9
STFC^3 (Ours)	**62.0**	**80.5**	**18.8**	**60.5**	**58.0**	**63.1**

Experiment

Table 2. Evaluation of the pose tracking task with the J-HMDB benchmarks. SM, D and I represent using Sintel Movie [25], DAVIS2017 and ImageNet as the training data respectively.

Method	Supervised?	Training Data	Backbone	PCK@0.1	PCK@0.2
Thin-Slicing Network [37]	√	J-HMDB+I	Self-Designed	68.7	92.1
PAAP [15]	√	J-HMDB+I	VGG-16	51.6	73.8
ResNet-18 [12]	√	ImageNet	ResNet18	59.0	80.6
MoCo [11]	×	ImageNet	ResNet18	58.1	75.6
VINCE [7]	×	Kinetics	ResNet18	58.4	75.7
Identity	×	-	-	43.1	64.5
ColorPointer [40]	×	Kinetics	ResNet18	45.2	69.6
TimeCycle [42]	×	VLOG	ResNet18	57.3	78.1
mgPFF [21]	×	SM+D+J-HMDB	ResNet18	58.4	78.1
UVC [26]	×	Kinetics	ResNet18	58.6	79.6
CRW [17]	×	Kinetics	ResNet18	58.8	80.2
VFS [44]	×	Kinetics	ResNet18	60.5	79.5
STFC^3 (Ours)	×	Kinetics	ResNet18	**62.0**	**80.5**

Experiment

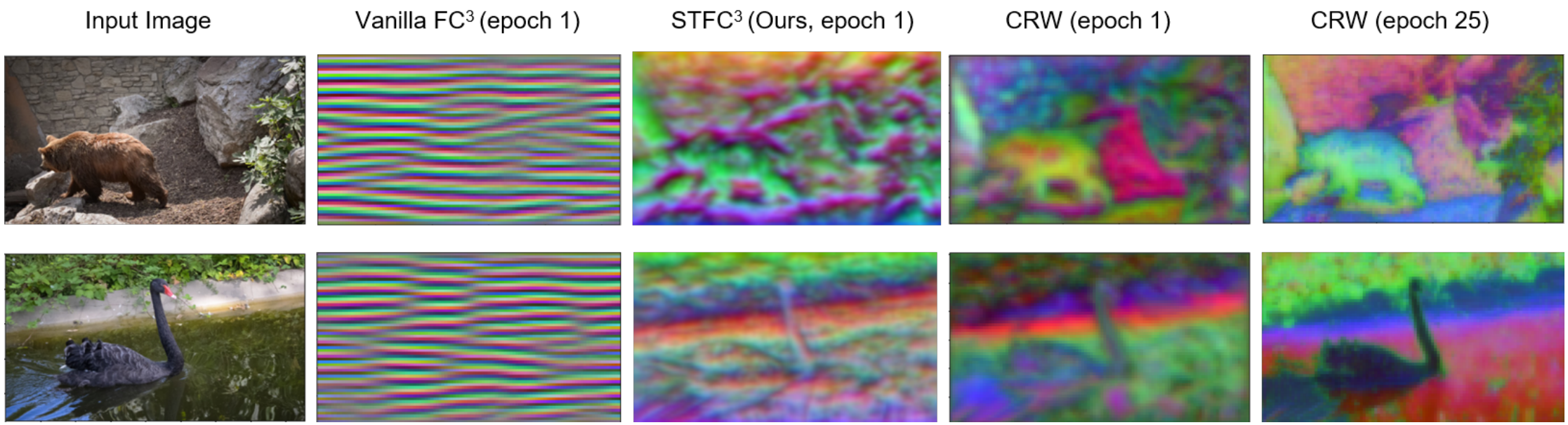
Table 3. Face landmark tracking results on the 300VW dataset [36], where the lower ↓ is better.

Method	RMSE ↓	Supervision
STRRN [47]	5.31	300W [35]
SBR [6]	5.77	300W [35] + ImageNet
ResNet-18	22.8	ImageNet
MoCo [11]	23.3	self-supervision
VINCE [7]	23.4	self-supervision
CRW [17]	21.6	self-supervision
UVC [26]	19.9	self-supervision
STFC^3 (Ours)	**18.8**	self-supervision

Table 4. Evaluation on the DAVIS-2017 dataset for video object segmentation. I, C, D, K, P, M, Y represents ImageNet, COCO, DAVIS2017, Kinetics, PASCAL-VOC, Mapillary and YouTube-VOS. The methods with * are under fully-supervised learning setting. All the self-supervised methods are based on ResNet-18.

Method	Train Data	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
PReMVOS* [29]	I/C/D/P/M	77.8	73.9	81.8
STM* [32]	I/D/Y	81.8	79.2	84.3
CFBI* [46]	I/C/D	83.3	80.5	86.0
ResNet-18* [13]	I	62.9	60.6	65.2
MoCo [11]	I	60.8	58.6	63.1
VINCE [7]	K	60.4	57.9	62.8
Colorization [40]	K	34.0	34.6	32.7
TimeCycle [42]	VLOG	48.7	46.4	50.0
CorrFlow [23]	OxUvA	50.3	48.4	52.2
UVC+track [26]	K	59.5	57.7	61.3
MAST [22]	Y	65.5	63.3	67.6
VFS[44]	K	66.6	64.0	69.4
CRW[17]	K	67.6	64.8	70.2
STFC^3 (Ours)	K	60.5	58.0	63.1

Experiment



Visualization of the top three PCA components of the feature map learned by different self-supervisedly pretrained models.

CRW learns invariant **patch-level** region features (smooth), better for segmentation

STFC^3 (ours) performs **pixel-level** learning (distinctive), better for keypoint tracking

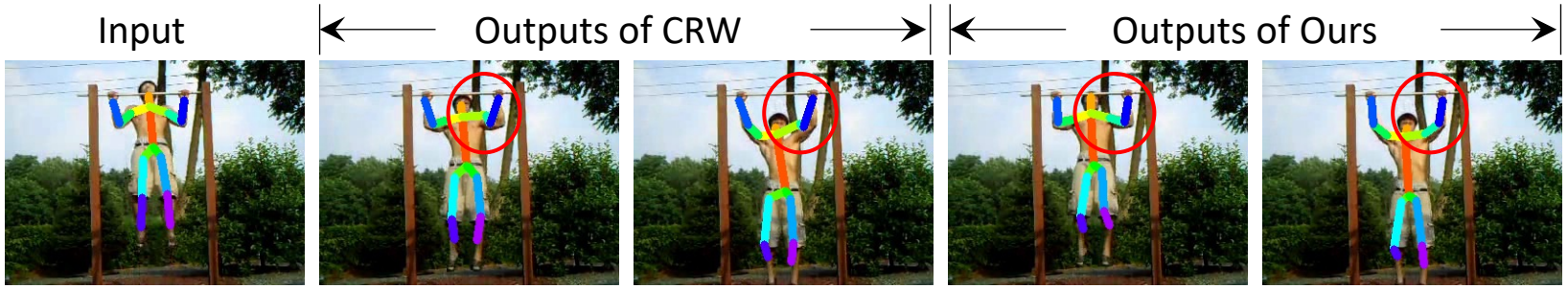
Experiment

Qualitative results of our method on J-HMDB for pose tracking

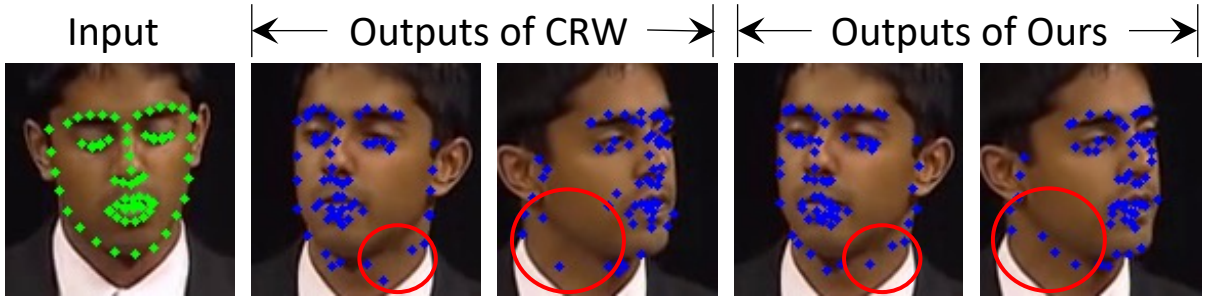
Experiment

Qualitative results of our method on 300VW for face landmark tracking and DAVIS-2017 for video object segmentation

Experiment



J-HMDB



300VW

Visualization comparison of our method with CRW

Conclusion

Explore various ***fully convolutional cycle-consistency*** methods for self-supervised video correspondence learning

Analyze the shortcut issue caused by ***position encoding***, and propose a ***spatial transformation*** approach to address it

Achieve state-of-the-art results on pose tracking and face landmark tracking

Future work

Contrastive learning: pixel level (ours) + frame level [1,2]

[1] DINO, Caron et al. ICCV 2021 [2] VFS, Xu et al. ICCV 2021