Current Works in Computer Vision

Learning Correspondence from the Cycle-consistency of Time

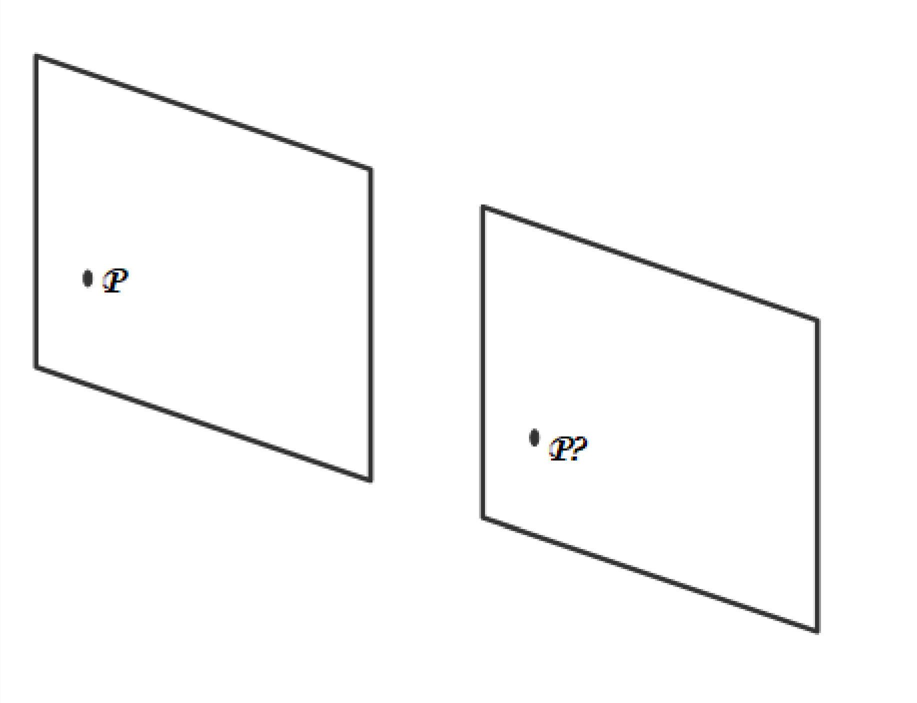
Xiaolong Wang   Allan Jabri   Alexei A. Efros

Outline

- Introduction
- Related Works
- Approach
- Experiments
- Limitations and Conclusion
- References

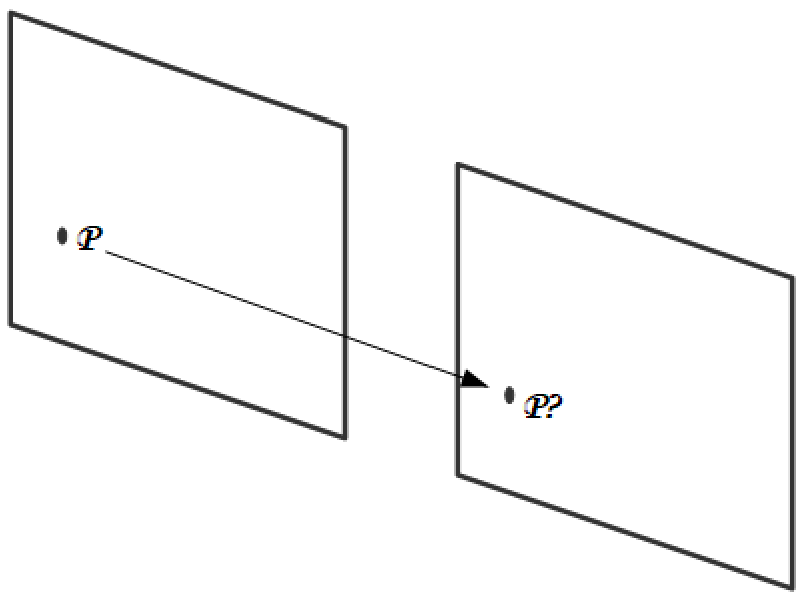
Introduction

- Given two images depicting roughly the same scene.



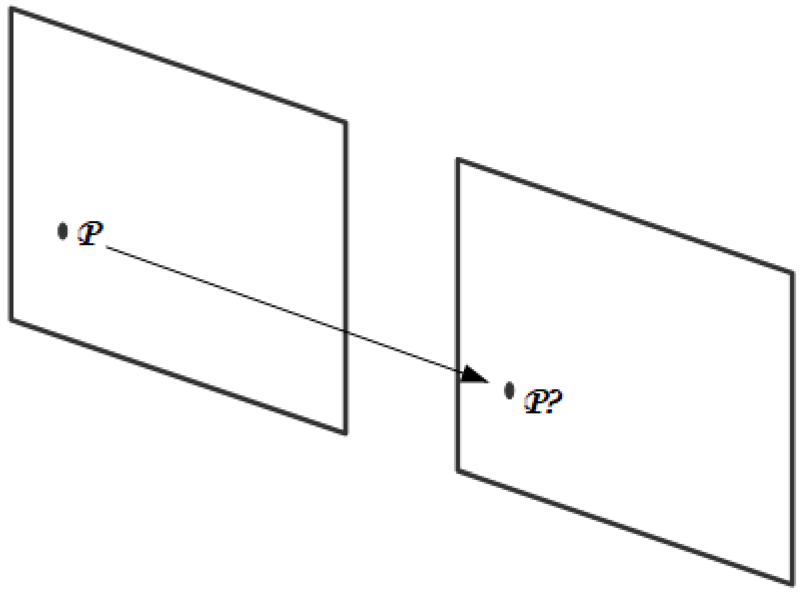
Introduction

- Given two images depicting roughly the same scene.
- Finding the corresponding pixel of p in the second image



Introduction

- Given two images depicting roughly the same scene.
- Finding the corresponding pixel of p in the second image.
- One of the most important problems in computer vision.



Introduction

- **Low-level learning approach**

  difficult to be generalized.
- **High-level learning approach**

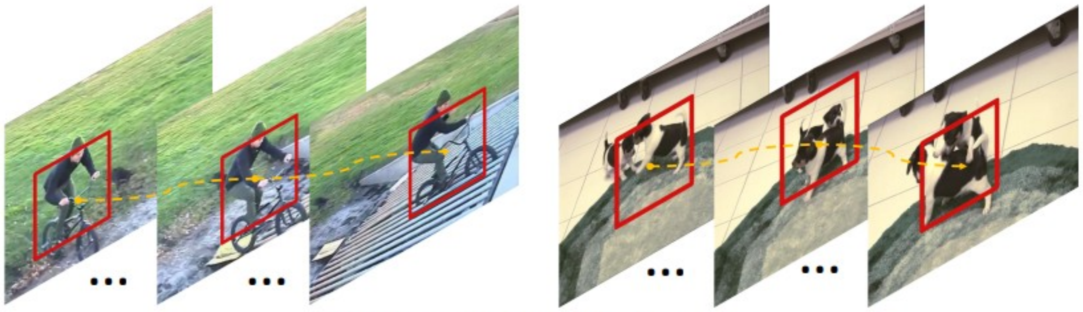
  expensive at large scale.
- **Unsupervised learning of correspondence**

Related Work

- Tracking
- Optical flow
- Self-supervised representation learning from video
- Temporal Continuity in Visual Learning
- Forward/backward cycle Consistency

Related Work

- **Tracking**
  - Two types of tracking.
  - Repeated recognition tracking.
  - Tracking by matching.
- **Optical flow**
  - Correspondence at the pixel level.
  - difficulties with long-range correspondence.
  - Mid-level optical flow.



(a) Unsupervised Tracking in Videos

- **Self-supervised representation learning from video**
  - Off-the-shelf tools.
  - Limitation by those tools.
  - Joint learning as a solution.

Related Work

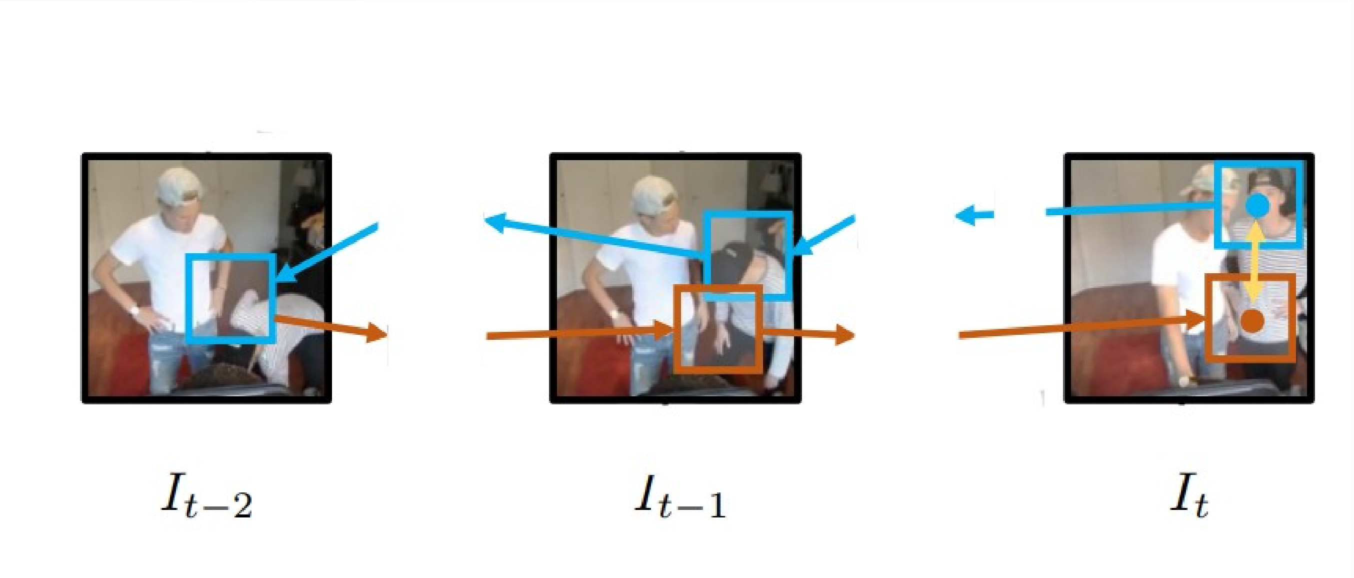
- **Forward/Backward and Cycle Consistency**
  - How to achieve cycle-consistency when facing some challenges.
  - The usage of cycle-consistency over multiple steps.

- **Temporal Continuity in Visual Learning**
  - Importance of Temporal stability.
  - Computational methods.
  - Slow feature learning with fixation without supervision.

Approach

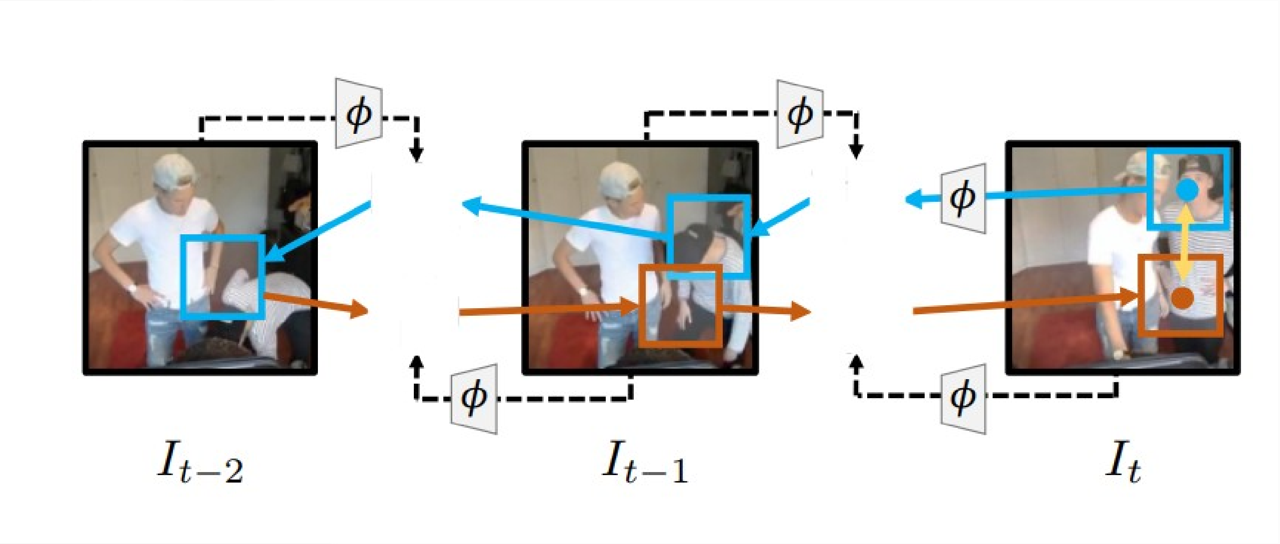
- **Cycle Consistency Losses**
  - Recurrent Tracking Formulation
  - Learning Objectives
- **Architecture for Mid-level correspondence**
  - Spatial Feature Encoder
  - Differentiable Tracker
  - End-to-end Joint Training

Approach



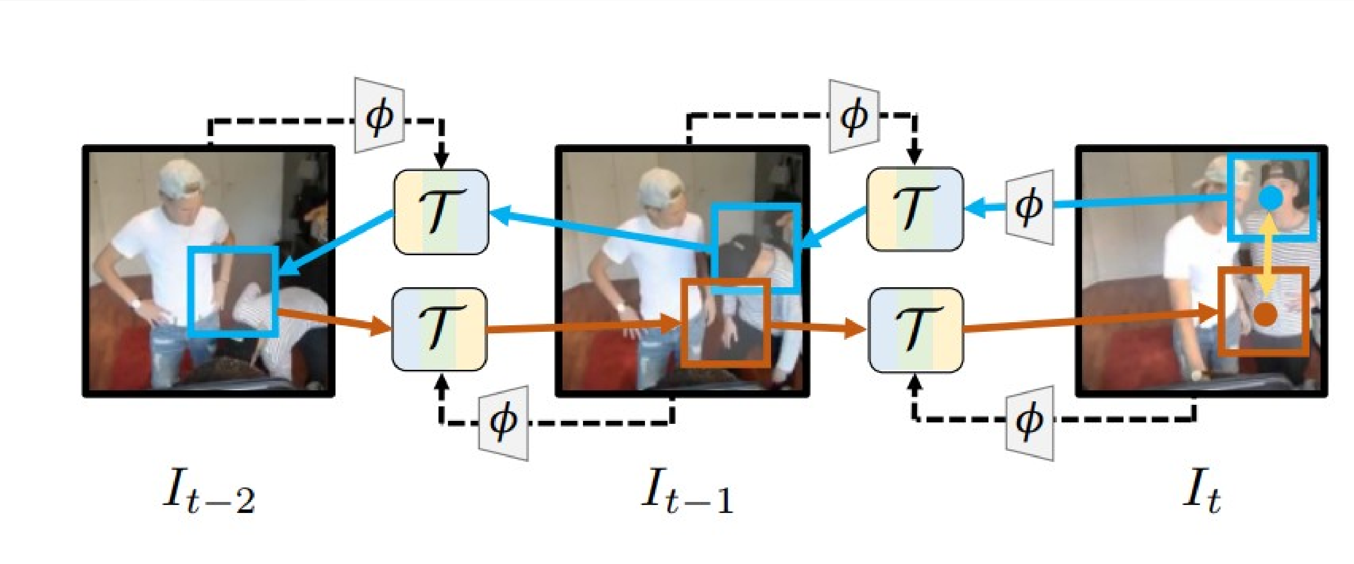
- extract a patch p_t from image I_t

Approach



- Extract a patch p_t from image I_t
- Learn feature space ϕ by tracking the patch backwards and then forwards in time.
- Minimization of the cycle consistency loss.

Approach



- Extract a patch p_t from image I_t
- Learn feature space ϕ by tracking the patch backwards and then forwards in time.
- Minimization of the cycle consistency loss.
- $\mathcal{T}$ returns best match feature region in the target image.

Cycle-Consistency Losses
Recurrent Tracking Formulation

Track operation applied iteratively in a forward manner:

$$\mathcal{T}^{(i)}(x^I_{t-i}, x^p) = \mathcal{T}(x^I_{t-1}, \mathcal{T}(x^I_{t-2}, ...\mathcal{T}(x^I_{t-i}, x^p)))$$

Cycle-Consistency Losses
Recurrent Tracking Formulation

Track operation applied iteratively in a forward manner:

$$\underbrace{\mathcal{T}^{(i)}}_{i\text{ times forwards}}(x^I_{t-i}, x^p) = \mathcal{T}(\underbrace{x^I_{t-1}}_{\text{Finishing at}}, \mathcal{T}(x^I_{t-2}, \ldots \mathcal{T}(\underbrace{x^I_{t-i}}_{\text{starting from}}, x^p)))$$

Cycle-Consistency Losses
Recurrent Tracking Formulation

Track operation applied iteratively in a forward manner:

$$\mathcal{T}^{(i)}(\boxed{x^I_{t-i}}, \boxed{x^p}) = \mathcal{T}(x^I_{t-1}, \mathcal{T}(x^I_{t-2}, \ldots \mathcal{T}(x^I_{t-i}, x^p)))$$

$$x^I_{t-k:t} = \phi(I_{t-k:t})$$

$$x^p_t = \phi(p_t)$$

Pixels mapped to a feature space by an encoder ϕ.

Cycle-Consistency Losses
Recurrent Tracking Formulation

Track operation is applied iteratively in a forward manner:

$$\mathcal{T}^{(i)}(x^I_{t-i}, x^p) = \mathcal{T}(x^I_{t-1}, \mathcal{T}(x^I_{t-2}, ...\mathcal{T}(x^I_{t-i}, x^p)))$$

Summary:

- Input: features of both image & a patch.
- Output: the most similar patch feature to the input patch.
- Can be applied iteratively in both forward and backward manner.

Track operation is applied iteratively in a backward manner:

$$\mathcal{T}^{(-i)}(x^I_{t-1}, x^p) = \mathcal{T}(x^I_{t-i}, \mathcal{T}(x^I_{t-i+1}, ...\mathcal{T}(x^I_{t-1}, x^p)))$$

Learning Objectives

The following is sums of the overall learning objectives over k cycles:

$$\mathcal{L} = \sum_{i=1}^{k} \mathcal{L}_{sim}^{i} + \boxed{\lambda}\mathcal{L}_{skip}^{i} + \boxed{\lambda}\mathcal{L}_{long}^{i}.$$

Weight parameter set to equal:= 0.1

Learning Objectives
Feature Similarity

The following is sums of the overall learning objectives over k cycles:

$$\mathcal{L} = \sum_{i=1}^{k} \boxed{\mathcal{L}_{sim}^{i}} + \lambda \mathcal{L}_{skip}^{i} + \lambda \mathcal{L}_{long}^{i}.$$

$$\mathcal{L}_{sim}^{i} = -\langle x_t^p, \mathcal{T}(x_{t-i}^I, x_t^p) \rangle$$

- Similarity measure between feature query patch & the localized patch.
- Negative Frobenius inner product.

Learning Objectives
Skip Cycle

The following is sums of the overall learning objectives over k cycles:

$$\mathcal{L} = \sum_{i=1}^{k} \mathcal{L}_{sim}^{i} + \lambda \boxed{\mathcal{L}_{skip}^{i}} + \lambda \mathcal{L}_{long}^{i}.$$

Error in alignment measure

$$\mathcal{L}_{skip}^{i} = \boxed{l_\theta}(x_t^p, \mathcal{T}(x_t^I, \mathcal{T}(x_{t-i}^I, x_t^p))).$$

- Long-range matching is allowed.
- It is achieved by skipping frames.
- Above an example of skipping to the i^{th} frame away.

Learning Objectives
Skip Cycle

The following is sums of the overall learning objectives over k cycles:

$$\mathcal{L} = \sum_{i=1}^{k} \mathcal{L}^i_{sim} + \lambda \boxed{\mathcal{L}^i_{skip}} + \lambda \mathcal{L}^i_{long}.$$

Error in alignment measure

$$\mathcal{L}^i_{skip} = \boxed{l_\theta}(x^p_t, \mathcal{T}(x^I_t, \mathcal{T}(x^I_{t-i}, x^p_t))).$$

Learning Objectives
Tracking

The following is sums of the overall learning objectives over k cycles:

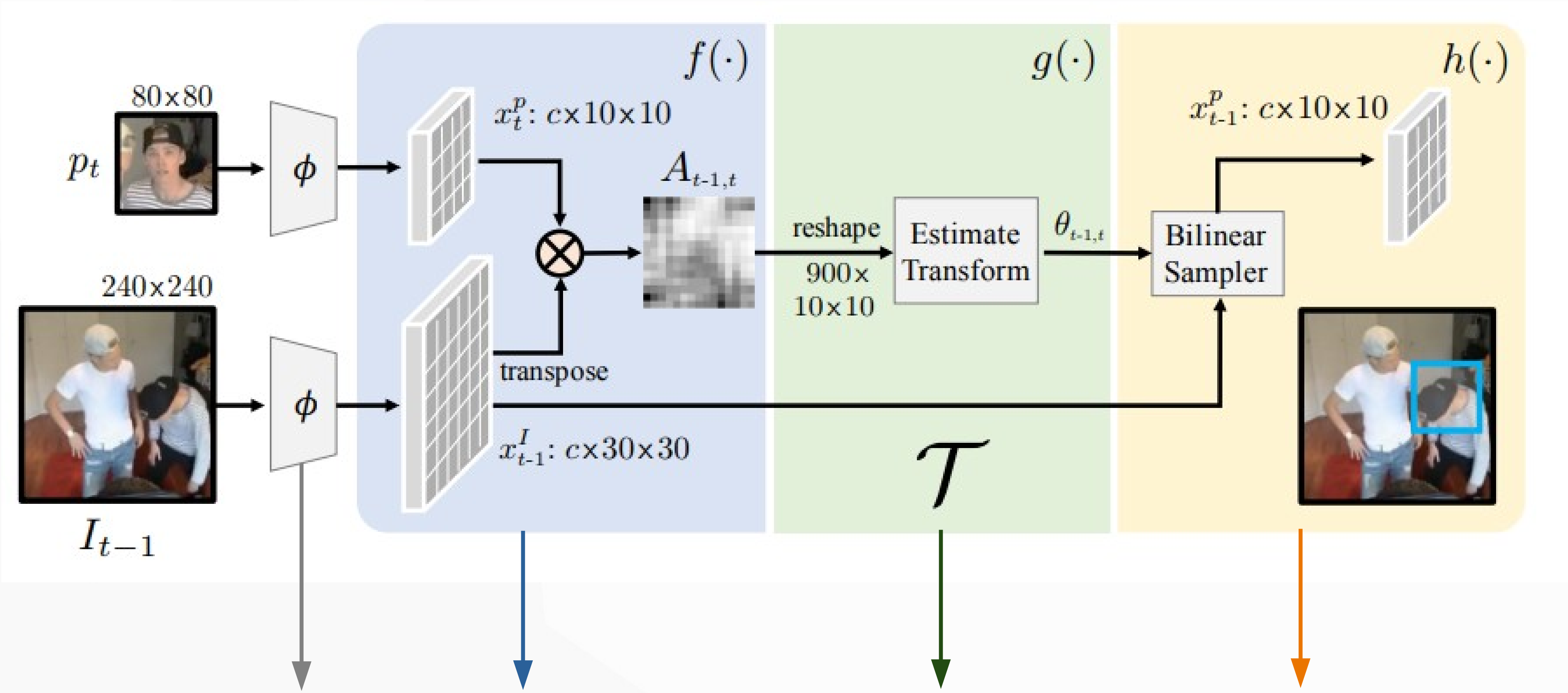
$$\mathcal{L} = \sum_{i=1}^{k} \mathcal{L}_{sim}^{i} + \lambda \mathcal{L}_{skip}^{i} + \lambda \boxed{\mathcal{L}_{long}^{i}}.$$

Error in alignment measure

$$\mathcal{L}_{long}^{i} = \boxed{l_\theta}(x_t^p, \mathcal{T}^{(i)}(x_{t-i+1}^I, \mathcal{T}^{(-i)}(x_{t-1}^I, x_t^p))).$$

- The tracker chase the features.
- At first it goes i^{th} steps backwards.
- Then goes forwards till it reaches the initial query.

Architecture for Mid-level Correspondence



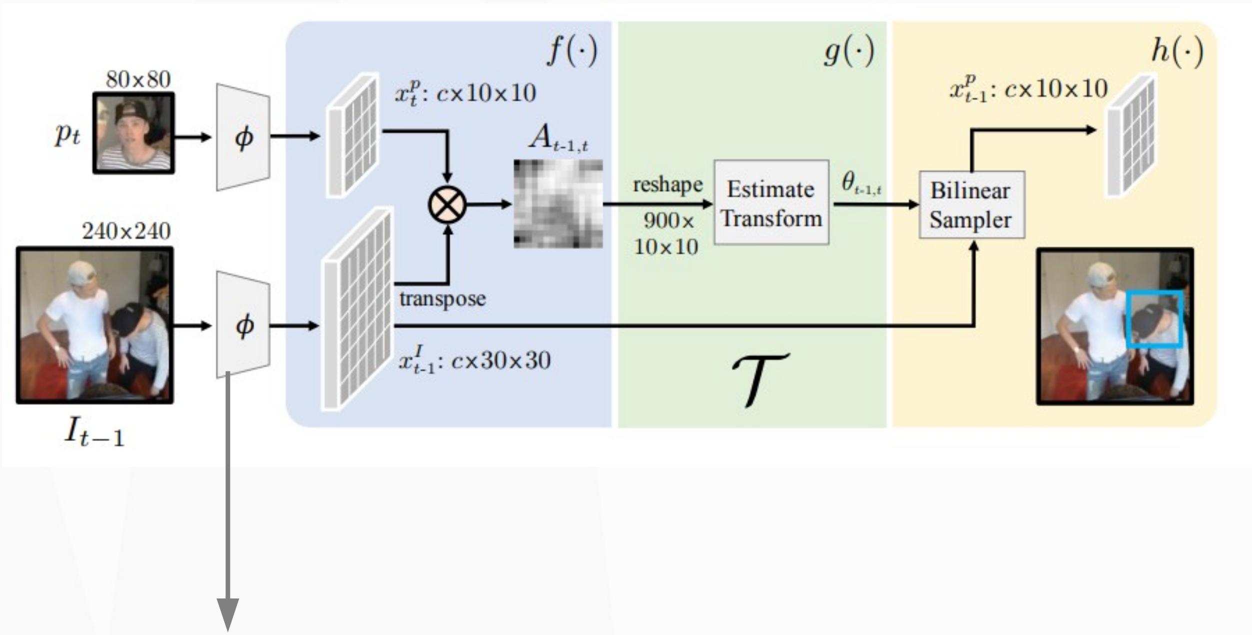
Spatial Feature Encoder

Affinity function

Localizer

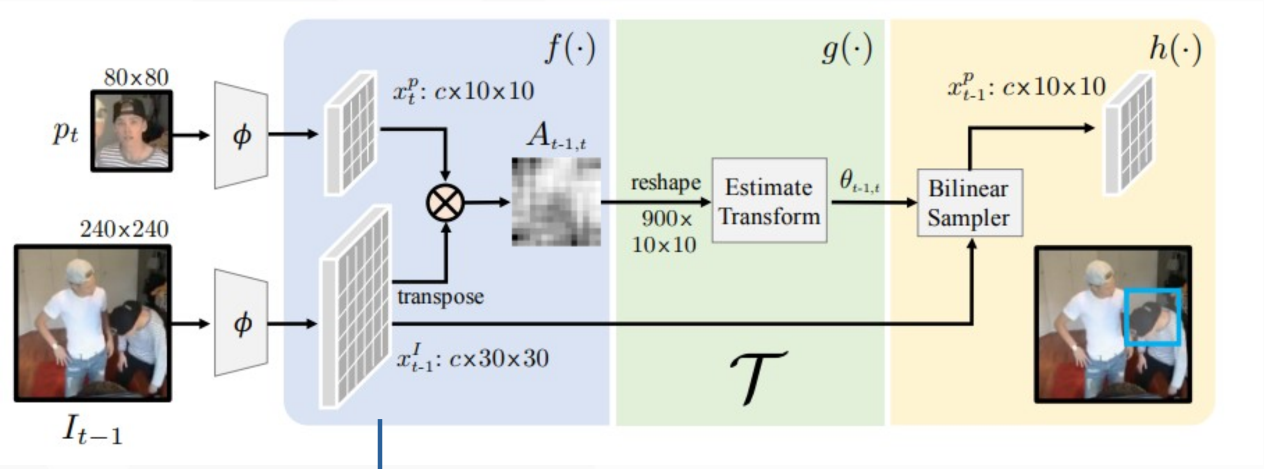
Bilinear Sampler

Architecture for Mid-level Correspondence spatial Feature Encoder



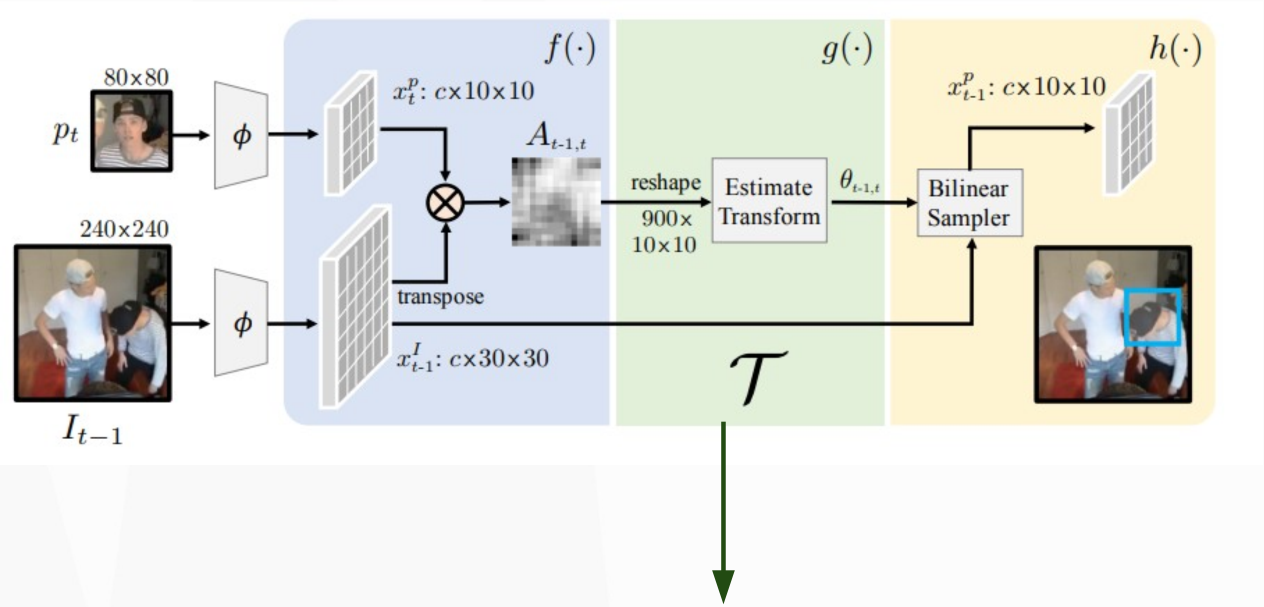
- **Spatial Feature Encoder**
  - determines the type of correspondence (mid-level in this case).
  - Maps pixels into feature space.
  - In this research ResNet-50 without res_5 was used.

Differentiable Tracker
Affinity function



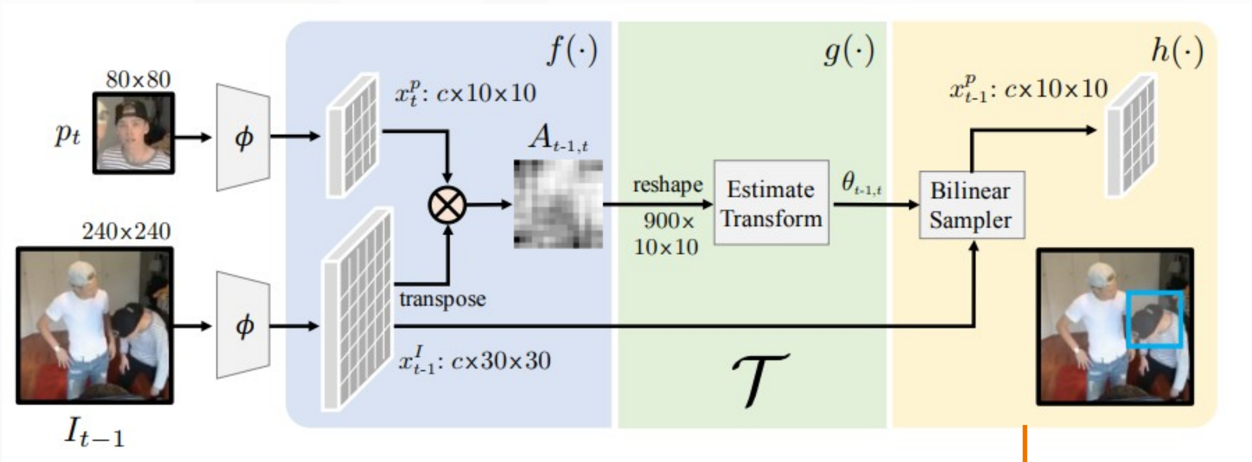
- **Affinity function**
  - A similarity measure between two coordinates of spatial features.
  - Dot product between embeddings.
  - $f : \mathbb{R}^{c\times30\times30} \times \mathbb{R}^{c\times10\times10} \rightarrow \mathbb{R}^{900\times100}$.

Differentiable Tracker Localizer



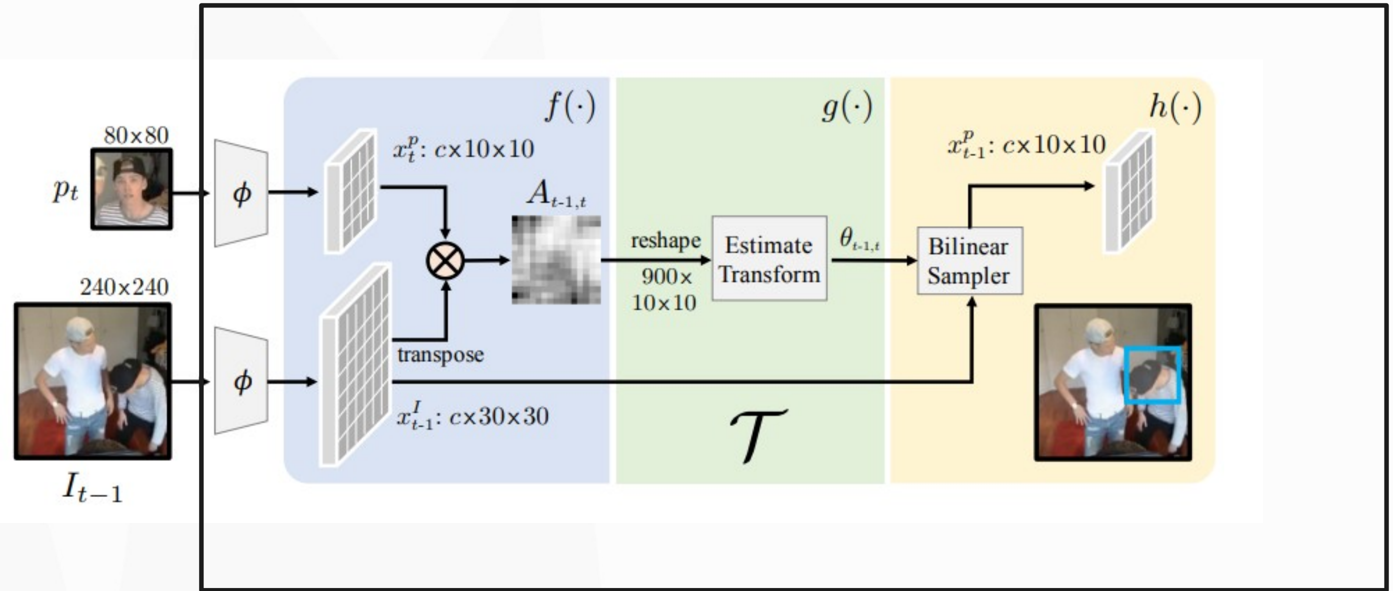
- **Localizer**
  - Uses the resulting Affinity matrix A to find the best match for the corresponding patch.
  - Outputs localization parameters θ .
  - Consist of 2 convolutional layers & 1 linear layer.

Differentiable Tracker
Bilinear sampler



- **Bilinear sampler**
  - Uses both the image feature and the localization parameters.
  - Produce a new feature patch.
  - $h : \mathbb{R}^{c\times 30\times 30} \times \mathbb{R}^3 \rightarrow \mathbb{R}^{c\times 10\times 10}.$

End-to-end Training



- The spatial encoder and the tracking operation together forms a differentiable patch tracker.
- This allows end-to-end training.
- $$x^I, x^p = \phi(I), \phi(p)$$
  $$\mathcal{T}(x^I, x^p) = h(x^I, g(f(x^I, x^p))).$$

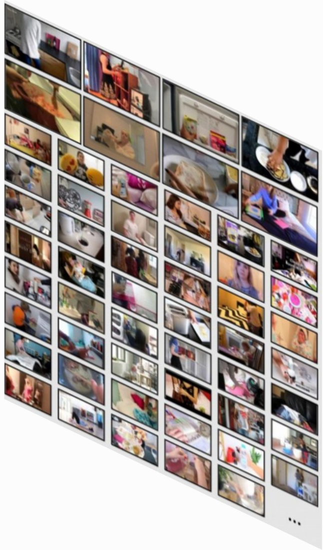
Experiments

- **Setup and baselines**

  \- Training - Inference - Propagation results -Baselines
- **Davis-2017 (experiment I & IV)**
- **JHMDB (II)**
- **VIP (III)**

Experiments
Training

- Model trained with VLOG dataset.
- No annotation or pre-training was used.
- 114K videos with overall length of 344 hrs.
- They set the past frames to be 4
- They trained the model for 30 epochs.



Experiments
Inference

- The trained encoder was used to compute dense correspondences.

- Labels are given for the first frame of video.

- The initial labels were propagated to the rest of frames.

- Labels of pixels are quantified by C classes.
(e.g. for segmentation masks C is the number of semantic labels).

- Labels are propagated in the feature space.

Experiments
Inference

- The Experiments tasks depend on the labels type.

- For instance masks labels Davis-2017 was used.
  - Multiple instances are annotated at each video sequence.
- For human pose keypoints JHMDB was used.
  - Fully annotated for human poses & actions.
- For both instance masks and semantic-level VIP was used.
  - Semantic labels are used for different human parts.
  - Instance labels are used to differentiate humans.

Experiments
Propagation Results

- The feature can propagate The initial labels to the rest of frames.
- In the examples above instance masks labels were propagated. (DAVIS-2017).

Experiments
Baseline

- **Unsupervised**
  - Identity
  - Optical flow
  - SIFT flow
  - Transitive Invariance
  - DeepCluster
  - Video Colorization
- **Supervised**
  - ImageNet pre-trained
  - Fully-Supervised Methodes

Experiments
Baseline

- **Unsupervised**
  - Identity
  - Optical flow
  - SIFT flow
  - Transitive Invariance
  - DeepCluster
  - Video Colorization
- **Supervised**
  - ImageNet pre-trained
  - Fully-Supervised Methodes

- Self supervised.
- Trained with Kinetics.
- Self-supervision via color propagation.

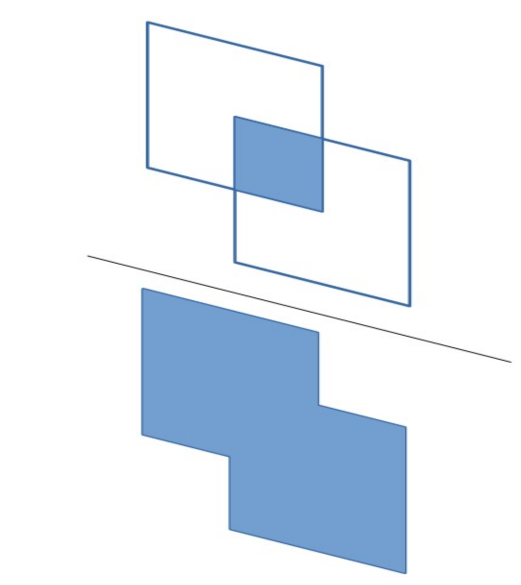
Experiments
Baseline

- **Unsupervised**
  - Identity
  - Optical flow
  - SIFT flow
  - Transitive Invariance
  - DeepCluster
  - Video Colorization
- **Supervised**
  - ImageNet pre-trained
  - Fully-Supervised Methodes

- Supervised method.
- ResNet-50.
- Trained on ImageNet.

Experiment I
Evaluation Metrics

- **Region similarity (IoU)** $\mathcal{J}$
  - Intersection over union
- **Contour-based accuracy** $\mathcal{F}$
  - How well both objects fits together.



Experiment I
Instance Propagation on DAVIS-2017

model	Supervised	$\mathcal{J}$(Mean)	$\mathcal{F}$(Mean)
Identity		22.1	23.6
Random Weights (ResNet-50)		12.4	12.5
Optical Flow (FlowNet2) [22]		26.7	25.2
SIFT Flow [39]		33.0	35.0
Transitive Inv. [74]		32.0	26.8
DeepCluster [8]		37.5	33.2
Video Colorization [69]		34.6	32.7
Ours (ResNet-18)		40.1	38.3
Ours (ResNet-50)		**41.9**	**39.4**
ImageNet (ResNet-50) [18]	✓	50.3	49.0
Fully Supervised [81, 7]	✓	55.1	62.1

4.4% in $\mathcal{J}$
6.2% in $\mathcal{F}$

2% worse

7.3% in $\mathcal{J}$
6.7% in $\mathcal{F}$

- ImageNet preforms better.
- ImageNet has advantage cause of the use of curated object-centric annotation.

Experiment II
Evaluation Metrics

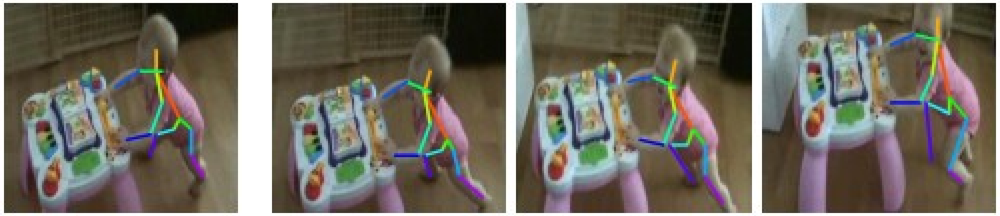
- **The standard metric PCK**
  - Measures the accuracy of the localization of the body joints.
  - In other words measures how many keypoints close to the ground truth in precentage.

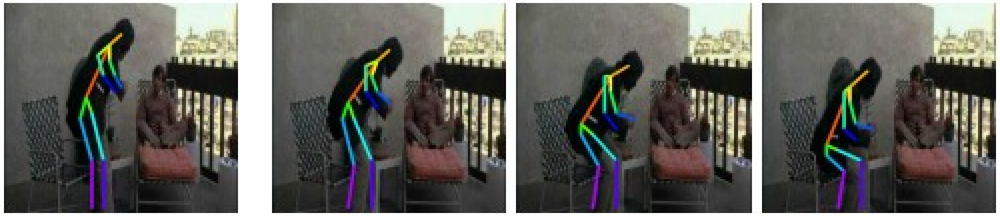
Experiment II
Pose Keypoint Propagation on JHMDB

model	Supervised	PCK@.1	PCK@.2
Identity		43.1	64.5
Optical Flow (FlowNet2) [22]		45.2	62.9
SIFT Flow [39]		49.0	68.6
Transitive Inv. [74]		43.9	67.0
DeepCluster [8]		43.2	66.9
Video Colorization [69]		45.2	69.6
Ours (ResNet-18)		57.3	78.1
Ours (ResNet-50)		**57.7**	**78.5**
ImageNet (ResNet-50) [18]	✓	58.4	78.4
Fully Supervised [59]	✓	68.7	92.1

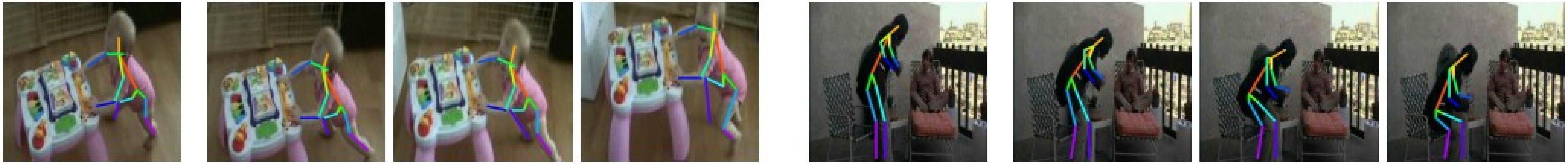
8.7% in PCK@.1
9.9% in PCK@.2

0.7% in PCK@.1



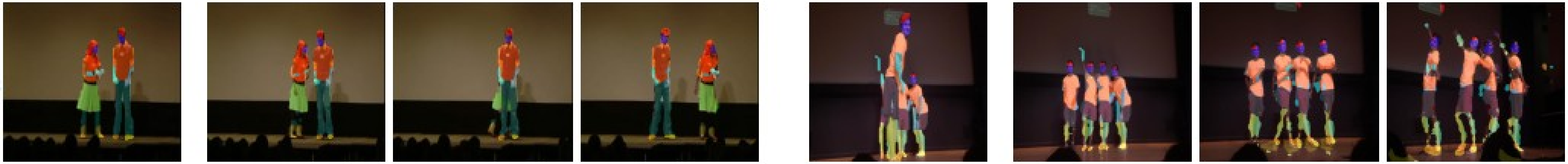


Experiment II
Pose Keypoint Propagation on JHMDB



Experiment III
Semantic and Instance Propagation on VIP

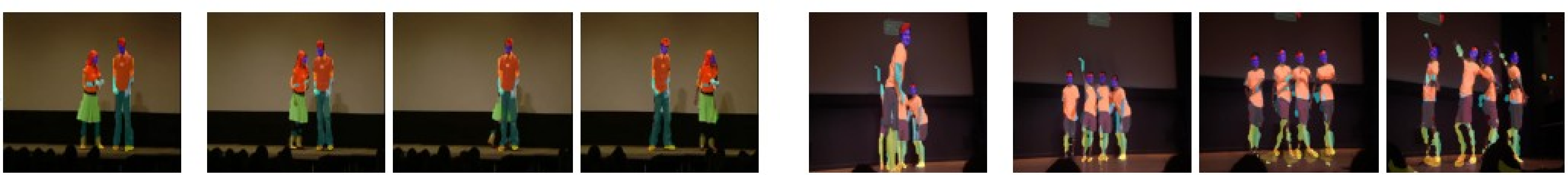
model	Supervised	mIoU	AP^{r}_{vol}
Identity		13.6	4.0
Optical Flow (FlowNet2) [22]		16.1	8.3
SIFT Flow [39]		$\boxed{21.3}$	10.5
Transitive Inv. [74]		19.4	5.0
DeepCluster [8]		$\boxed{21.8}$	8.1
Ours (ResNet-50)		**28.9**	$\boxed{\mathbf{15.6}}$
ImageNet (ResNet-50) [18]	✓	34.7	$\boxed{16.1}$
Fully Supervised [85]	✓	37.9	24.1



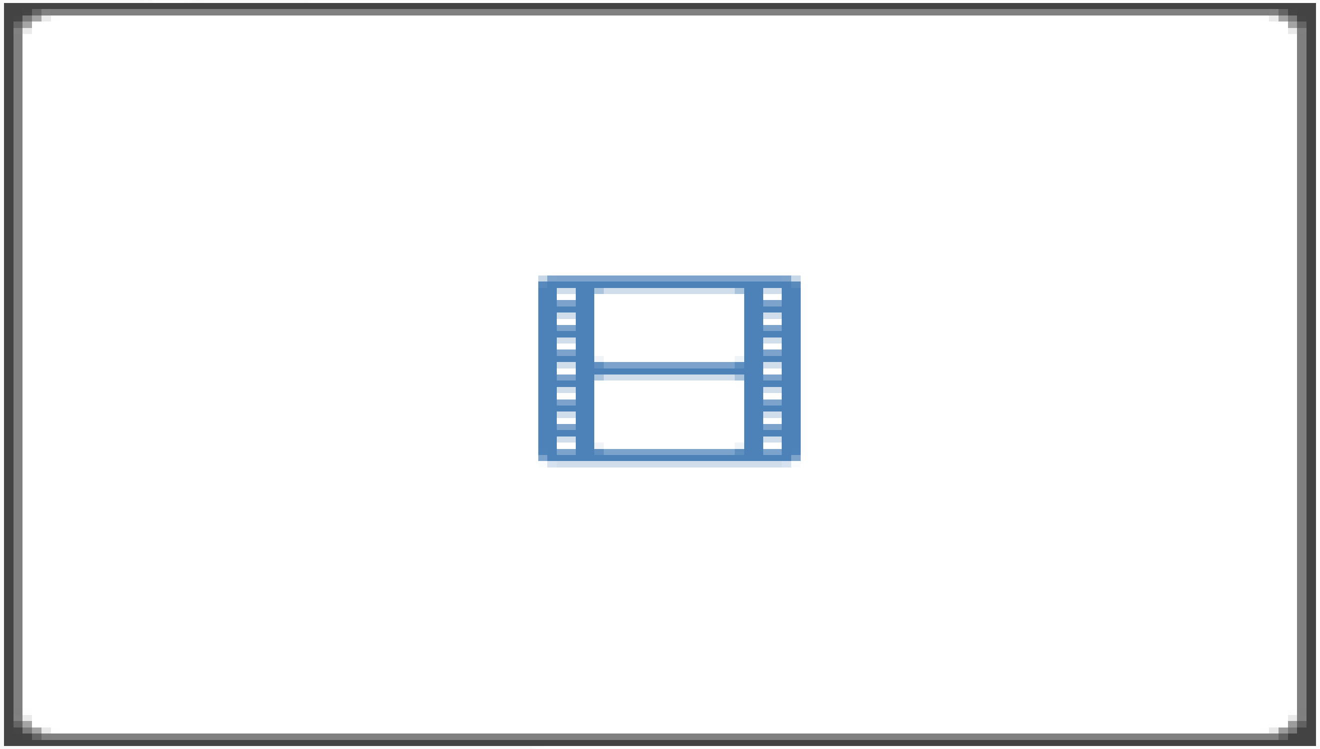
Experiment III
Semantic and Instance Propagation on VIP

model	AP^{r}_{vol}	IoU threshold 0.3	0.5	0.7
Ours (ResNet-50)	15.6	23.0	**12.7**	**5.4**
ImageNet (ResNet-50) [18]	**16.1**	**24.2**	11.9	4.8

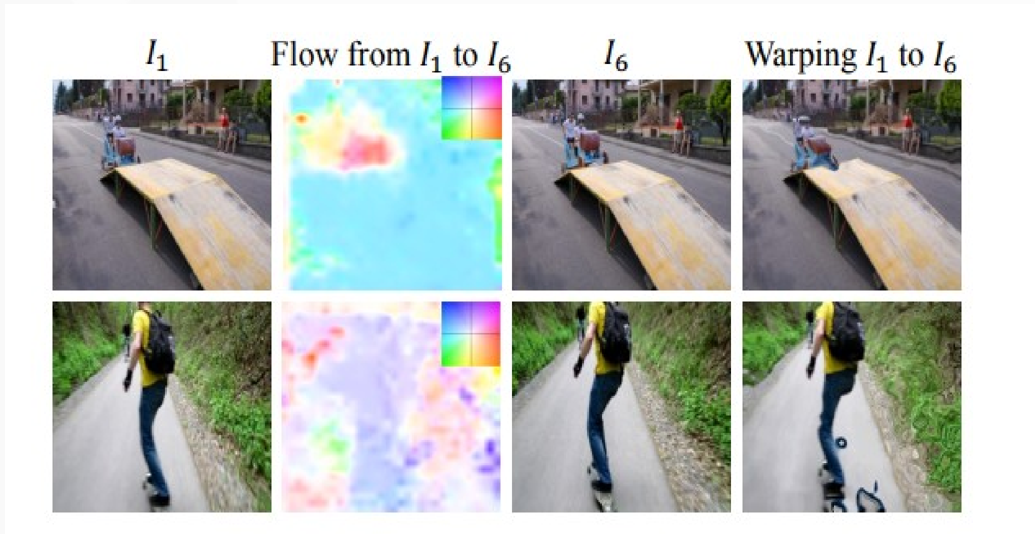
- ImageNet preforms better at smaller threshold.
- ImageNet has advantage at coarse corresponding.
- The “Ours” model preforms better at spatial precision.



Experiment IV
Texture Propagation



Experiment V
Video Frame Reconstructions



model	5-F	10-F
Identity	82.0	97.7
Optical Flow (FlowNet2) [22]	62.4	90.3
ImageNet (ResNet-50) [18]	64.0	79.2
Ours (ResNet-50)	**60.4**	**76.4**

Limitations and Conclusion

- Despite the method should keep improving with more data, in practice, after a considerable amount of time learning seems to enter state of little change.
- Correspondence learning using cycle consistency was shown to outperform most of the unsupervised methods.
- Yet in practice it did not outperform supervised approaches such as ImageNet.

References

- A. Jabri, Alexei A. Efros and Xiaolong Wang, Learning Correspondence from the Cycle-consistency of Time.
- Xiaolong Wang and Abhinav Gupta, Unsupervised Learning of Visual Representations using Videos.
- David Fouhey, Weicheng Kuo, Alexei Efros and Jitendra Malik, The VLOG Dataset.
- Animated GIF from : Xiaolong Wang, https://github.com/xiaolonw/TimeCycle
- Videos from : Xiaolong Wang, https://youtu.be/vk1DwG75ewQ