Learning to Track Instances without Video Annotations

Yang Fu, Sifei Liu, Umar Iqbal, Shalini De Mello, Humphrey Shi,
Jan Kautz

CVPR 2021 (Oral)

Keywords

- Instance tracking
- Semi-supervised learning
- Contrastive Learning
- Maximum Entropy Regularization
- Video Instance Correspondence
- Test-time Adaptation

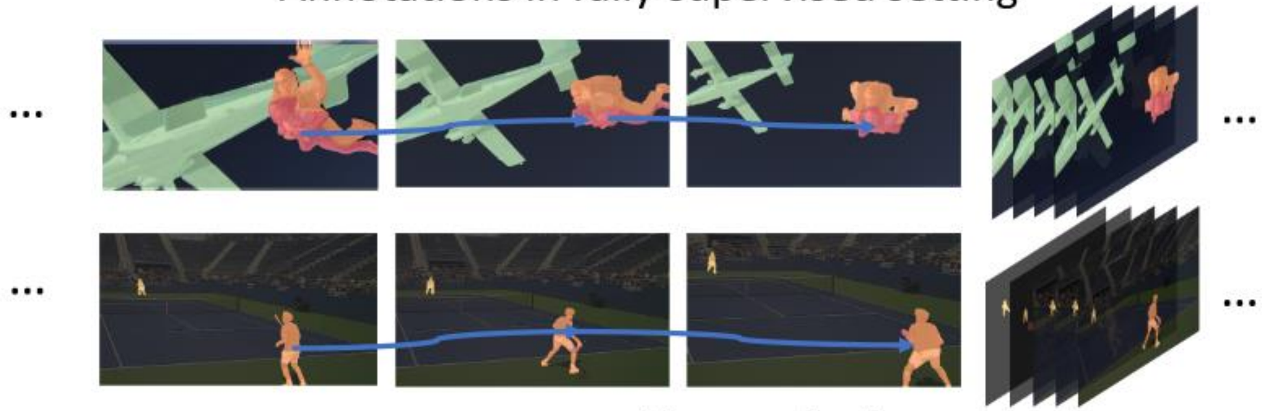
Main task
: Video Instance Segmentation

Linjie Yang, et al., Video Instance Segmentation, ICCV 2019

Problem Statement

- How to learn to track instances **without annotation** in video
- Annotation of videos requires **excessive labor** especially in a per-frame manner and becomes the **major bottleneck** for framewise video processing

Annotations in fully-supervised setting

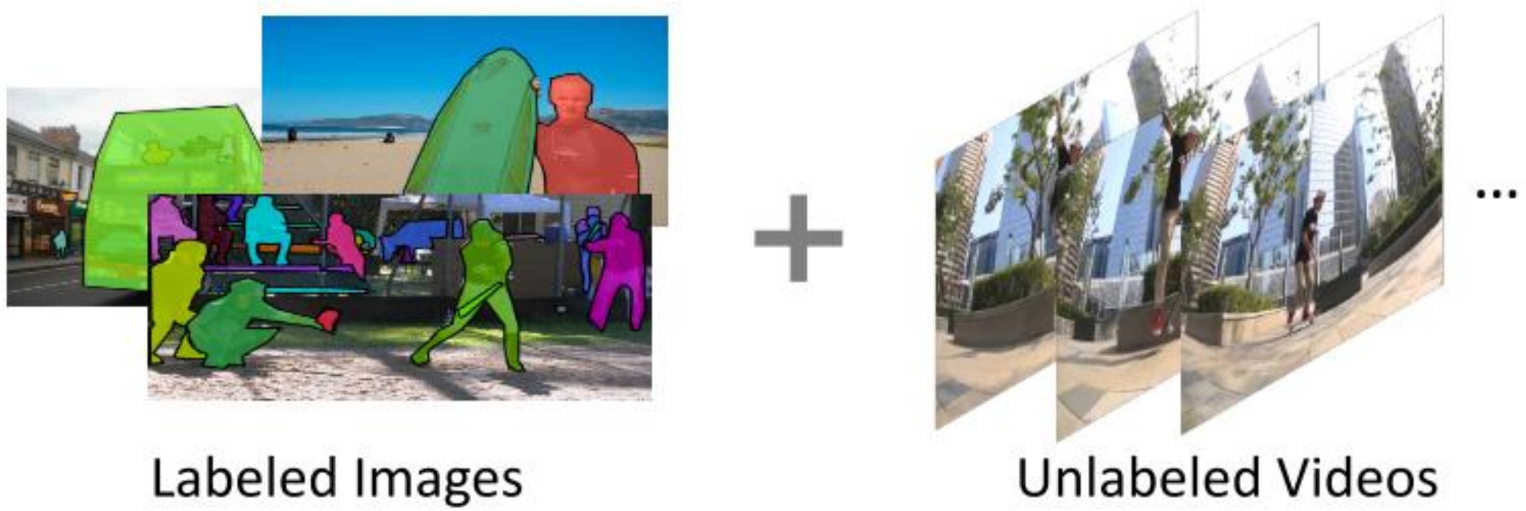


⋮ large-scale videos

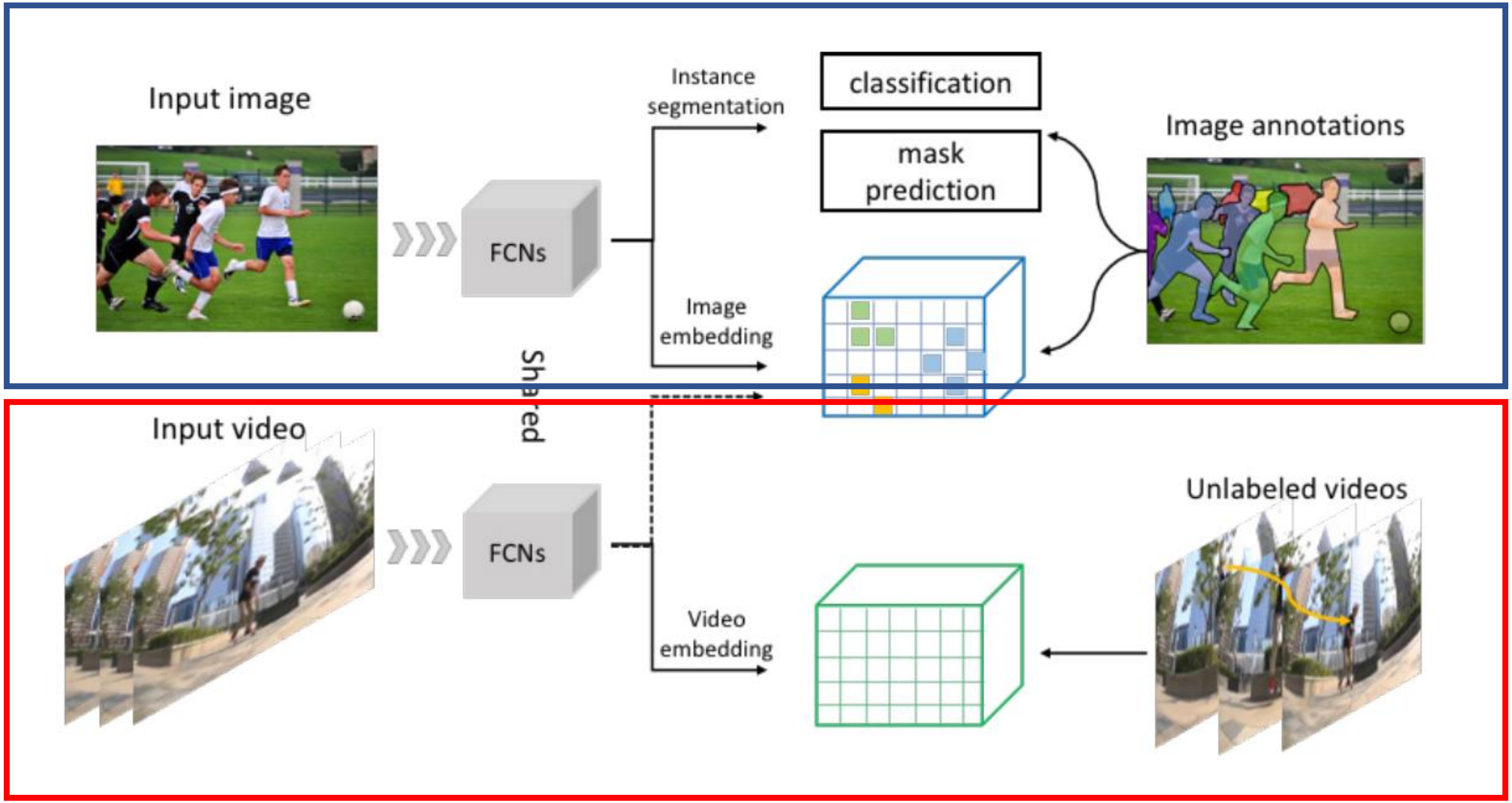
How to solve the Problem

1. Trained with images in a supervised manner

2. Enhance the tracking capability of the embedding by learning correspondence from unlabeled videos in a self-supervised manner



Proposed Method



Backgrounds
: SOLO

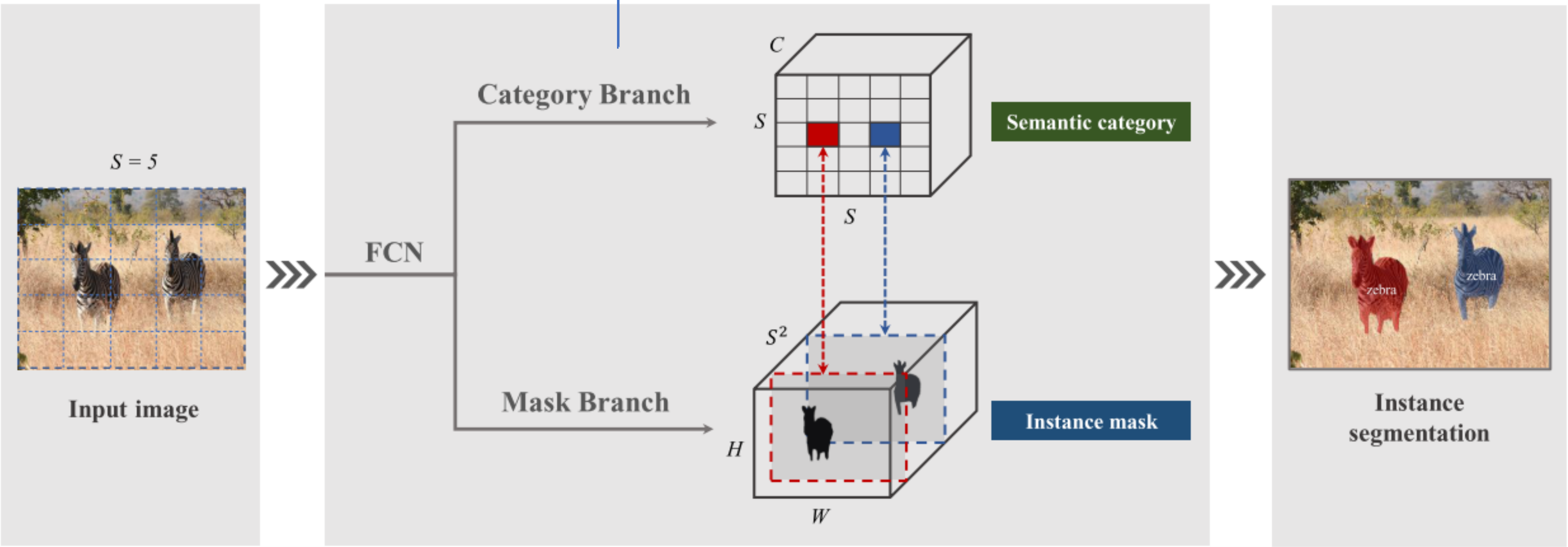
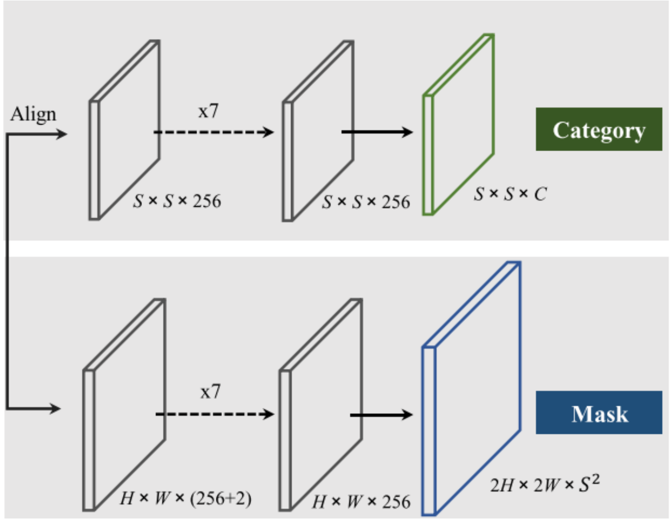


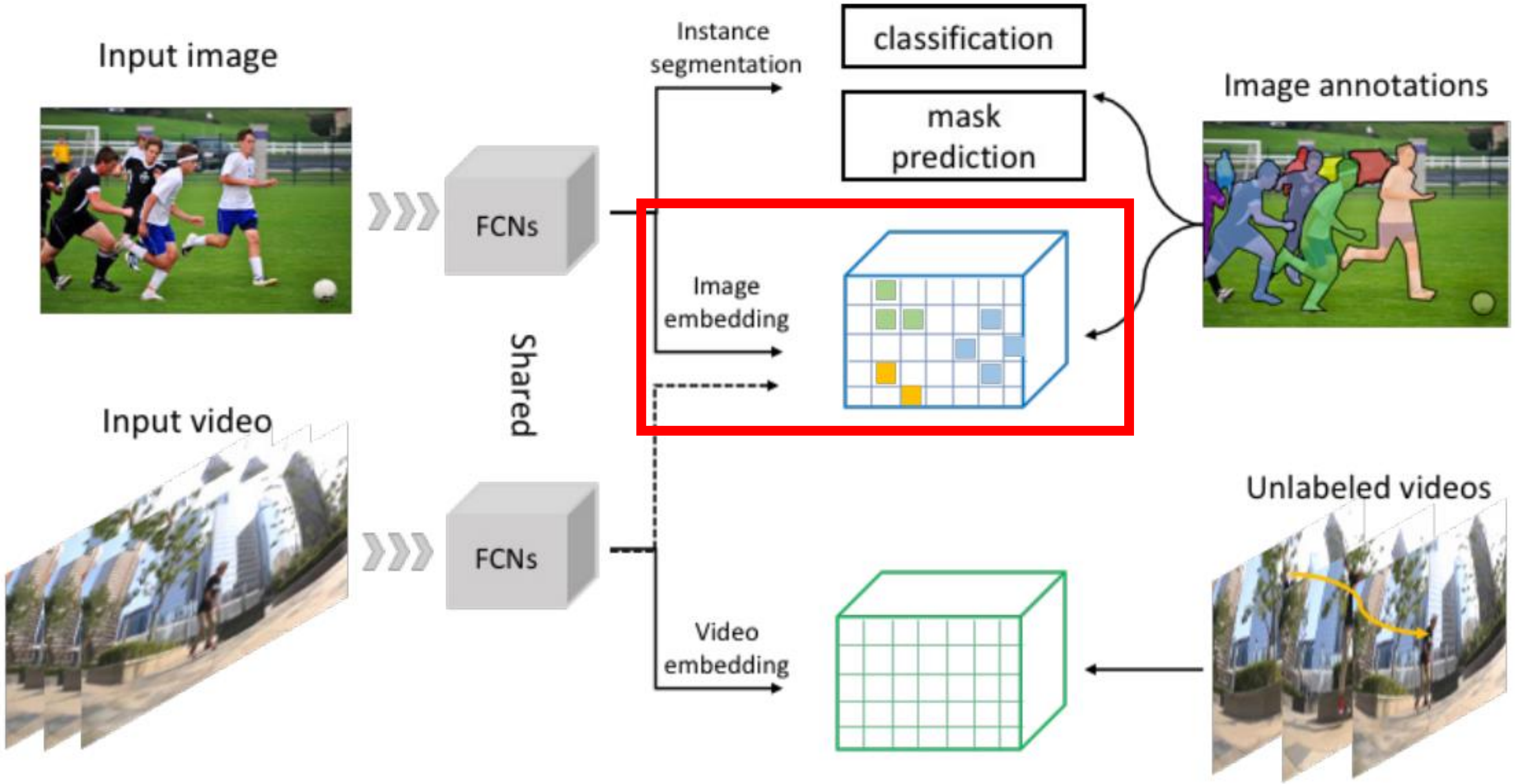
Fig. 2. SOLO framework. We reformulate the instance segmentation as two sub-tasks: category prediction and instance mask generation problems. An input image is divided into a uniform grids, *i.e.*, $S \times S$. Here we illustrate the grid with $S = 5$. If the center of an object falls into a grid cell, that grid cell is responsible for predicting the semantic category (top) and masks of instances (bottom). We do not show the feature pyramid network (FPN) here for simpler illustration.

Details in Category Branch and Mask Branch



Xinlong Wang, et al., Solo: Segmenting objects by locations, ECCV 2020

Proposed Method



Proposed Method

- Why it needs image, video embedding part?

  As assigning ids to each object, the model **need to know each instance** so that it **can track** those possible

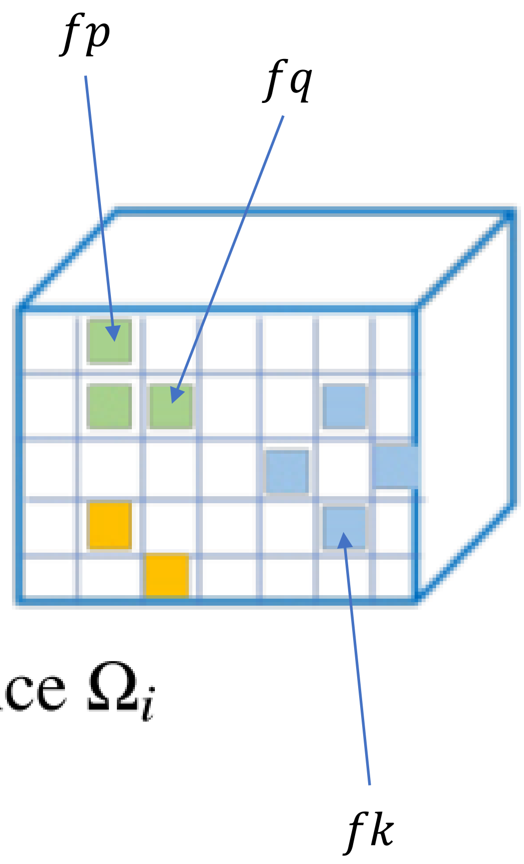
- How to learn embedding efficiently
  1. discriminative of different instances
  2. consistent regardless of the variations present in videos.
  3. focus more on appearance rather than location, since objects can move in time.

Proposed Method

: Image embedding (Instance Contrastive Loss)

- Objective : discriminative of different instances
- Solution : Contrastive Loss

$$\mathcal{L}_q = -\log \frac{\exp(f_p^\top \cdot f_q)}{\sum_{k \in \Omega_{\bar{i}}} \exp(f_k^\top \cdot f_q)}, \quad p, q \in \Omega_i$$



one query grid cell $x_q \in X$ with feature f_q from the i^{th} instance Ω_i

vector f_p from the same instance as the positive sample

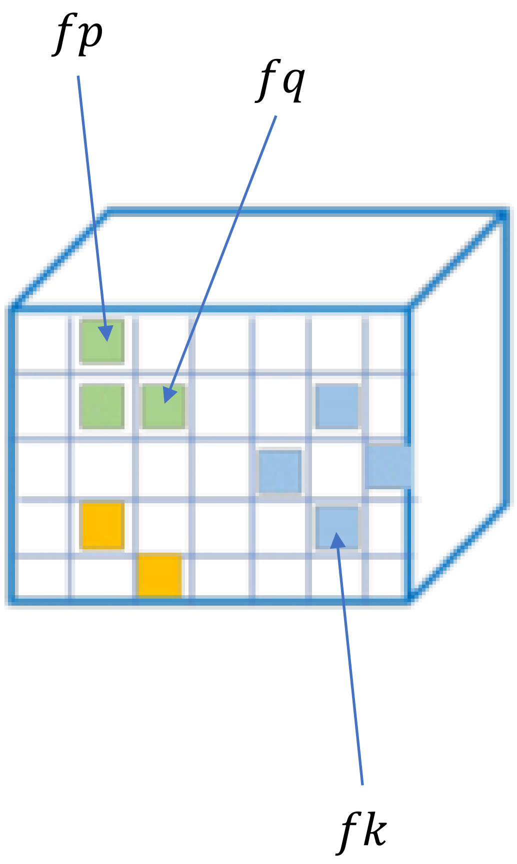
$\Omega_{\bar{i}}$ is the set of cells from all the other instances $\bar{i}$.

Aaron van den Oord et al., Representation learning with contrastive predictive coding. arXiv 2018

Proposed Method
: Image embedding (Instance Contrastive Loss)

- Objective : discriminative of different instances
- Solution : Contrastive Loss

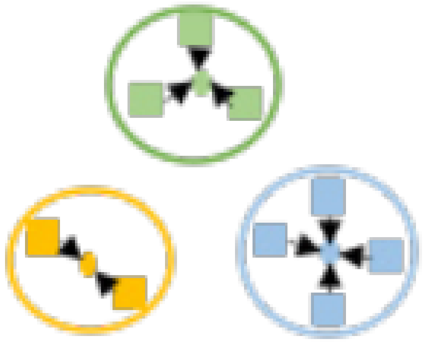
$$\mathcal{L}_q = -\log \frac{\exp(f_p^\top \cdot f_q)}{\sum_{k \in \Omega_{\bar{i}}} \exp(f_k^\top \cdot f_q)}, \quad p, q \in \Omega_i$$



Smaller instances will be **insufficiently trained** due to less positive samples !

Aaron van den Oord et al., Representation learning with contrastive predictive coding. arXiv 2018

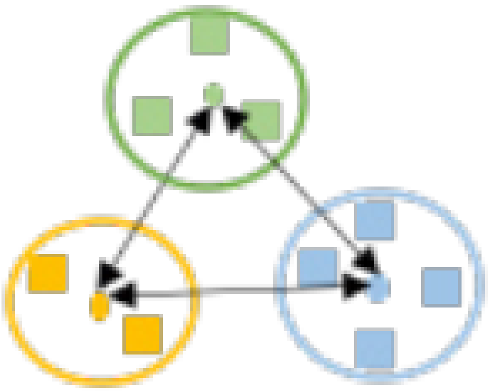
Proposed Method
: Image embedding (Instance Contrastive Loss)



$$C_i = \frac{1}{N_i} \sum_{q \in \Omega_i} f_q \qquad \mathcal{L}_i^{\text{center}} = \sum_{q \in \Omega_i} \| C_i - f_q \|_1 .$$

- Force the embedding feature vectors of the same instance to be similar

Proposed Method
: Image embedding (Instance Contrastive Loss)



$$S(i,j) = \frac{\exp(C_i^\top \cdot C_j)}{\sum_{k=0}^{K} \exp(C_i^\top \cdot C_k)}, \qquad \mathcal{L}^{\text{contra}} = \mathbf{CE}(S, I).$$

K is the number of instances in an image

- Push the center representation of all the instances further apart

Proposed Method
: Image embedding (Instance Contrastive Loss)

$$\mathcal{L}^{\mathrm{IC}} = \sum_{i=0}^{K} \mathcal{L}_i^{\mathrm{center}} + \lambda \mathcal{L}^{\mathrm{contra}}.$$

Proposed Method
: Image embedding (maximum entropy regularization)

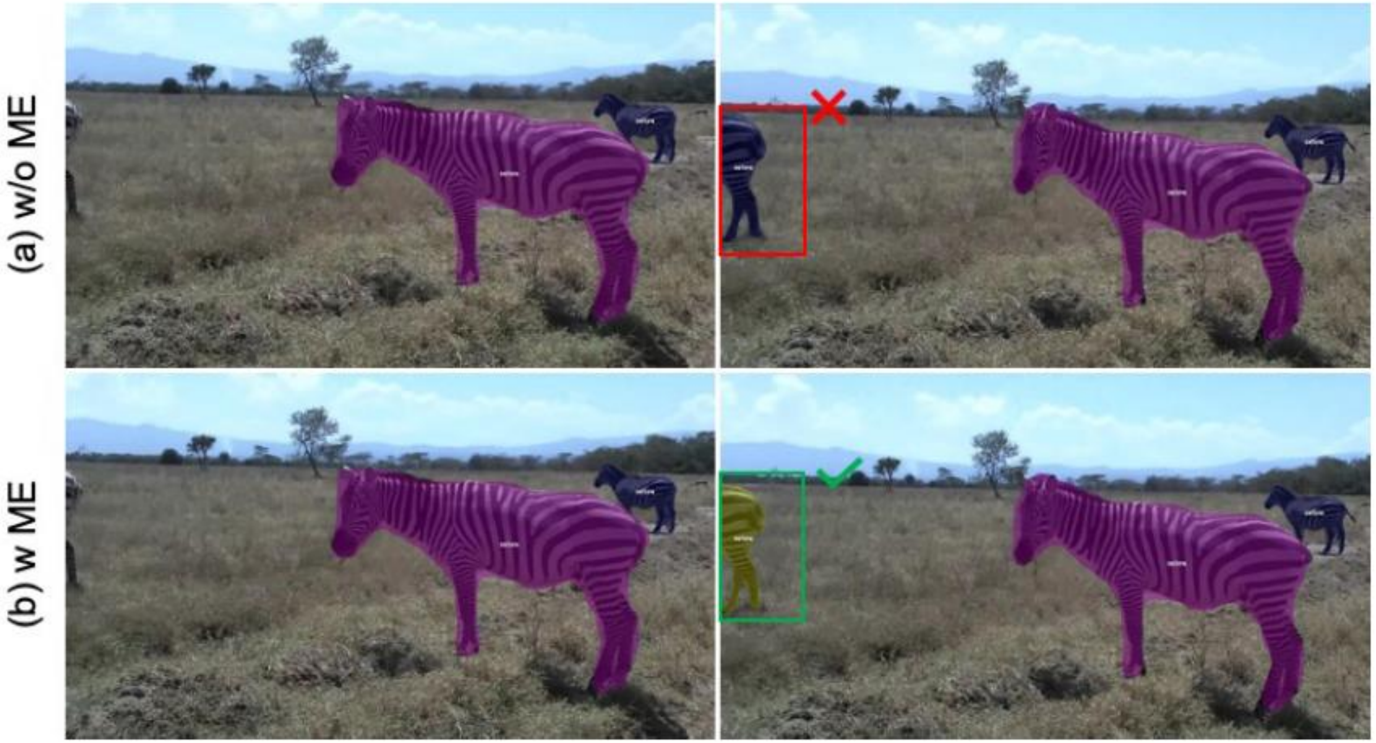
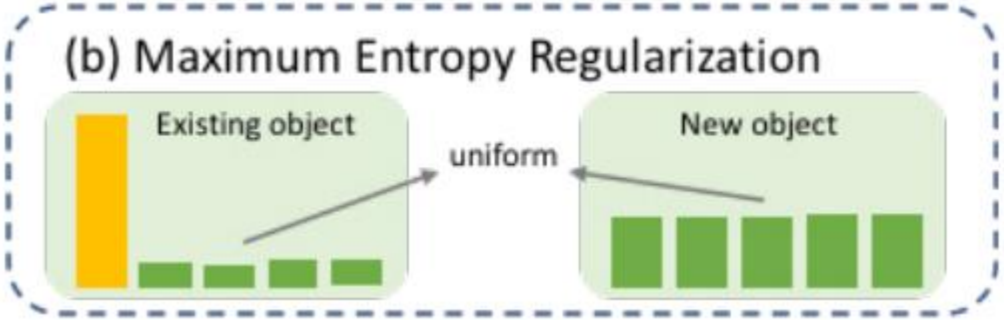


Figure 3. An illustration of failure case when a new object appears and the effectiveness of maximum entropy (ME) regularization. Row (a) and (b) are results without and with ME regularization. Best viewed in color and zoom in to see details.

- Our tracking approach is based on the assumption that any instance in the current frame also exists in the previous frame.
- It doesn't consider **newly emerged objects**.

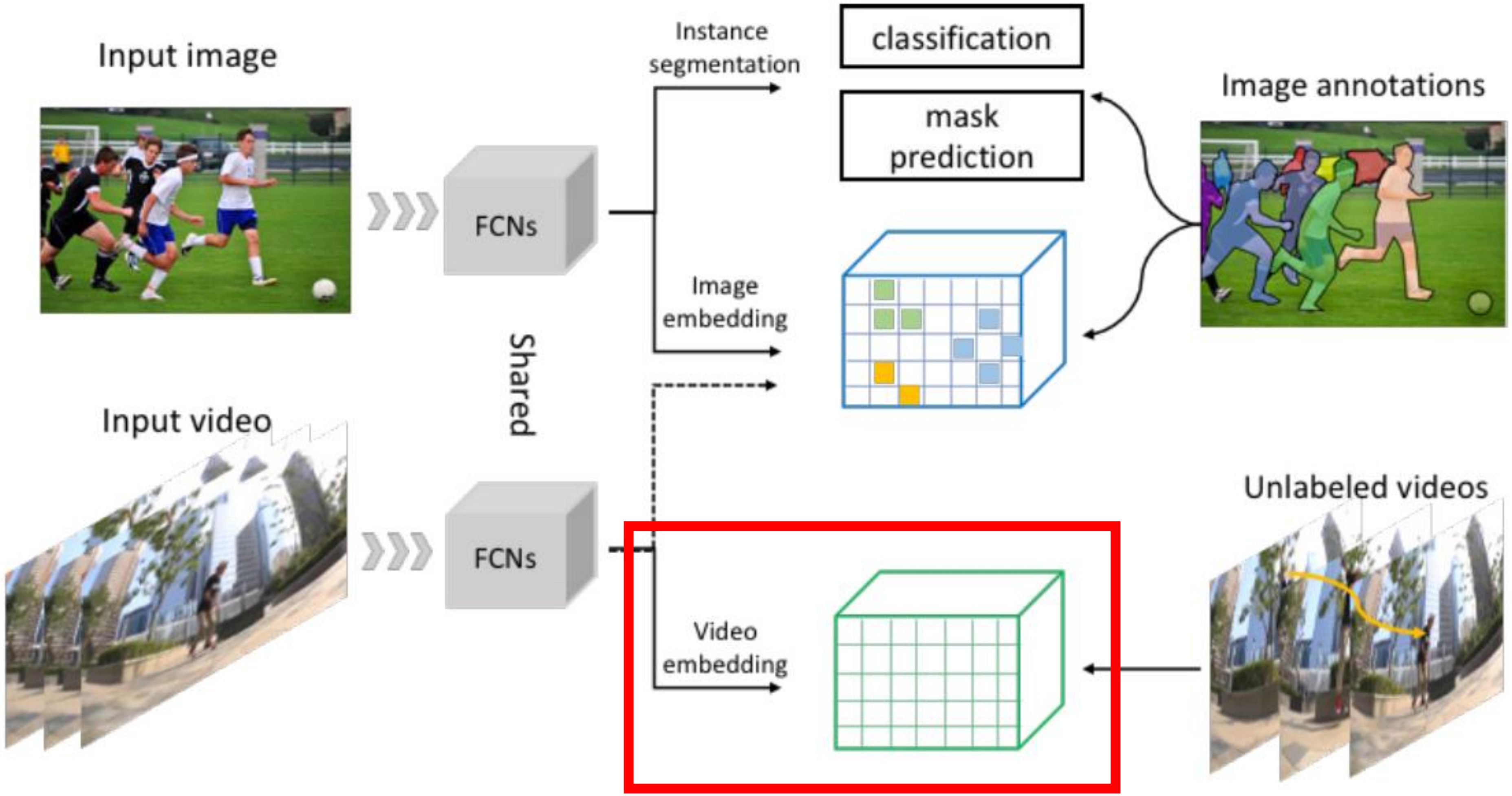
Proposed Method
: Image embedding (maximum entropy regularization)

$$H = -\sum_{i}^{K} \sum_{j \neq i}^{K} S(i,j) \log(S(i,j)),$$

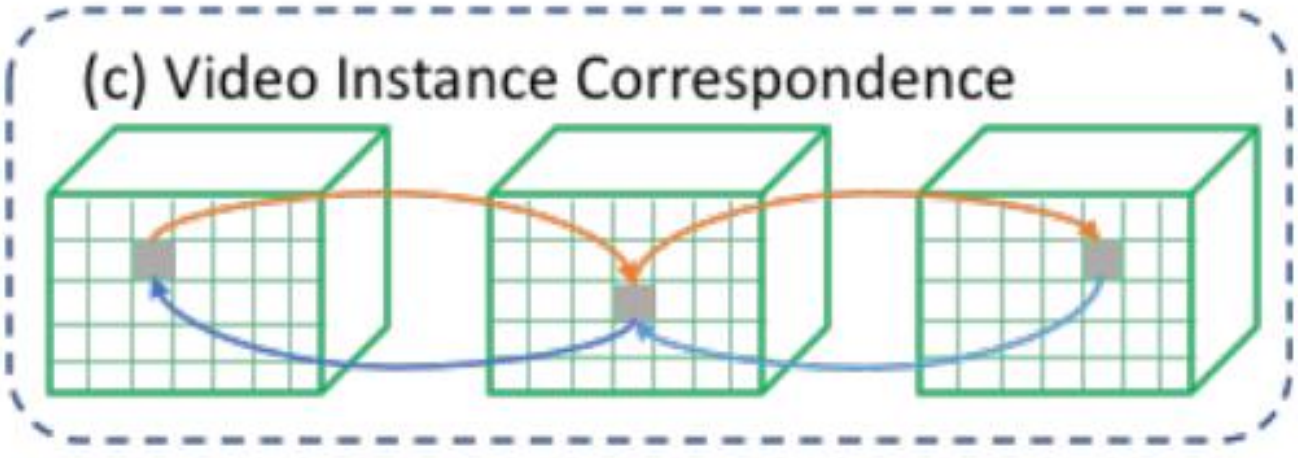


- K is the number of instances and S(i,j) is the probability of matching instance i to j
- High entropy H indicates uniform output probability

Proposed Method

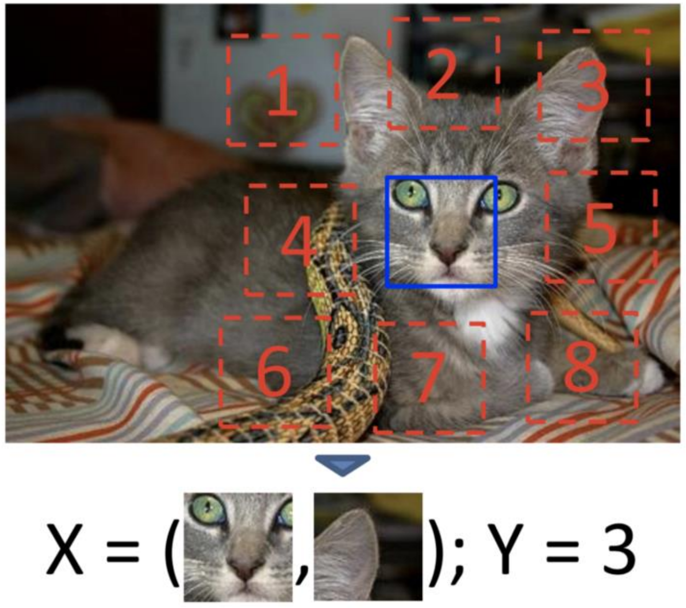


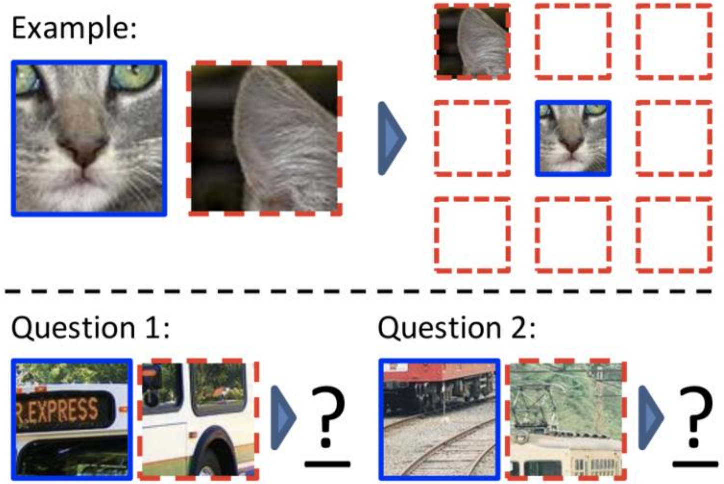
Proposed Method
: Video embedding (Video Instance Correspondence)



- With videos, we also need to address the **domain gap** that usually exists between image and videos
- Solution : self-supervised video correspondence learning

Background
: Self-supervise learning

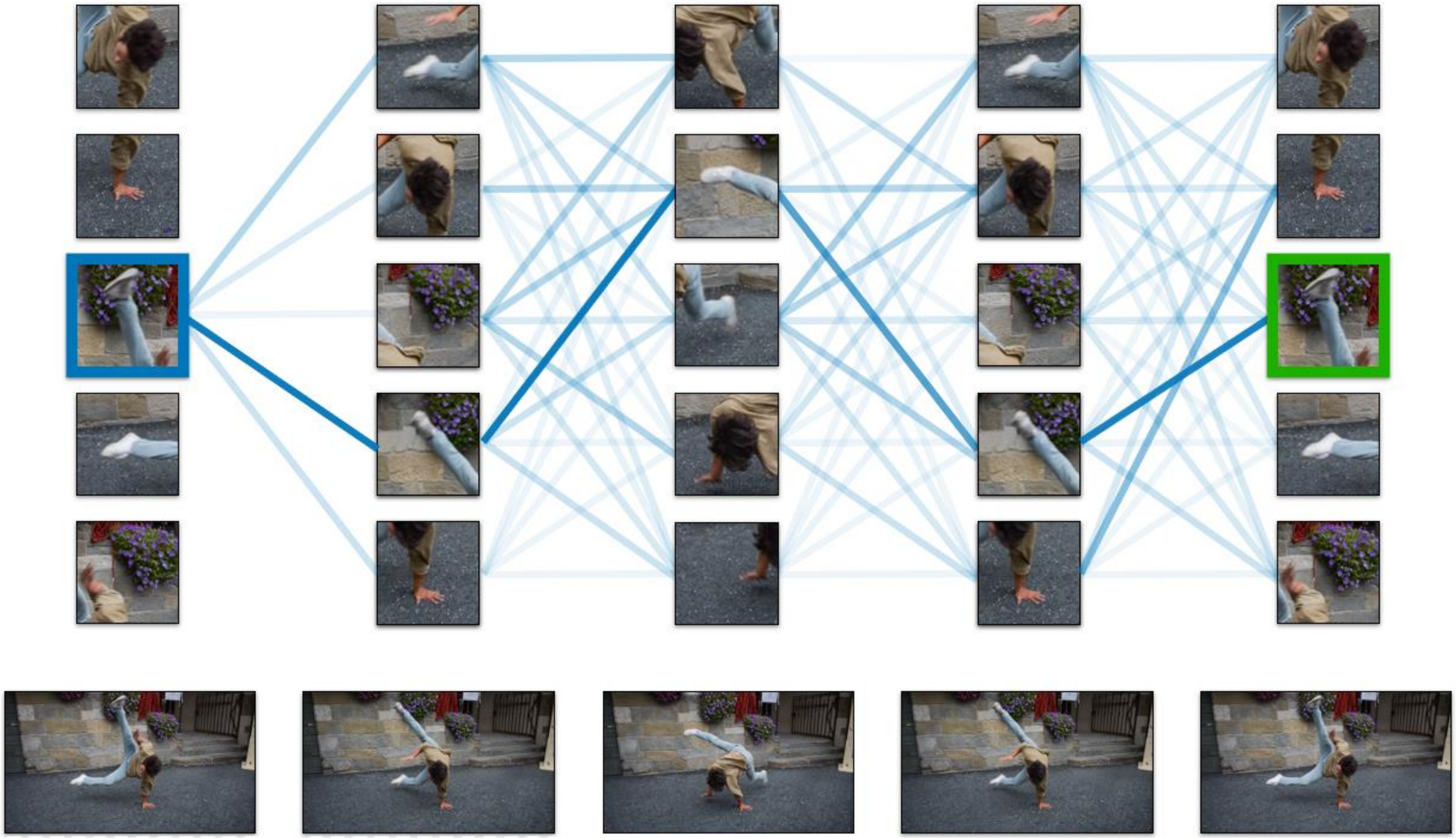




- Pretext task: get supervision from the **data itself**
- Self-supervised representation learning care about producing good features **generally helpful for many tasks**

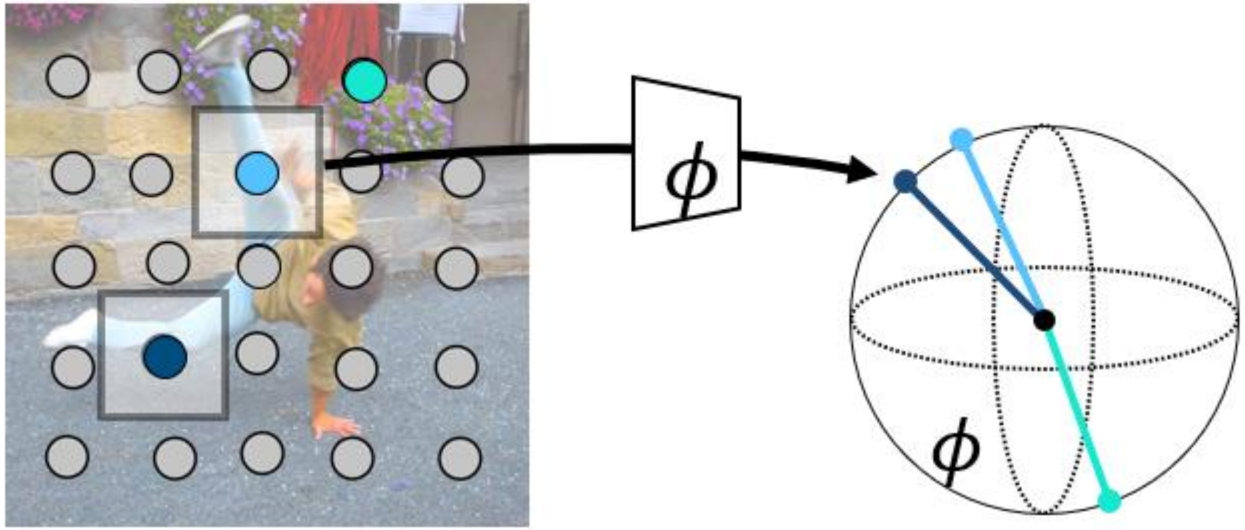
Carl Doersch ,et al. Unsupervised Visual Representation Learning by Context Prediction. ICCV 2015

Self-supervised Learning



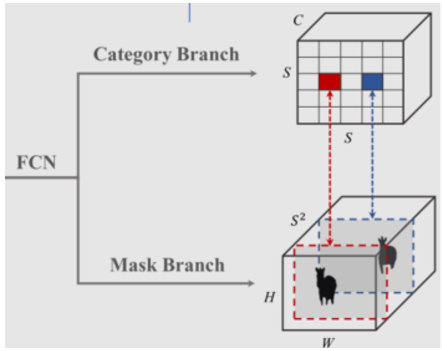
Allan Jabri ,et al. Space-Time Correspondence as a Contrastive Random Walk. NeurIPS 2020, Oral

Proposed Method
: Video embedding (Video Instance Correspondence)



frame

256 x 256 px



nodes

64 x 64 px

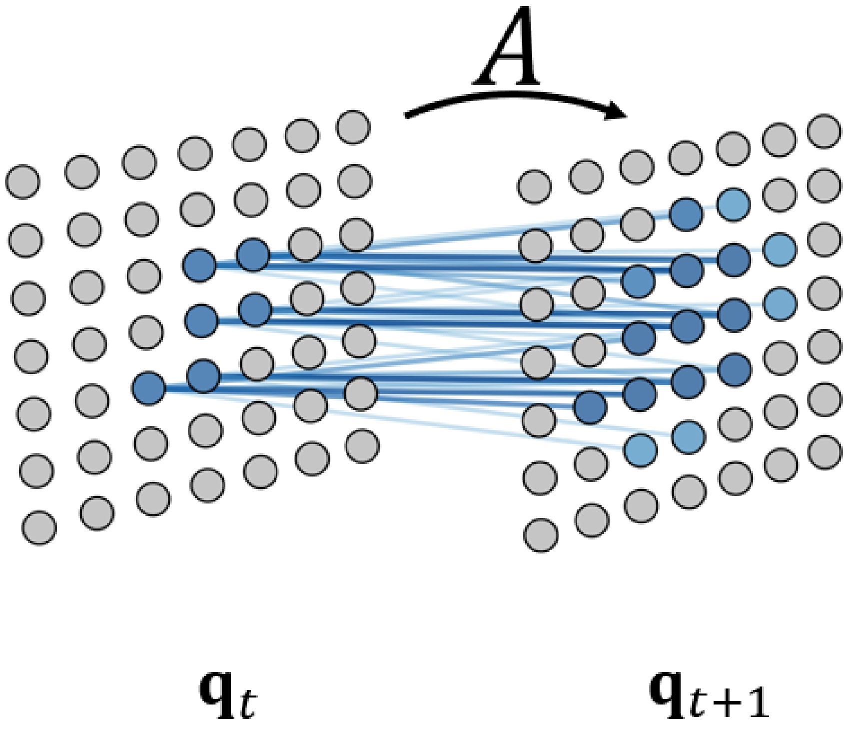
patches

embeddings

$\in \mathbb{R}^{128}$

Allan Jabri ,et al. Space-Time Correspondence as a Contrastive Random Walk. NeurIPS 2020, Oral

Video as a Graph



$$A_{ij} = \frac{e^{d_\phi(q_t^i,\, q_{t+1}^j)/\tau}}{\sum_l e^{d_\phi(q_t^i,\, q_{t+1}^l)/\tau}}$$

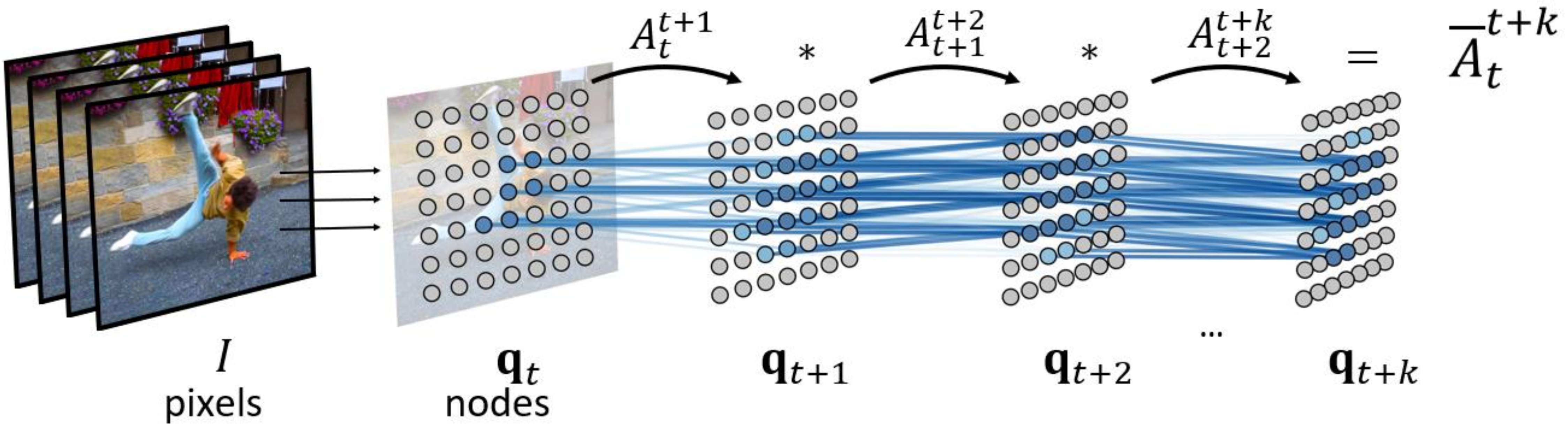
$$= P(X_{t+1} = j | X_t = i)$$

where $d_\phi(x, y) = \phi(x)^\top \phi(y)$

X_t is the position of walker at time t

Allan Jabri ,et al. Space-Time Correspondence as a Contrastive Random Walk. NeurIPS 2020, Oral

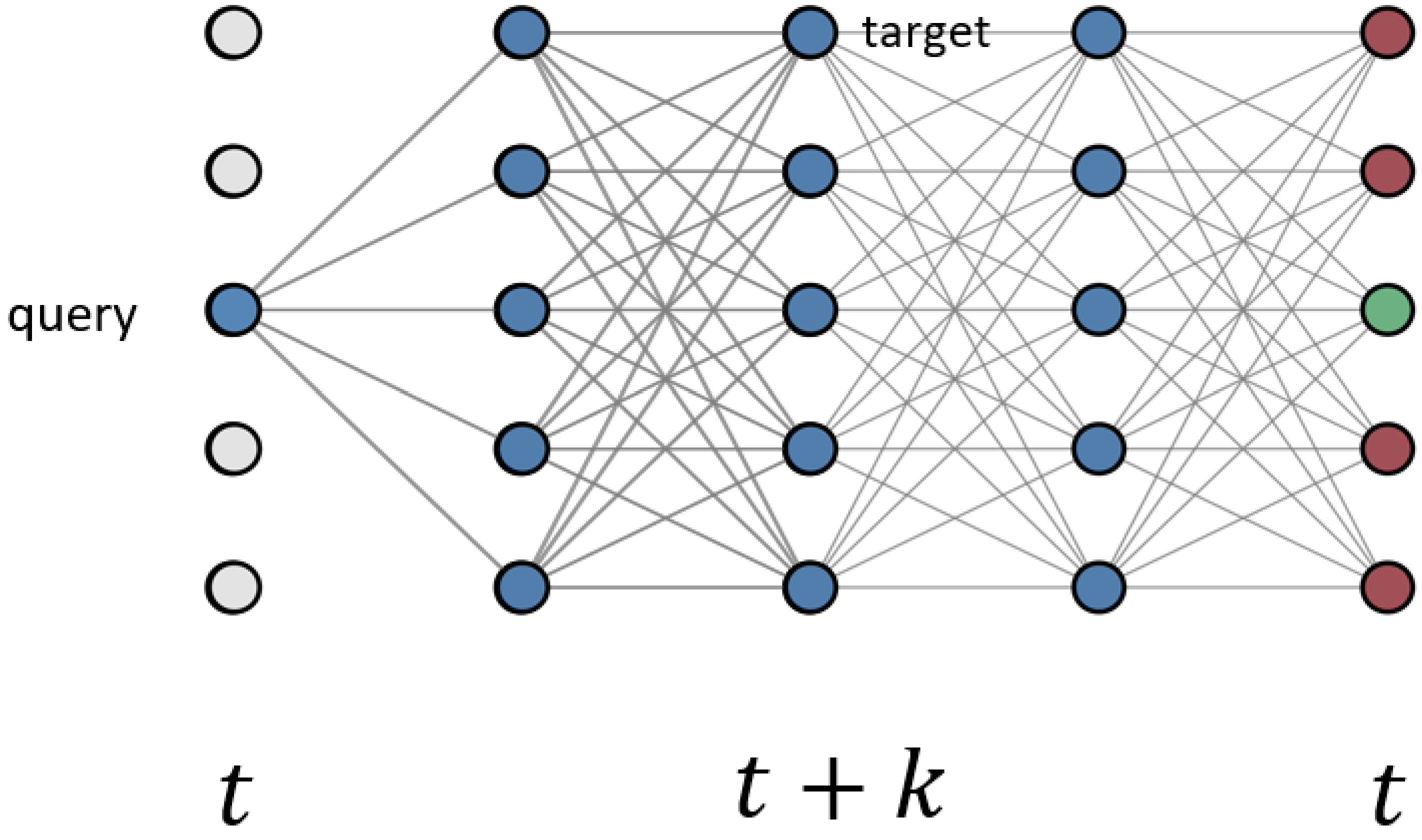
Proposed Method
: Video embedding (Video Instance Correspondence)



Learn representation ϕ = Fit transition probabilities $\bar{A}_t^{t+k}$

Allan Jabri ,et al. Space-Time Correspondence as a Contrastive Random Walk. NeurIPS 2020, Oral

Proposed Method
: Video embedding (Video Instance Correspondence)



Train on Palindromes

$$\mathcal{L}_{cyc}^{k} = \mathcal{L}_{CE}(\bar{A}_t^{t+k}\bar{A}_{t+k}^{t}, I)$$

Allan Jabri ,et al. Space-Time Correspondence as a Contrastive Random Walk. NeurIPS 2020, Oral

Results

Methods	Video Annotations	With Embed	Contrastive Loss	Max Entropy	Video Correspondence	AP	$AP_{0.5}$	$AP_{0.75}$	AR_1	AR_{10}
MaskTrack-RCNN [43]	✓	✓				29.0	47.5	32.2	28.7	32.4
SOLO [40]						23.9	43.3	21.5	26.7	37.3
SOLO-Track		✓	✓			28.4	50.0	30.4	27.6	34.4
SOLO Track		✓	✓	✓		29.7	52.8	29.9	30.7	34.9
SOLO-Track		✓	✓	✓	✓	**32.9**	**54.4**	**35.0**	**34.1**	**40.8**

Table 1. Ablation study with different proposed components on YouTube-VIS validation set. The best results are highlighted in bold.

Results

Methods	AP	$AP_{0.5}$	$AP_{0.75}$	AR_1	AR_{10}
Video + Image Annotations					
MaskTrack R-CNN [43]	29.0	47.5	32.2	28.7	32.4
SipMask [5]	24.1	42.0	26.0	26.2	28.6
Only Image Annotations					
Ours	29.7	52.8	29.9	30.7	34.9
Ours$^+$	**32.9**	**54.4**	**35.0**	**34.1**	**40.8**
After post-processing					
Video + Image Annotations					
IoUTracker+ [43]	29.4	48.5	30.6	32.1	34.2
SeqTracker [43]	31.8	52.2	35.8	32.2	34.4
MaskTrack R-CNN [43]	36.0	58.4	**40.2**	35.4	38.9
SipMask [5]	**37.7**	57.8	38.0	37.4	40.3
Only Image Annotations					
Ours	34.1	58.0	37.9	33.0	39.2
Ours$^+$	37.4	59.7	39.1	36.4	43.8
Ours*	**38.3**	**61.1**	39.8	36.9	**44.5**

Table 3. Comparison of the our approach with the SOTA methods on the YouTube-VIS validation set. "Ours" represents the model with instance embedding branch trained with IC loss and ME regularization. "Ours$^+$" stands for the model with the video correspondence module as well. "Our* is the model updated by test-time adaptation upon "Ours$^+$". The best results are highlighted in bold.

Supplementary

- Post Processing
  we also apply the post-processing procedure introduced in, which combines category confidence, bounding box Intersection over Union (IoU), embedding similarity and category consistency through a weighted sum.

$$s(n,m) = \mathrm{sim}(n,m) + \alpha \mathrm{c}(\mathrm{n}) + \beta \mathrm{IoU}(\mathrm{b_n}, \mathrm{b_m}) + \gamma \delta(\mathrm{c_n}, \mathrm{c_m}) \tag{1}$$

where $\mathrm{c}(\mathrm{n})$ is the classification confidence score of the nth object, $\mathrm{c_n}$ is the predicted category and $\delta(\mathrm{c_n}, \mathrm{c_m})$ is the Kronecker delta function, which returns one if and only if $\mathrm{c_n}$ is equal to $\mathrm{c_m}$, otherwise it returns zero.