Self-Supervised Learning for Vision

Javier Selva Castelló

Universitat de Barcelona

Motivation

➢ Supervision is **costly**!

Photo source: https://javilopez.eu/precariedad-laboral/call-center/

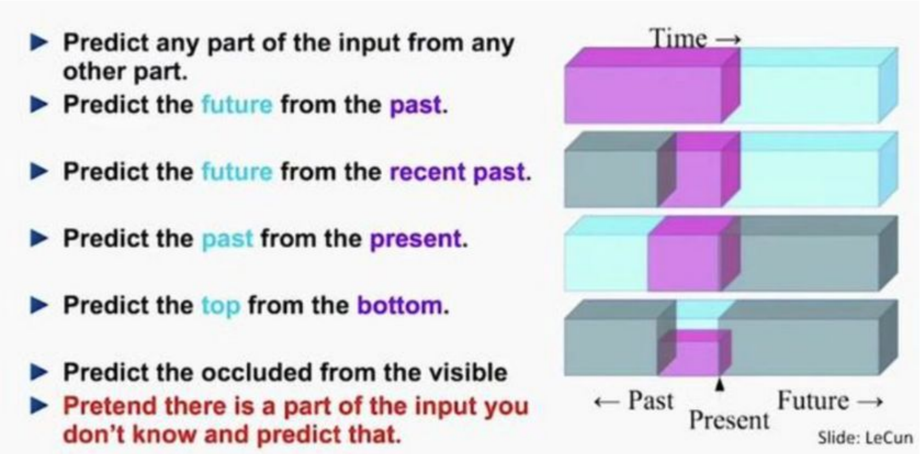
Motivation

- ➢ Supervision is **costly**!
- ➢ Billions of GB of **internet content**.

Photo source: https://i.workana.com/glossary/what-is-content-creation/

Motivation

- Supervision is **costly**!
- Billions of GB of **internet content**.
- Can we use the **data** itself **as supervision**?



(Famous illustration from Yann LeCun)

Motivation

- ➢ Supervision is **costly**!
- ➢ Billions of GB of **internet content**.
- ➢ Can we use the **data** itself **as supervision**?
- ➢ Fine-tune pre-trained model on **supervised downstream task**.

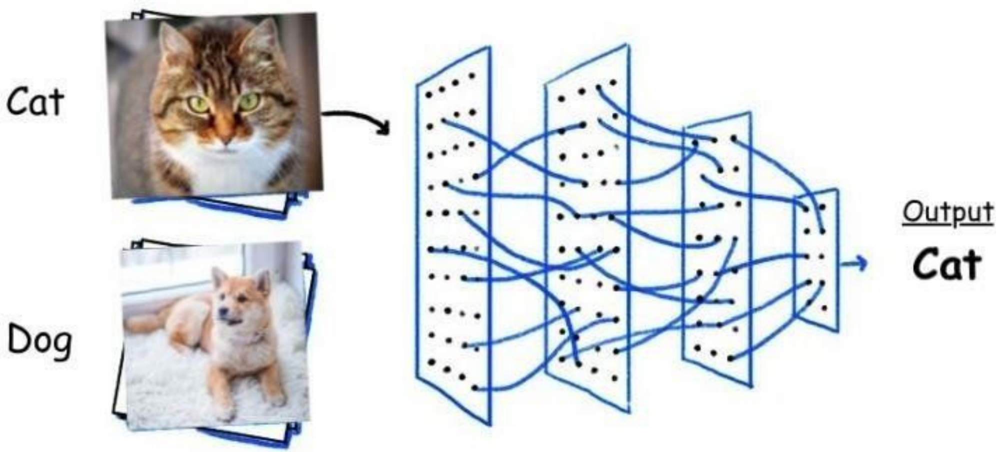
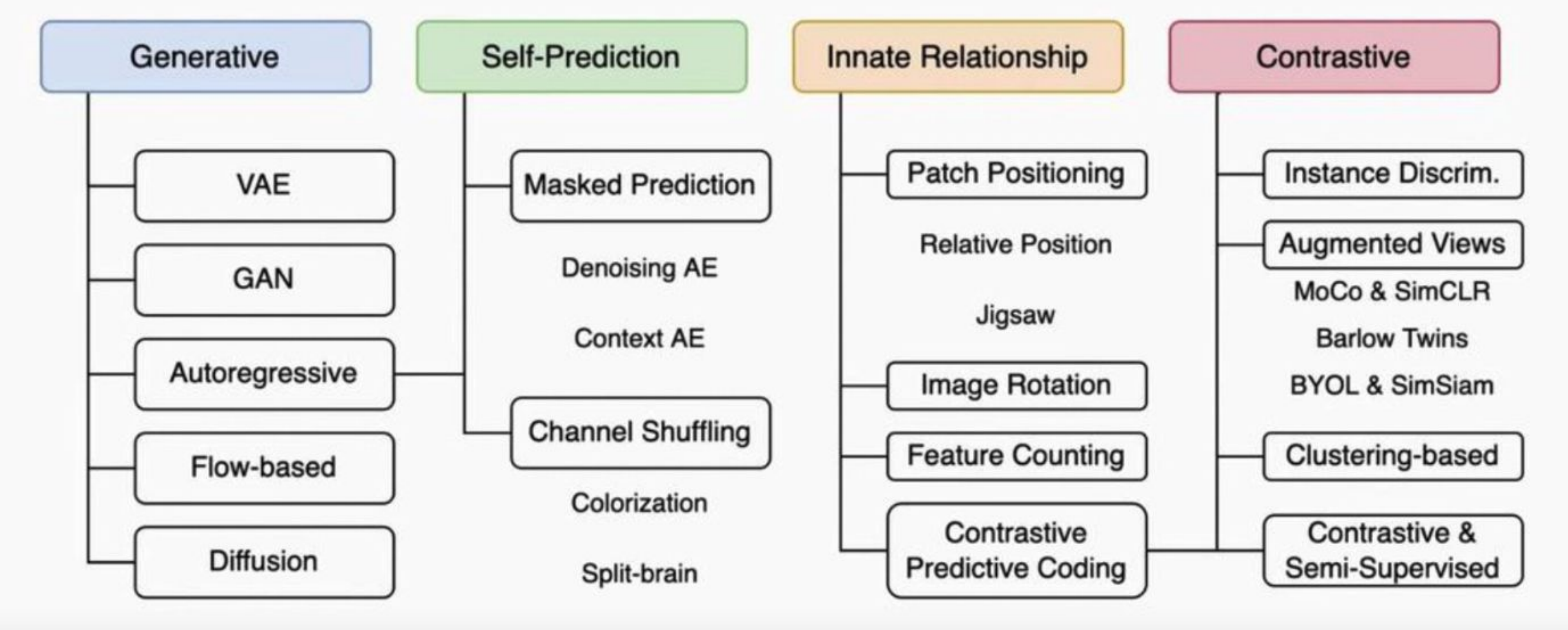


Photo source: https://www.researchgate.net/figure/Classification-of-two-classes-dog-and-cat-using-CNN_fig2_353247632

Many methods!!



Taxonomy from: Self-Supervised Learning: Self-Prediction and Contrastive Learning | Tutorial | NeurIPS 2021 https://www.youtube.com/watch?v=7l6fttRJzeU

Some background

Basics: Contrastive Learning

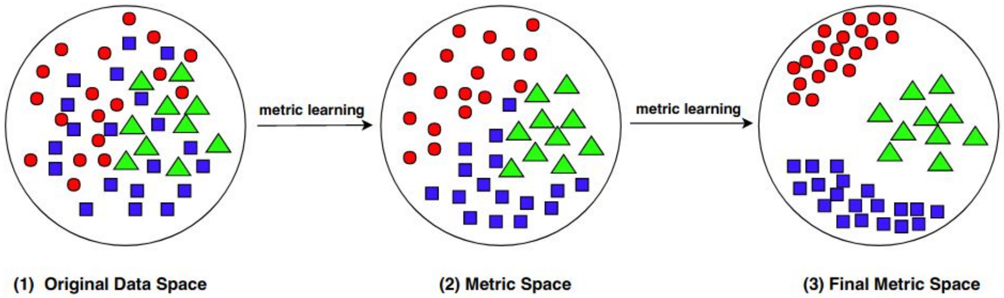
Similar samples should be **closer** in representation space.

COLLAPSE TO CONSTANT REPRESENTATION!

Different samples (negatives) should be **far away** in representation space.

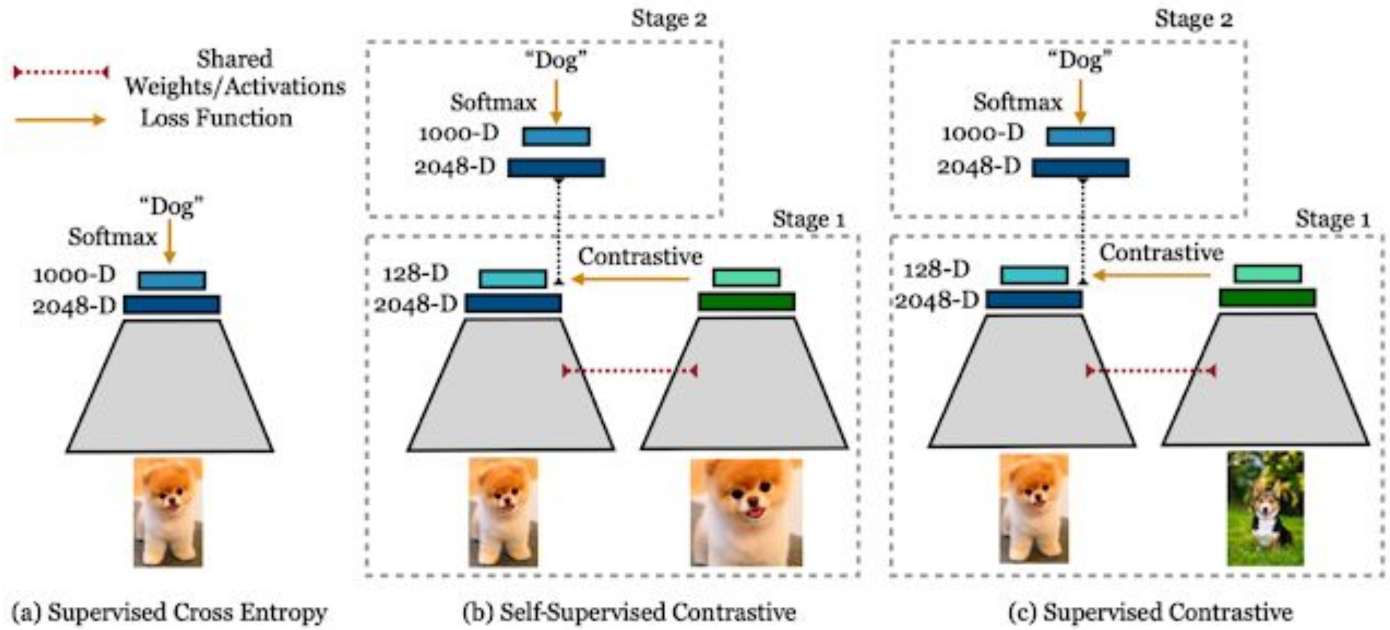
Problems:

- ➢ **Size** of negative set
- ➢ **Quality** of negatives



Wang et al. “Metric Learning based Framework for Streaming Classification with Concept Evolution”, IJCNN (2019)

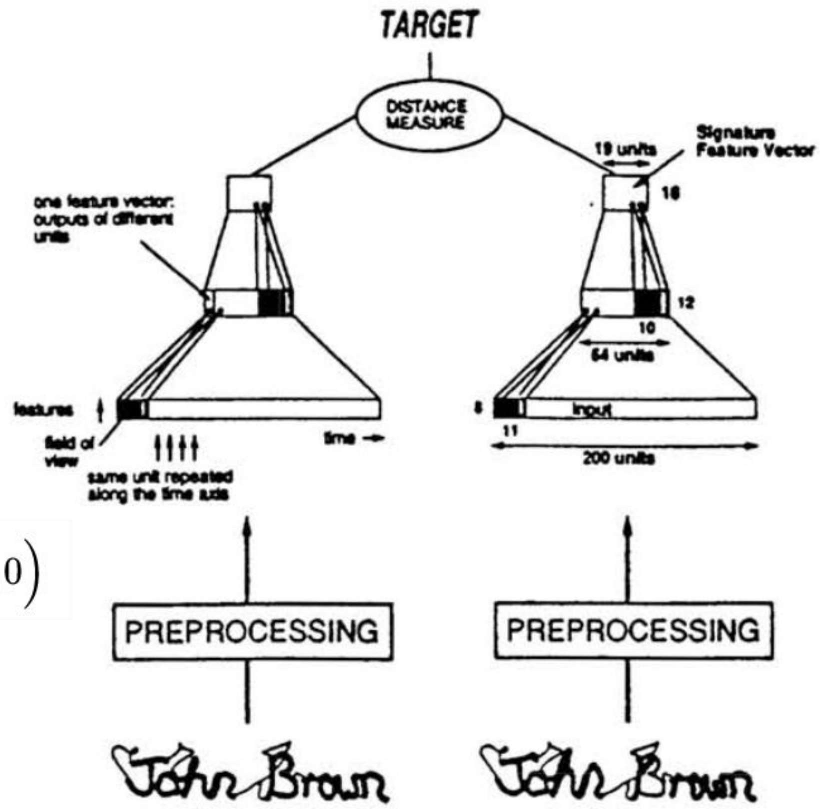
Basics: Contrastive Learning



Khosla et al. "Supervised Contrastive Learning", NeurIPS (2020)

Basics: Siamese Networks

- ➢ Two sister networks (same weights)
- ➢ Different inputs
- ➢ Trained contrastively.



Triplet loss:

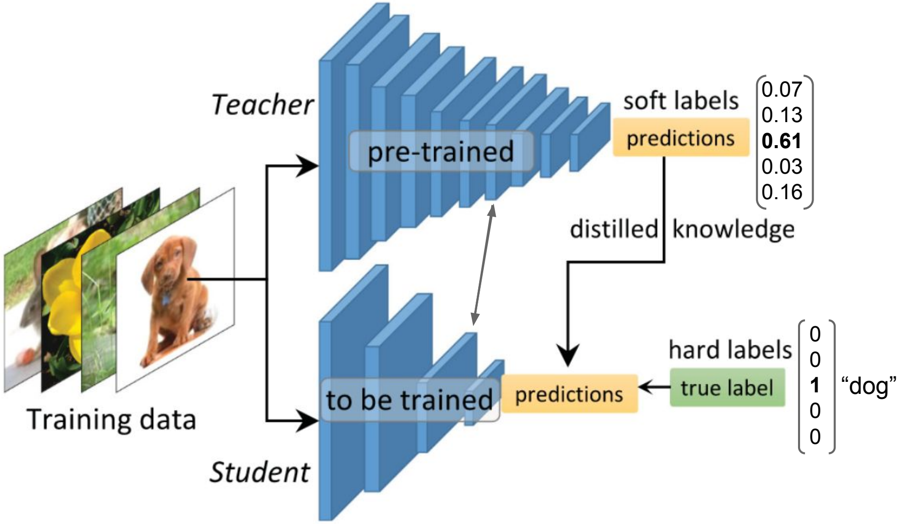
$$\mathcal{L}(A, P, N) = \max\left(\| \mathrm{f}(A) - \mathrm{f}(P)\|^2 - \| \mathrm{f}(A) - \mathrm{f}(N)\|^2 + \alpha, 0\right)$$

Bromley et al. “Signature verification using a ‘Siamese’ time delay neural network”. In ICML (1994).

Basics: Knowledge Distillation

- Mimic behaviour of a larger network by **transferring learned representations** to a smaller one.

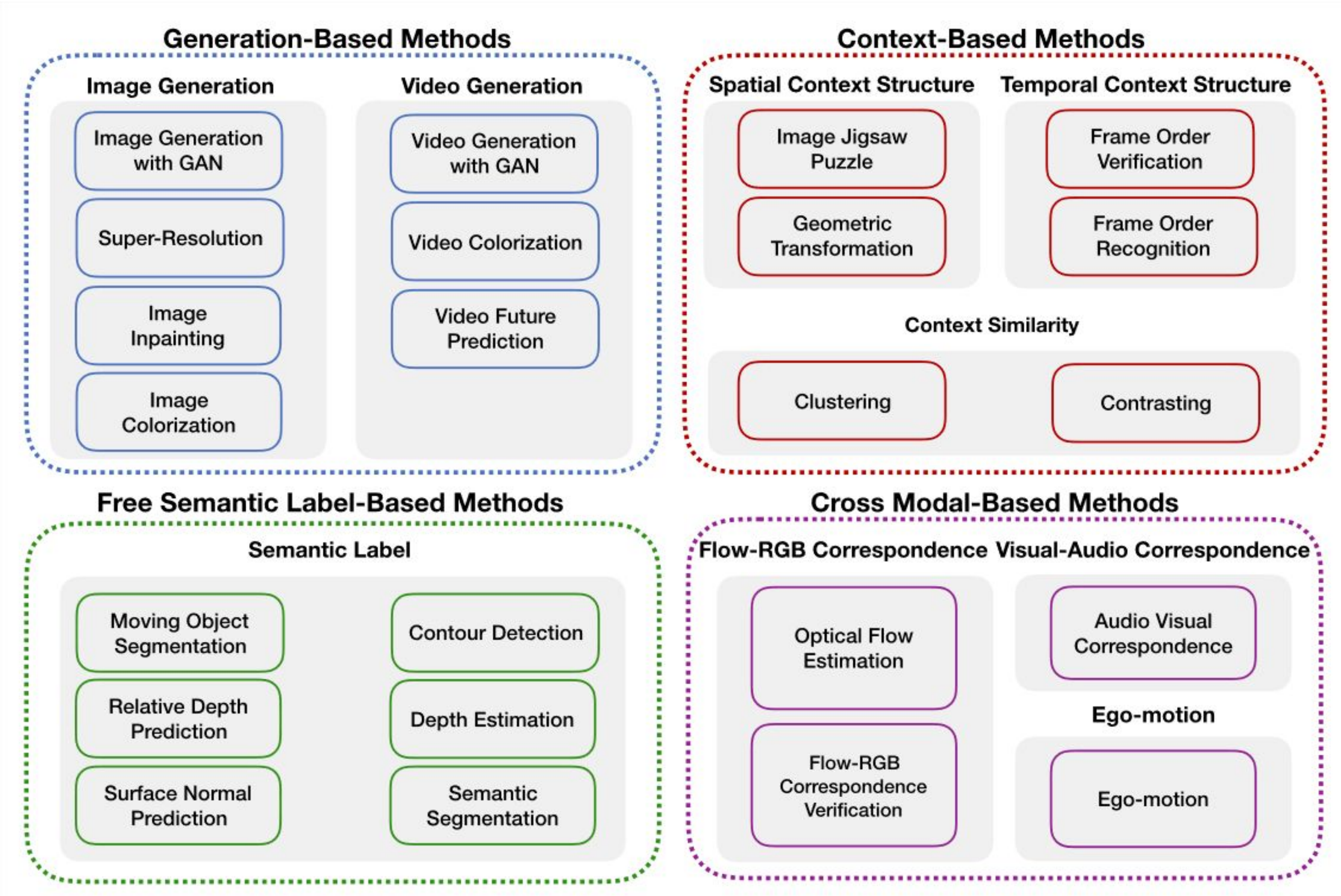
- Train large network.
- Train smaller network on output logits form the first network.



Classic Self-supervised Learning for Video

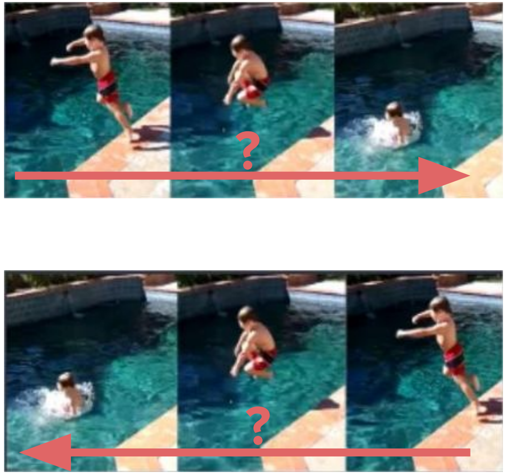
⚠️ Disclaimer! Multi-modality not considered

Classic SSL for Video

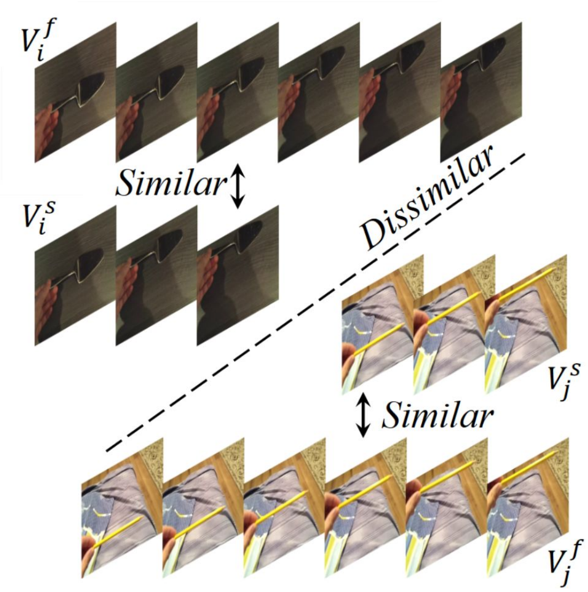


L. Jing & Y. Tian "Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey", TPAMI (Nov. 2021)

Temporal Consistency



Arrow of time

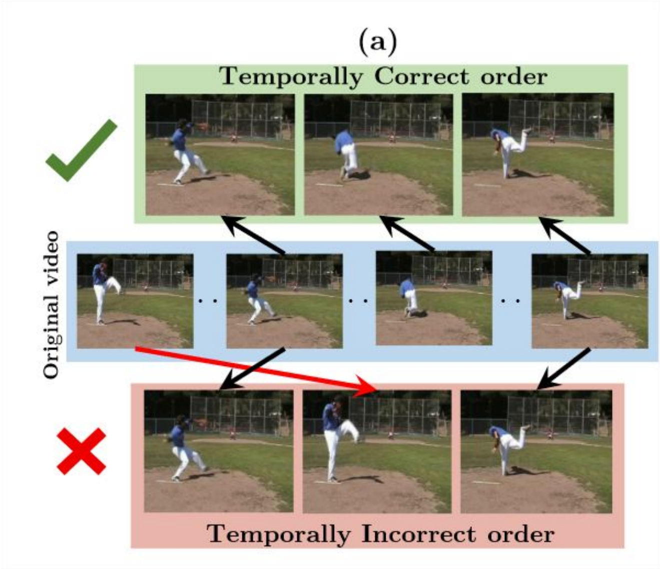


Tempo consistency

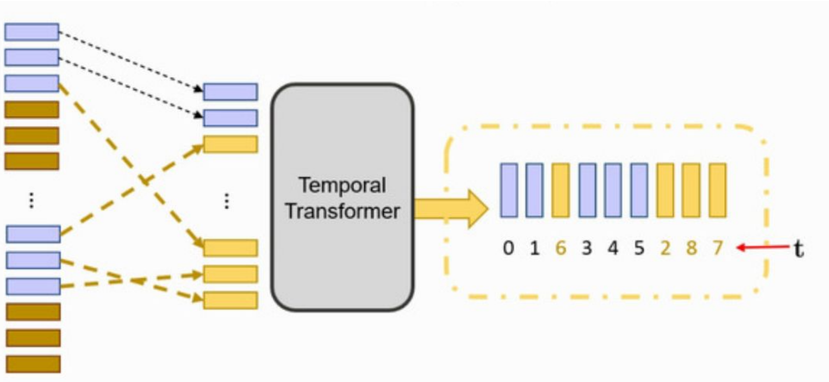
Cubic puzzles

Yang et al. "Video representation learning with visual tempo consistency." arXiv preprint (2020).
Kim et al. “ Self-supervised video representation learning with space-time cubic puzzles.”, AAAI (2021)

Jigsaw Puzzle



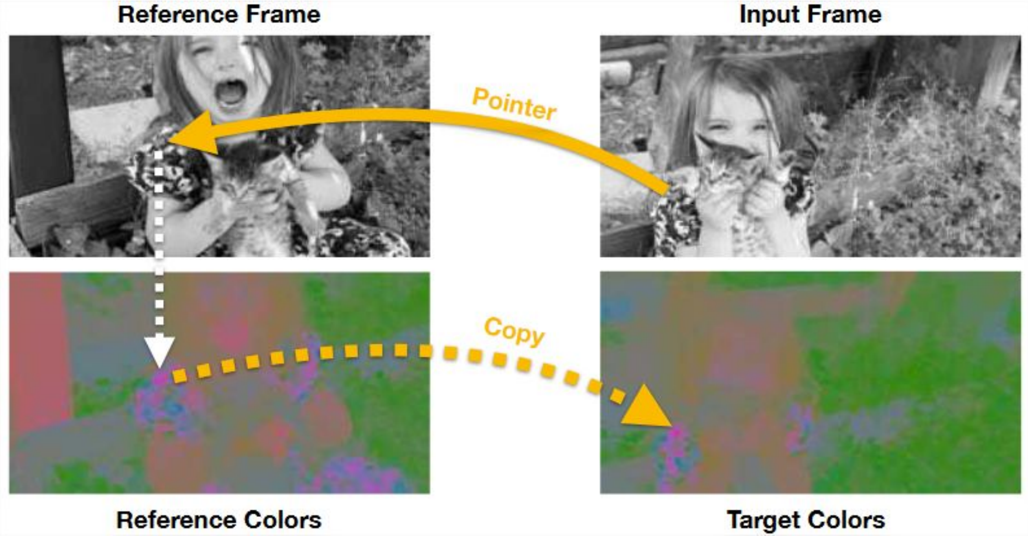
Binary

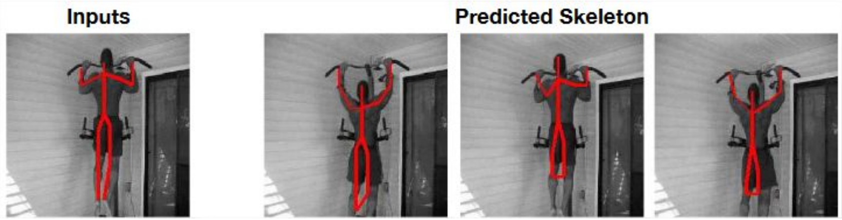


Multi-class

Misra et al. “Shuffle and learn: unsupervised learning using temporal order verification”, ECCV (2016)
Li et al. “HERO: Hierarchical Encoder for Video+ Language Omni-representation Pre-training”. In EMNLP (2020).

Colorization

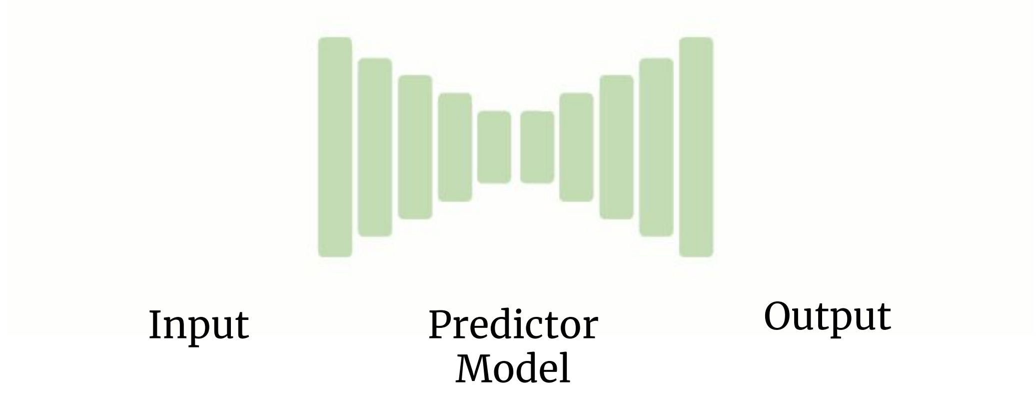




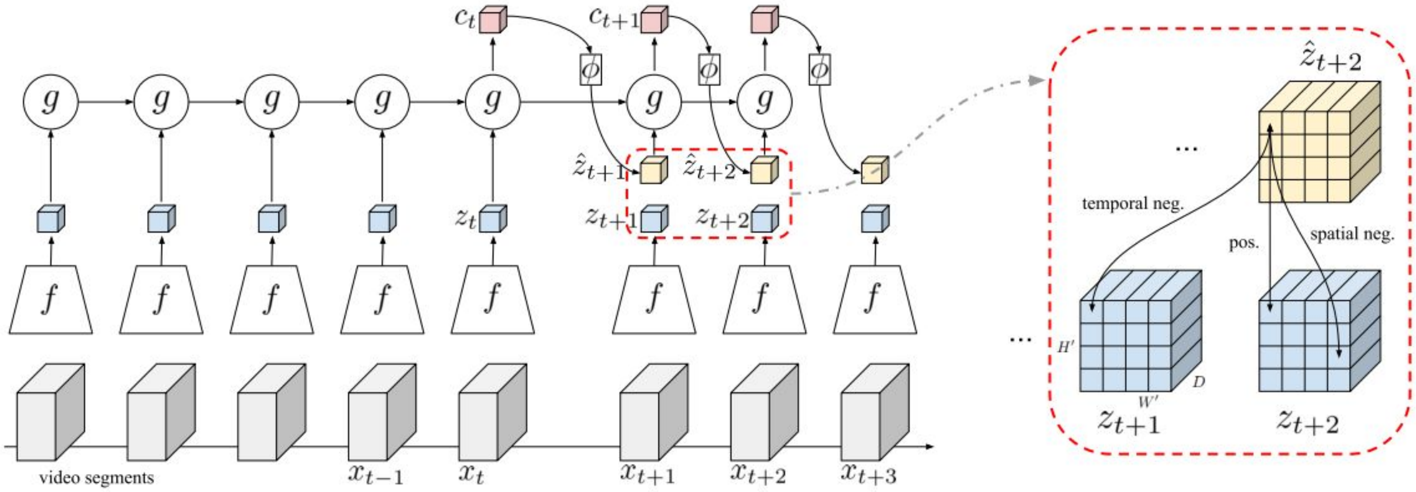
Vondrick et al. “Tracking Emerges by Colorizing Videos”, ECCV (2018).

Video Prediction

Given a video sequence, generate the next frames.



Dense Predictive Coding



Predictive Coding: anticipate the future clip representations given a context of multiple clips.

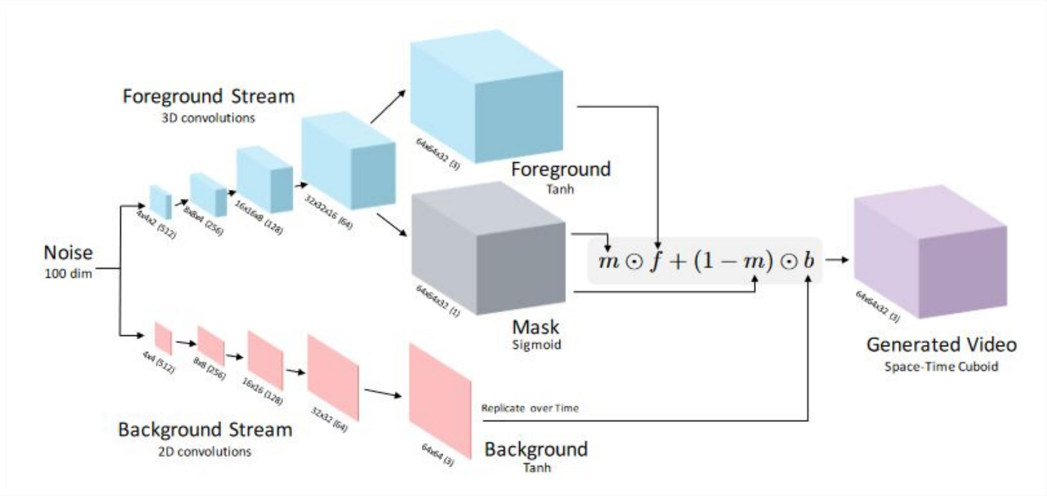
Sample negatives from:

- ➢ other videos
- ➢ same video on different time
- ➢ same time on different positions.

$$\mathcal{L} = -\sum_{i,k}\left[\log \frac{\exp(\hat{z}_{i,k}^{\top} \cdot z_{i,k})}{\sum_{j,m} \exp(\hat{z}_{i,k}^{\top} \cdot z_{j,m})}\right]$$

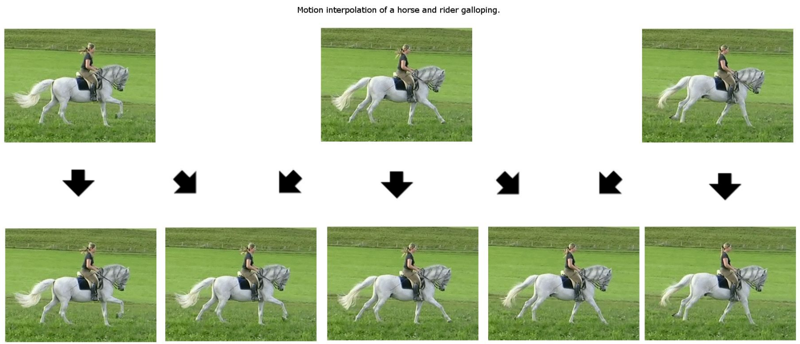
Han et al. “Video Representation Learning by Dense Predictive Coding”. In ICCVW (2019).

Generative Video Models

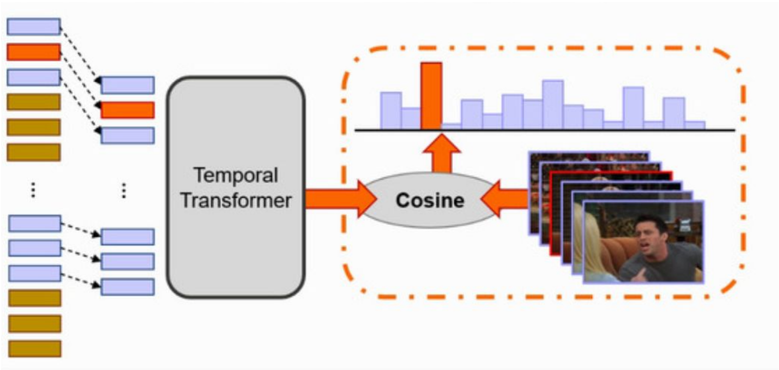


Vondrick et al. “Generating Videos with Scene Dynamics”, NeurIPS (2016).

Interpolation



Generative



Contrastive

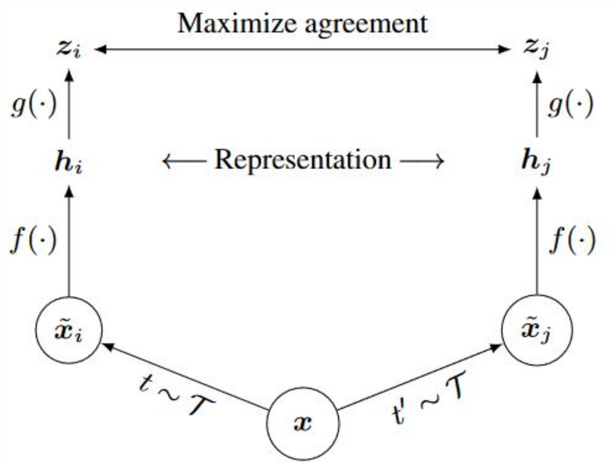
Li et al. “HERO: Hierarchical Encoder for Video+ Language Omni-representation Pre-training”. In EMNLP (2020).

Current Trends on Visual SSL

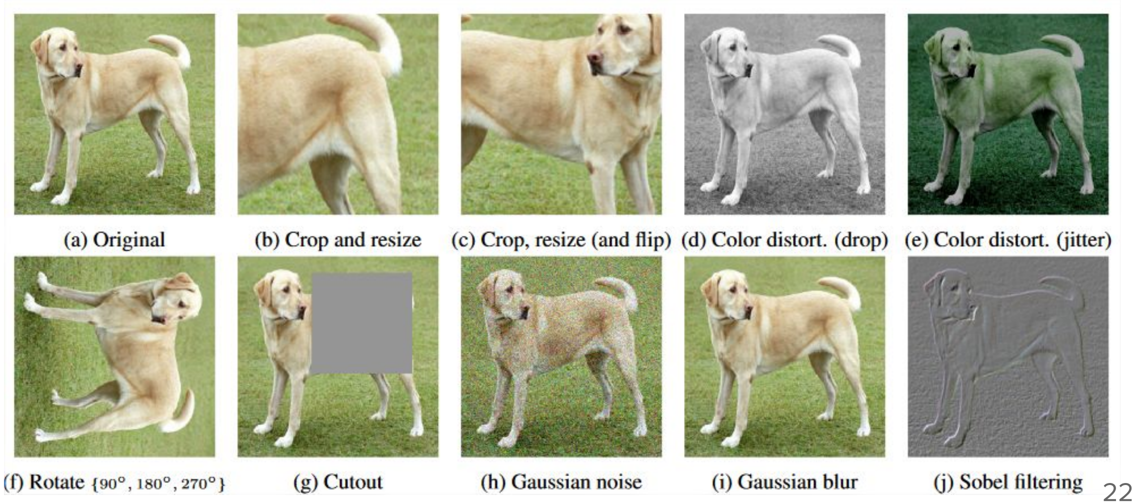
⚠️ Disclaimer! This is not a survey 😉

Contrastive: SimCLR

➢ Siamese setting.
➢ Data augmentation.
➢ Contrastive loss.
➢ Negatives from batch.



$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(z_i, z_k)/\tau)}$$



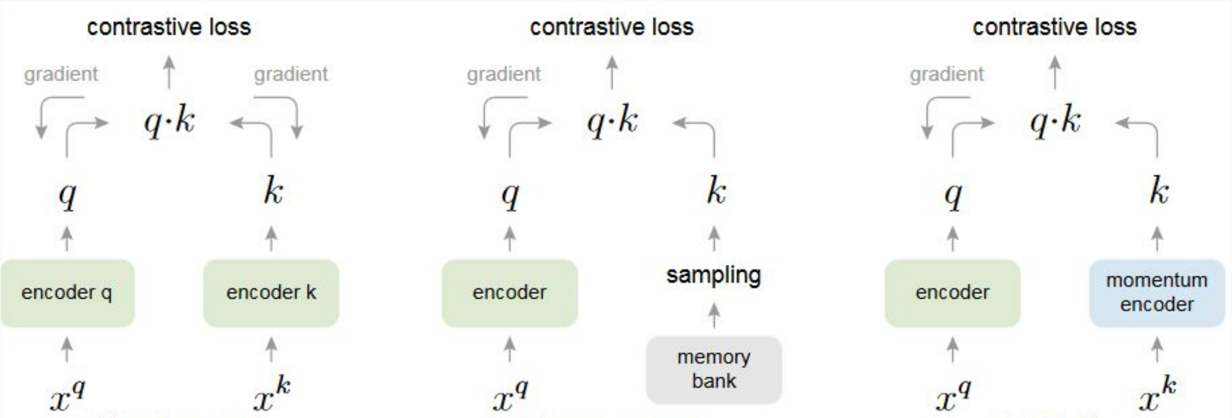
Chen et al. “A Simple Framework for Contrastive Learning of Visual Representations”. In ICML (2020).

Contrastive: MoCo

- ➢ Similar to SimCLR
- ➢ Uses queue to store past batches
- ➢ Draw negatives from queue

- ➢ Unstable training!
- ➢ Use momentum encoder

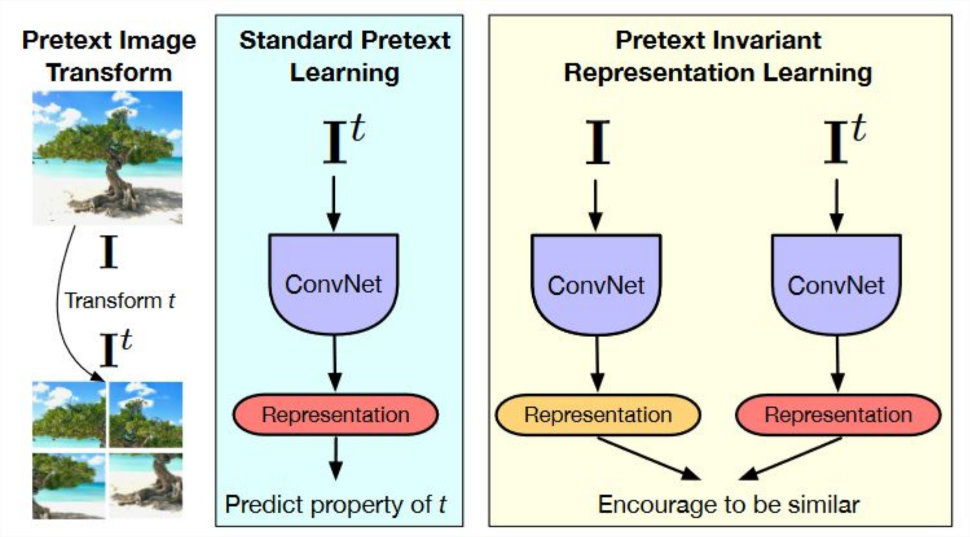
$$\theta_{\mathrm{k}} \leftarrow m\theta_{\mathrm{k}} + (1-m)\theta_{\mathrm{q}}$$

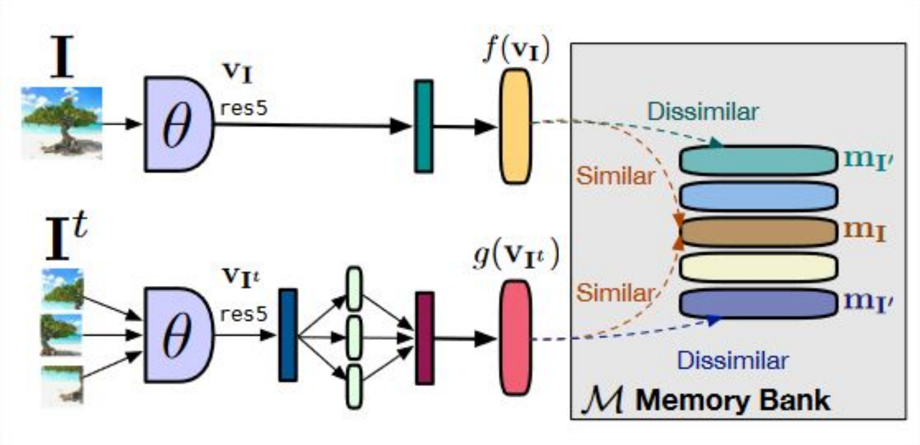


(a) end-to-end (b) memory bank (c) MoCo

He et al. "Momentum Contrast for Unsupervised Visual Representation Learning". In CVPR (2020).
Wu et al. "Unsupervised Feature Learning via Non-Parametric Instance Discrimination", CVPR (2018)

Contrastive: PIRL





Pretext tasks ➔ **covariant** representations

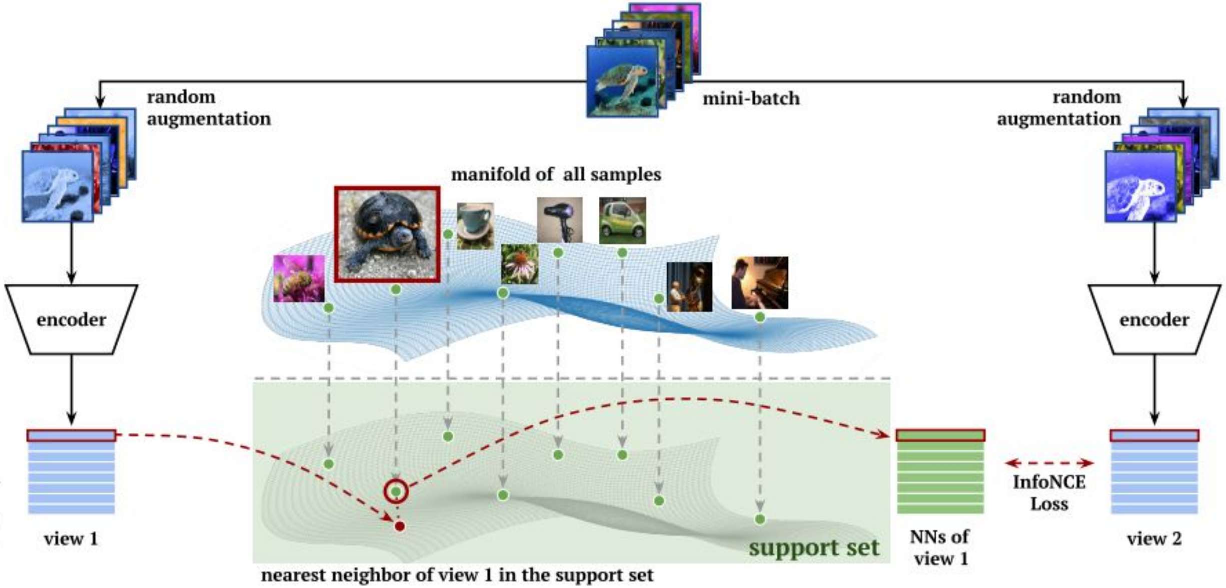
Use **siamese** and **contrastive** to make them **invariant**.

Misra et al. "Self-Supervised Learning of Pretext-Invariant Representations". In CVPR (2020).

Clustering: NNCLR

- **Contrastive** setting
- Negatives from mini-batch
- Positive is a **similar sample** from a **memory** bank

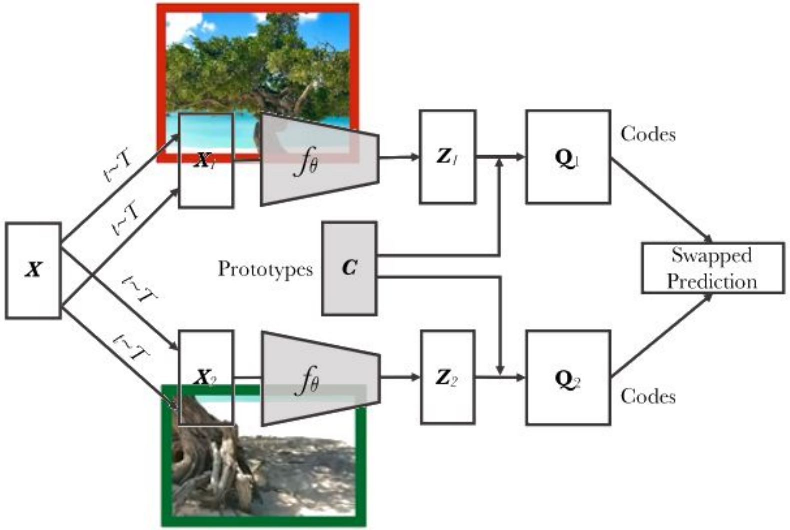
$$\mathcal{L}_i^{\text{NNCLR}} = -\log \frac{\exp\left(\text{NN}(z_i, Q) \cdot z_i^+ / \tau\right)}{\sum_{k=1}^{n} \exp\left(\text{NN}(z_i, Q) \cdot z_k^+ / \tau\right)}$$



Dwibedi et al. “Nearest-Neighbor Contrastive Learning of Visual Representations”. In ICCV (2021).

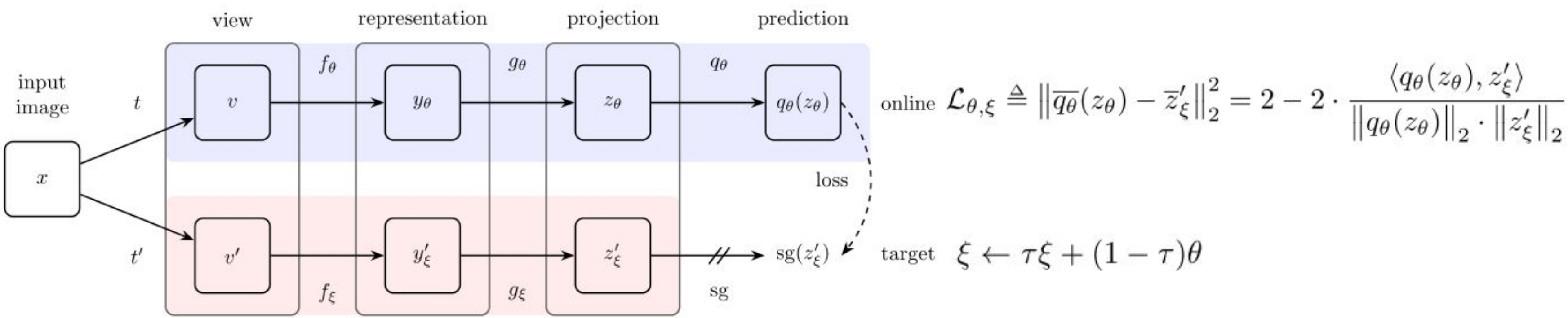
Clustering: SwAv

- ➢ NO MORE NEGATIVES!
- ➢ Learnable prototypes
- ➢ Collapse!
    - ○ Cluster using **Sinkhorn-Knopp**
    - ○ **Uniform sample distribution**
    - ○ **Soft assignment**
- ➢ Swapped prediction.
- ➢ **Multi-crop**



Caron et al. “Unsupervised Learning of Visual Features by Contrasting Cluster Assignments”. In NeurIPS (2020).

Distillation: BYOL



- ➢ Knowledge **distillation** setting
- ➢ **Asymmetric** Siamese architecture
- ➢ Both networks **start from scratch**
- ➢ Teacher is a **moving average**
- ➢ Minimize distance between augmentations

WHY DOES THIS NOT COLLAPSE??

Grill et al. “Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning”. In NeurIPS (2020).

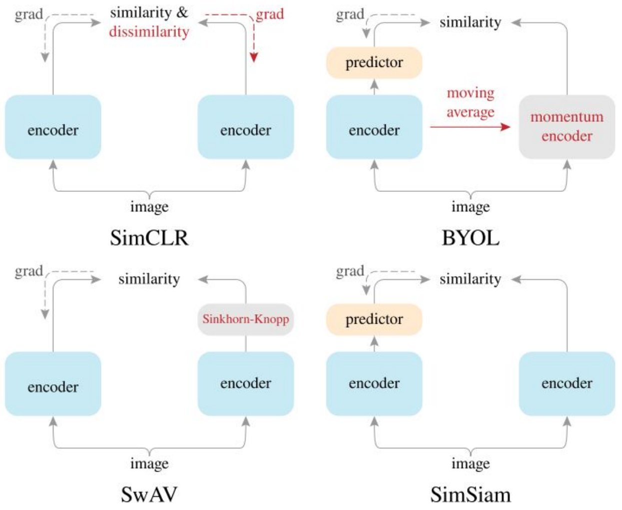
Distillation: SimSiam

Which components are **crucial for avoiding collapse**?

Stop-gradient

Momentum encoder and **predictor** improve accuracy.

- **Expectation Maximization** algorithm:
  - E: the teacher generates various samples (η^t)
  - M: fitting the parameters of the online net to approximate those representations.



$$\mathcal{L}(\theta, \eta) = \mathbb{E}_{x,\mathcal{T}}\Big[\big\|\mathcal{F}_\theta(\mathcal{T}(x)) - \eta_x\big\|_2^2\Big]$$

$$\theta^t \leftarrow \arg\min_\theta \mathcal{L}(\theta, \eta^{t-1})$$

$$\eta^t \leftarrow \arg\min_\eta \mathcal{L}(\theta^t, \eta)$$

$$\eta_x^t \leftarrow \mathcal{F}_{\theta^t}(\mathcal{T}'(x)).$$

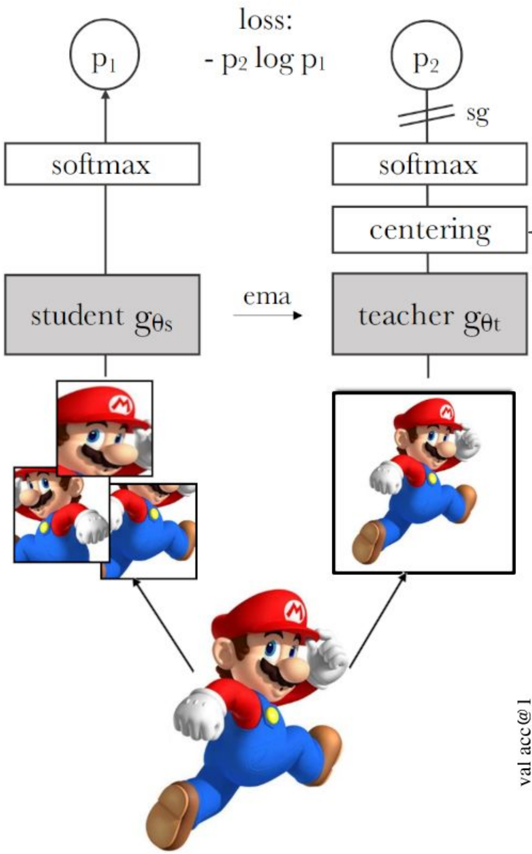
$$\theta^{t+1} \leftarrow \arg\min_\theta \mathbb{E}_{x,\mathcal{T}}\Big[\big\|\mathcal{F}_\theta(\mathcal{T}(x)) - \mathcal{F}_{\theta^t}(\mathcal{T}'(x))\big\|_2^2\Big]$$

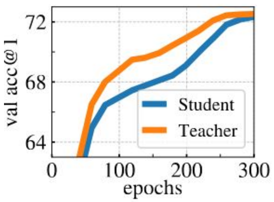
https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

Chen & He “Exploring Simple Siamese Representation Learning”. In CVPR (2021).

Distillation: DINO

- ➢ Instance classification (CE)
- ➢ Multi Crop
- ➢ Transformer Architecture
- ➢ Centering & Sharpenning

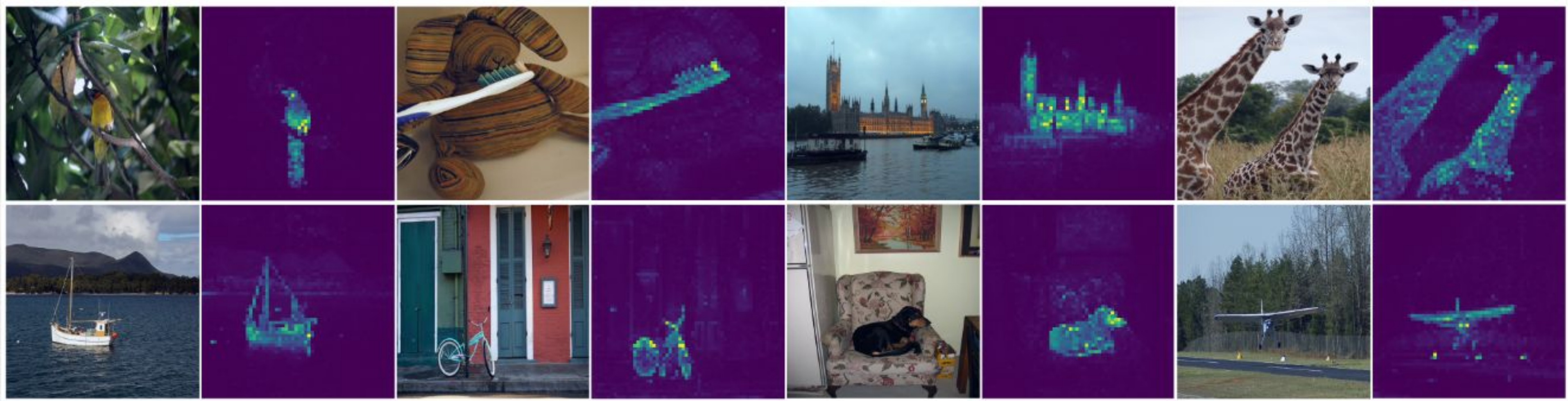




The momentum teacher behaves similarly to a **Polyak-Ruppert averaging**, i.e. a net ensemble. This may explain why the teacher is actually teaching something to the student.

Caron et al. “Emerging Properties in Self-Supervised Vision Transformers”. In ICCV (2021).

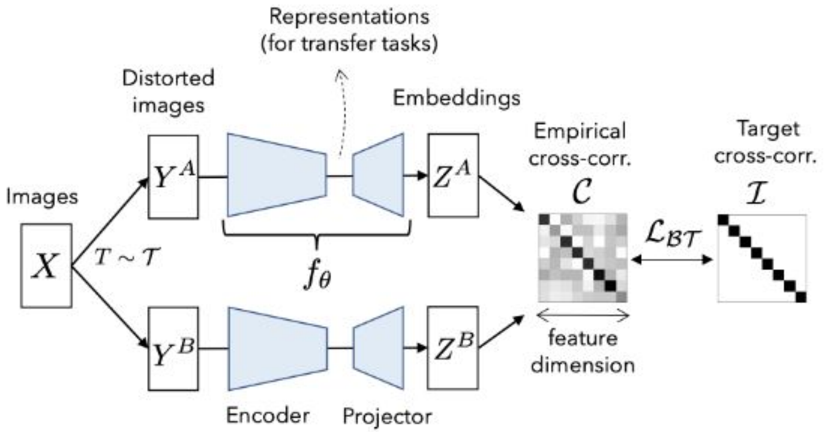
Distillation: DINO



Caron et al. “Emerging Properties in Self-Supervised Vision Transformers”. In ICCV (2021).

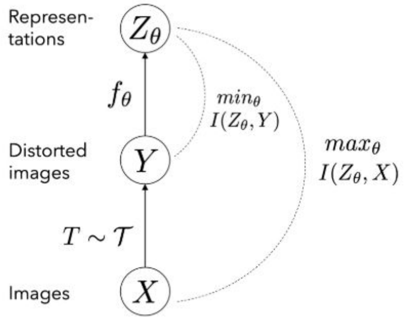
Redundancy Reduction: Barlow Twins

- ➢ Symmetric Siamese
- ➢ Correlation matrix
  - ○ **Invariance** (same feature should be similar in both embeddings)
  - ○ **Reduces redundancy** (no two features should be similar)



$$\mathcal{L}_{\mathcal{BT}} \triangleq \underbrace{\sum_i (1-\mathcal{C}_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j\neq i} {\mathcal{C}_{ij}}^2}_{\text{redundancy reduction term}}$$

Information Bottleneck Principle: Maximize input-output mutual information while being invariant to distortions

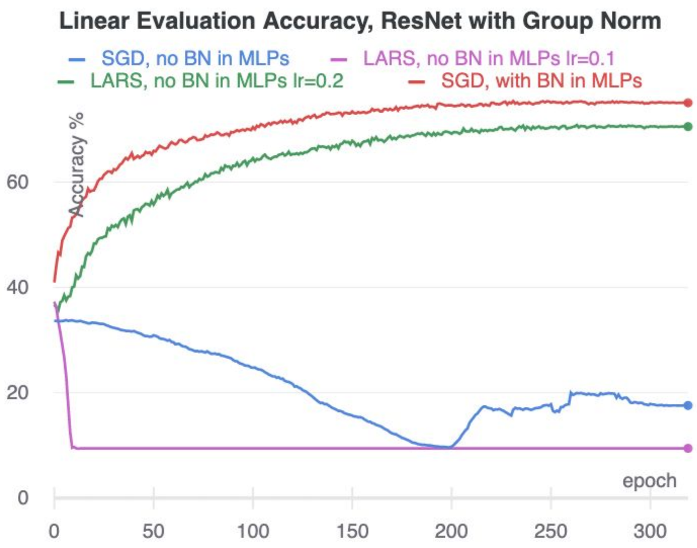


Information Bottleneck Principle

Zbontar et al. “Barlow Twins: Self-Supervised Learning via Redundancy Reduction”. In ICML (2021).

Intuitions

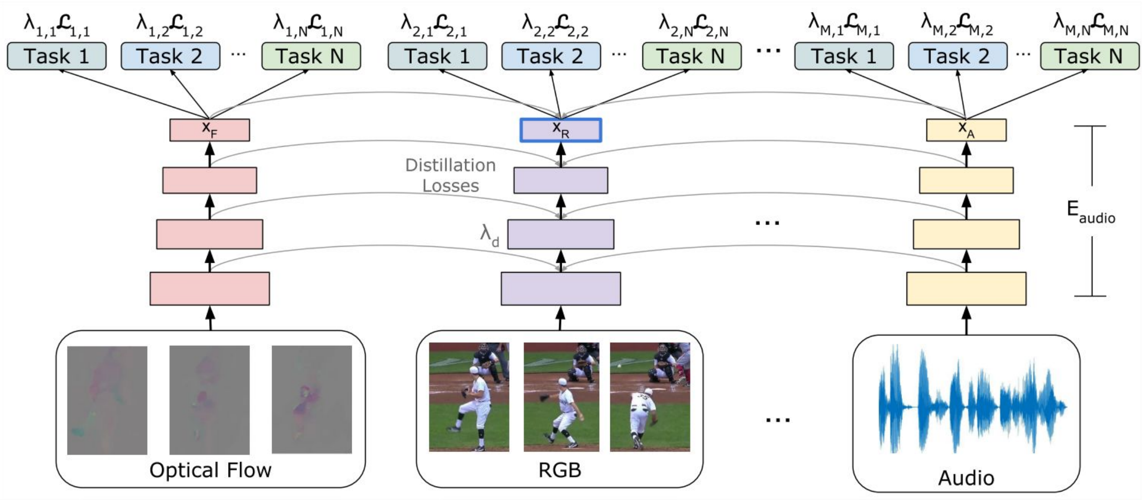
- ➢ Batch Normalization
- ➢ LARS optimizer
- ➢ Asymmetry (projector/predictor with large LR)
- ➢ Momentum Network
- ➢ Weight Decay
- ➢ Stop-gradient

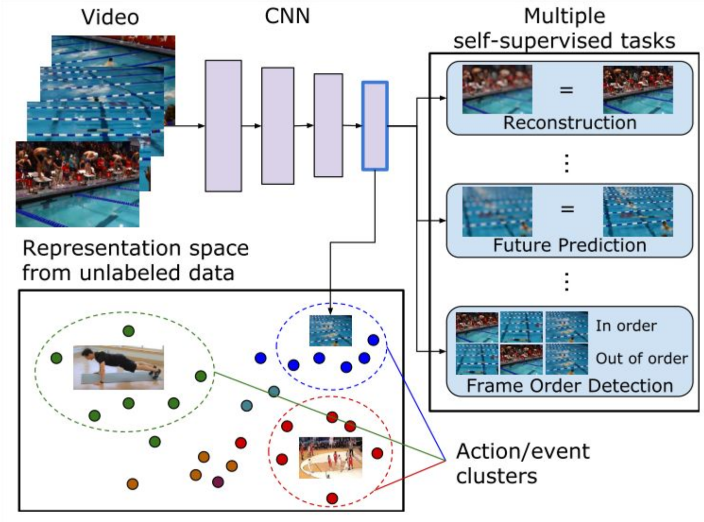


Zbontar et al. “Understanding Self-Supervised and Contrastive Learning with ‘Bootstrap Your Own Latent’”. Blog Post (2021).
Tian et al. “Understanding self-supervised Learning Dynamics without Contrastive Pairs”, ICML (2021)

What about video?

Evolved SSL Losses

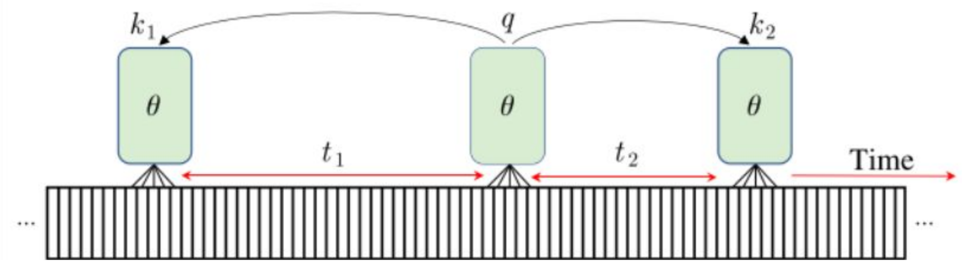


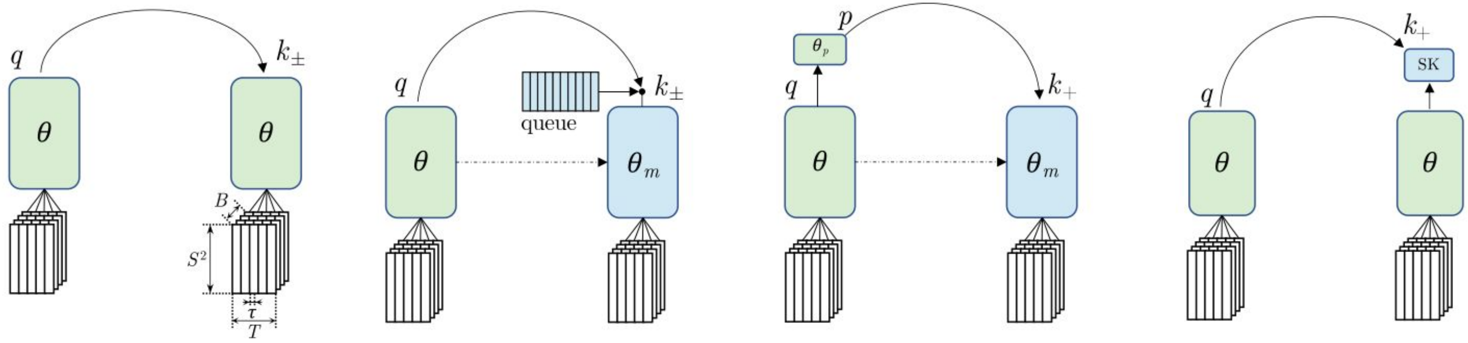


- ➢ **Multiple streams** (flow, grayscale, audio…)
- ➢ **Distilled** at different depths onto the RGB stream.
- ➢ **Multi-task** (reconstruction, prediction, temporal ordering, multi-modal contrastive/alignment).
- ➢ **Evolutionary algorithm** to weight the losses.
- ➢ **Clustered** output forced to follow **Zipf's law** through **KL divergence**

Piergiovanni et al. "Evolving Losses for Unsupervised Video Representation Learning". In CVPR (2020).

Large-scale study



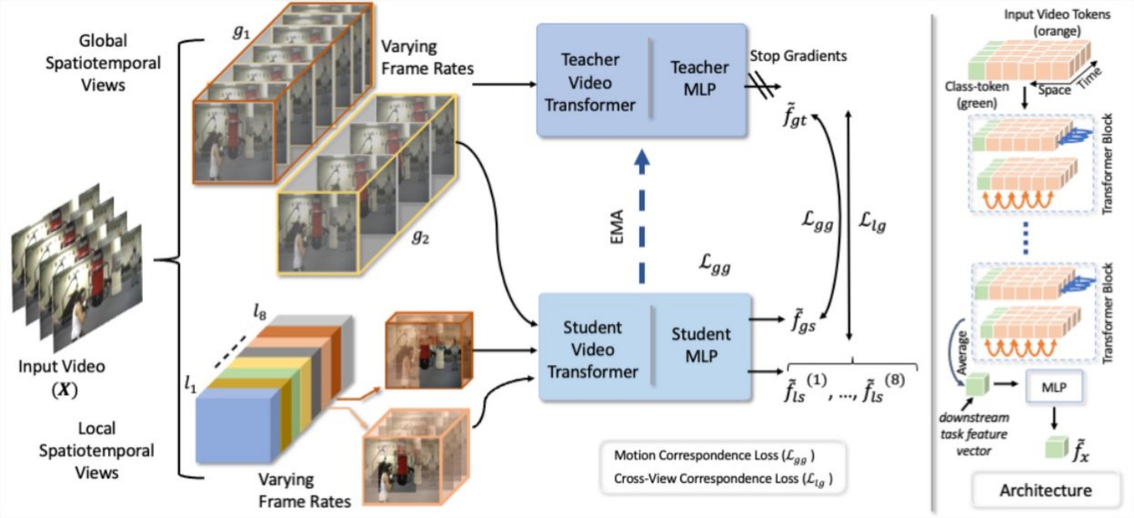


(a) SimCLR (b) MoCo (c) BYOL (d) SwAV

method	pre-train	*linear protocol* K400	*finetuning accuracy* UCF101	AVA (mAP)	Charades (mAP)	SSv2
supervised	*scratch*	**74.7**	68.8	11.7	7.4	48.8
supervised	K400-240K	-	**94.8**	22.2	**34.7**	52.8
SimCLR	K400-240K	62.0 (−12.7)	87.9 (−6.9)	17.6 (−4.6)	11.4 (−23.3)	52.0 (−0.8)
SwAV	K400-240K	62.7 (−11.5)	89.4 (−5.4)	18.2 (−4.0)	10.7 (−24.0)	51.7 (−1.1)
BYOL	K400-240K	68.3 (−6.4)	93.8 (−1.0)	**23.4** (+1.2)	21.0 (−13.7)	**55.8** (+3.0)
MoCo	K400-240K	67.3 (−7.4)	92.8 (−2.0)	20.3 (−1.9)	33.5 (−1.2)	54.4 (+1.8)

Feichtenhofer et al. “A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning”. In CVPR (2021).

Self-supervised Video Transformer



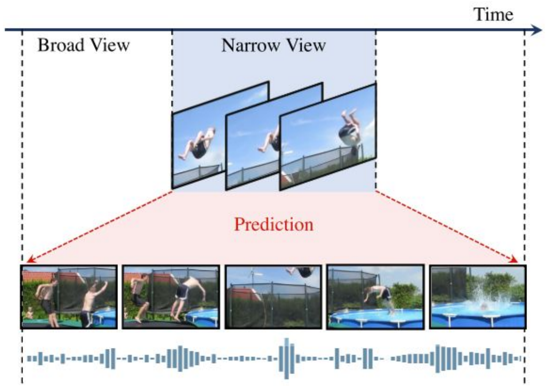
Extension to **DINO** for video. Use of multiple **global views** (varying frame-rate) and multiple **local views**. Matching at global-global (motion) and local-global (cross-view).

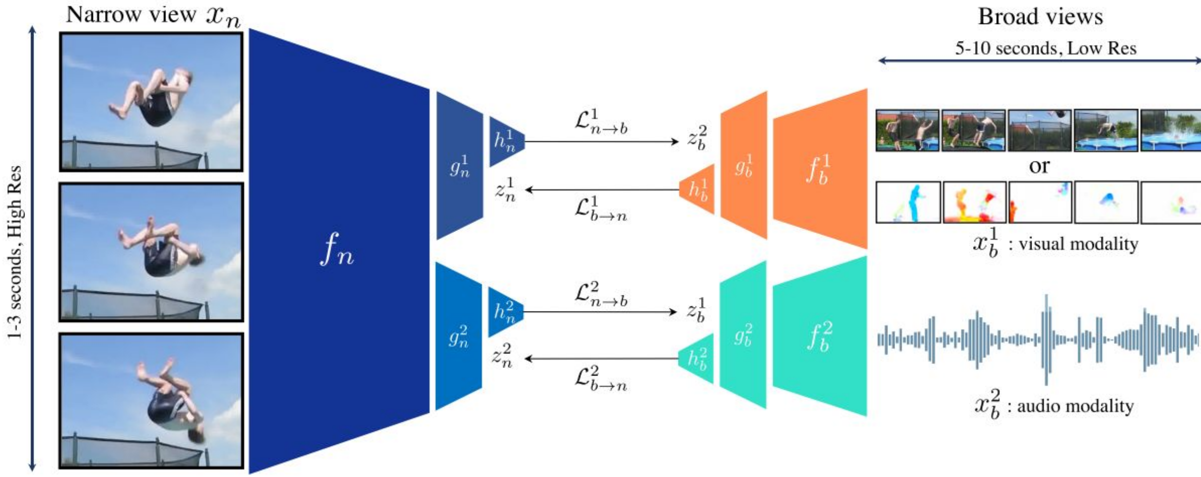
$l \rightarrow g$	$g \rightarrow g$	$l \rightarrow l$	$g \rightarrow l$	UCF-101	HMDB-51
✓	✗	✗	✗	84.11	50.72
✗	✓	✗	✗	81.95	49.04
✓	✓	✗	✗	**84.64**	**52.17**
✓	✓	✓	✗	83.11	51.23
✓	✓	✗	✓	84.71	51.88
✓	✓	✓	✓	83.69	51.71

Ranashinghe et al. "Self-Supervised Video Transformer". In CVPR (2022).

Brave

- ➢ Focus on leveraging **multi-modality**.
- ➢ **Temporal crops** (local and global):
  - ○ Teacher performs global → local prediction.
  - ○ Student performs local → global prediction.
- ➢ Pairwise predictors.
- ➢ Synchronization matters!





Recasens et al. “Broaden Your Views for Self-Supervised Video Learning”. In ICCV (2021).

Thanks!

QUESTIONS?