JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

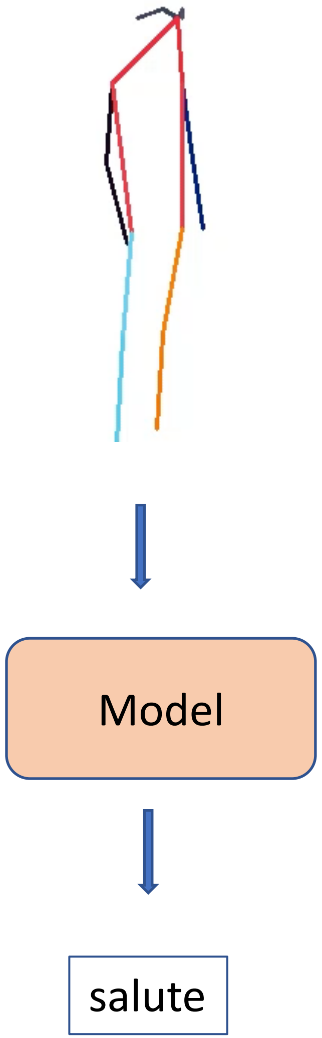
HaLP: Hallucinating Latent Positives for Skeleton-based Self-Supervised Learning of Actions

Anshul Shah[1], Aniket Roy[1*], Ketul Shah[1*], Shlok Mishra[2], David Jacobs[23], Anoop Cherian[4], Rama Chellappa[1]

[1]Johns Hopkins University
[2]University of Maryland, College Park
[3]Meta AI
[4]Mitsubishi Electric Research Laboratories
*Equal contribution

THU-AM-226

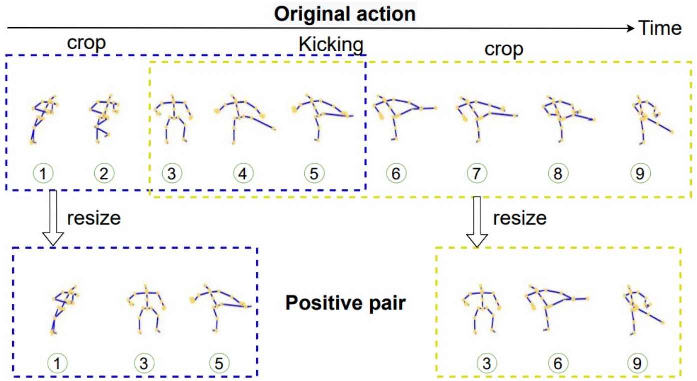
Skeleton-based action recognition

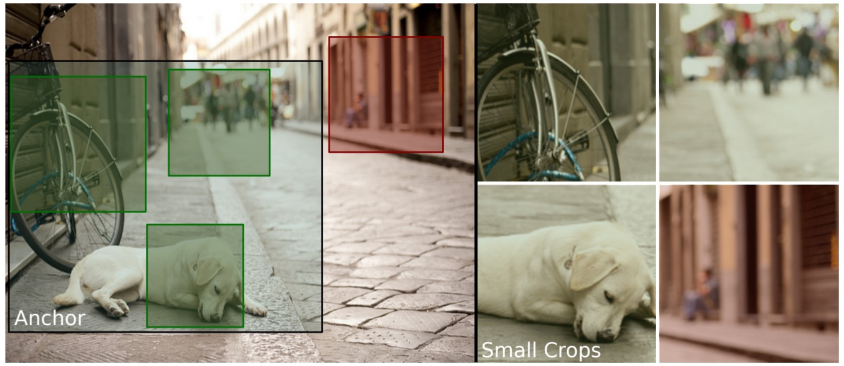


Skeletons vs RGB

- ✓ Convey the action succinctly
- ✓ Reduce the impact of scene and object biases
- ✓ Reduced privacy concerns

Key motivation of our approach



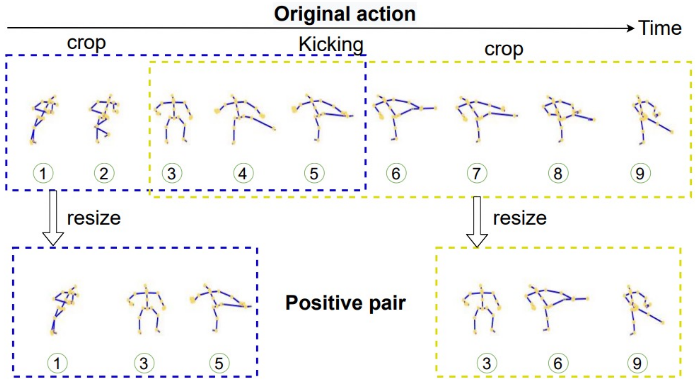


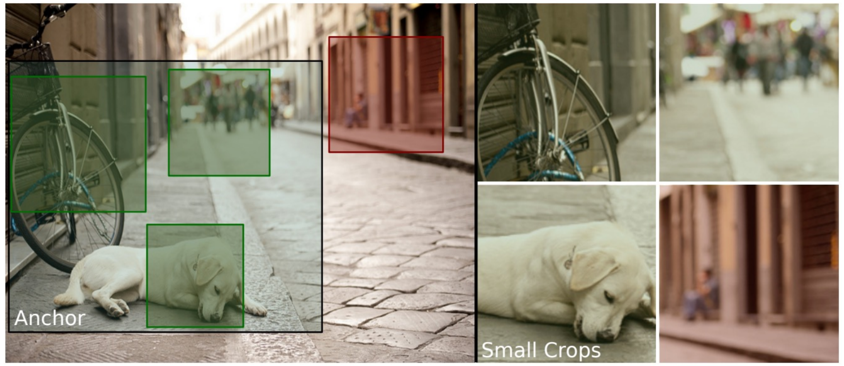
- Data augmentations play a key role in contrastive learning
    - Diversity and strength of augmentation
    - Multi-view / Multi-crop strategy is shown to be helpful
- Crafting plausible augmentations for skeletons is challenging

Skeleton-Contrastive 3D Action Representation Learning, Thoker et al. 2021.

Revisiting Contrastive Methods for Unsupervised Learning of Visual Representations, Van Gansbeke et al. 2021

Key motivation of our approach





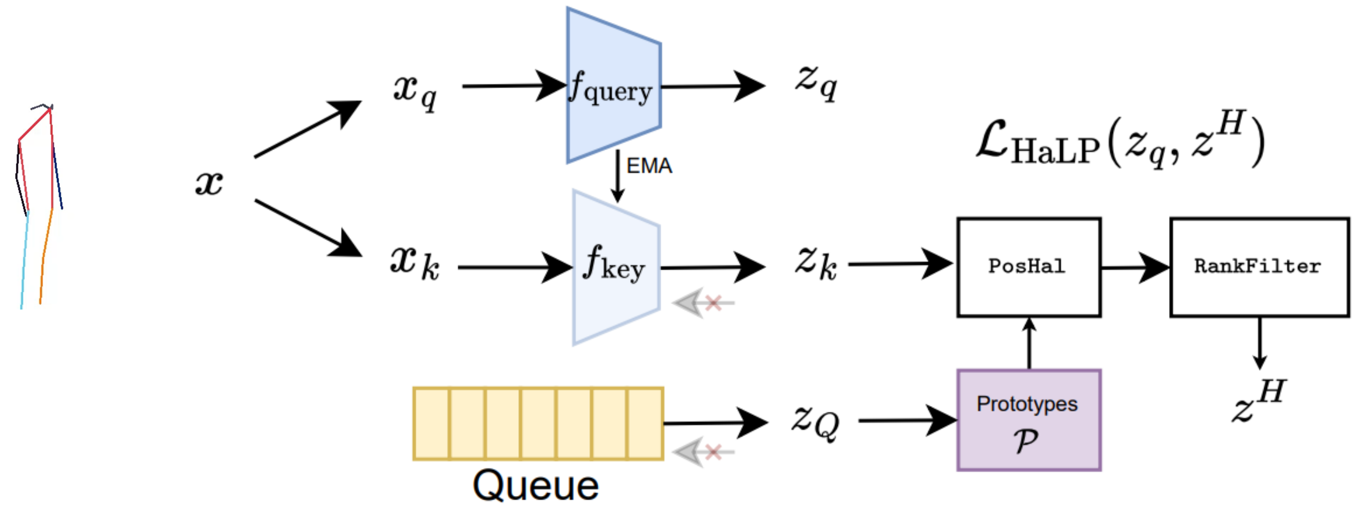
- Data augmentations play a key role in contrastive learning
  - Diversity and strength of augmentation
  - Multi-view / Multi-crop strategy is shown to be helpful
- Crafting plausible augmentations for skeletons is challenging

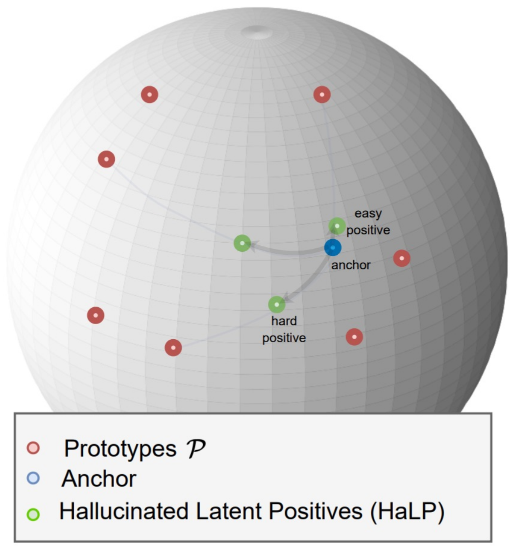
Can we hallucinate positives in the latent space ?

Skeleton-Contrastive 3D Action Representation Learning, Thoker et al. 2021.

Revisiting Contrastive Methods for Unsupervised Learning of Visual Representations, Van Gansbeke et al. 2021

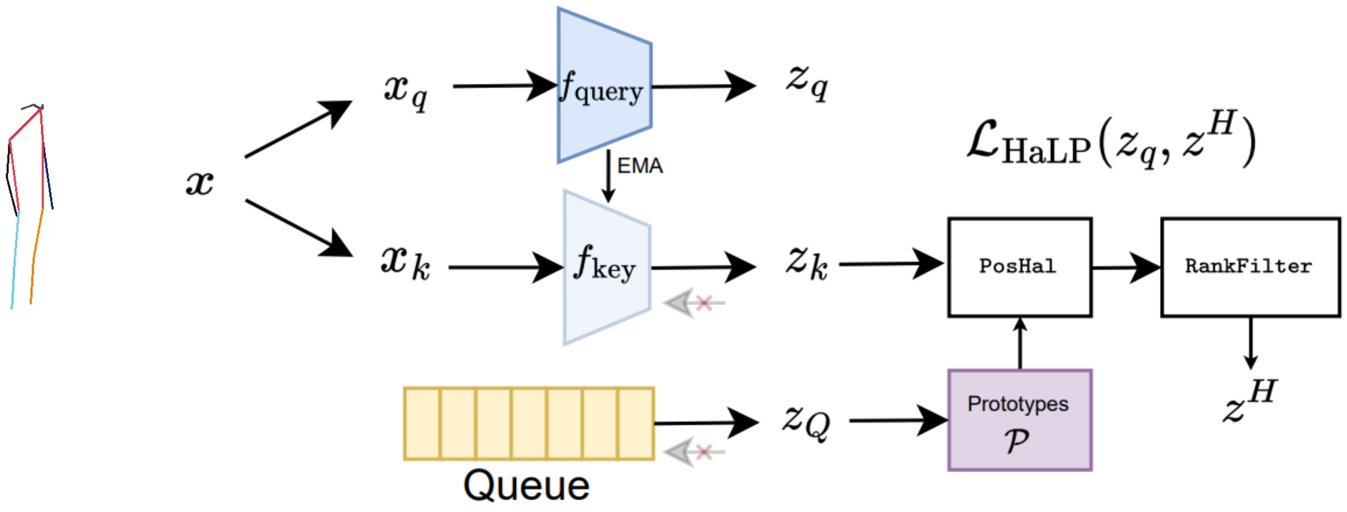
Hallucinate new positives in the input space

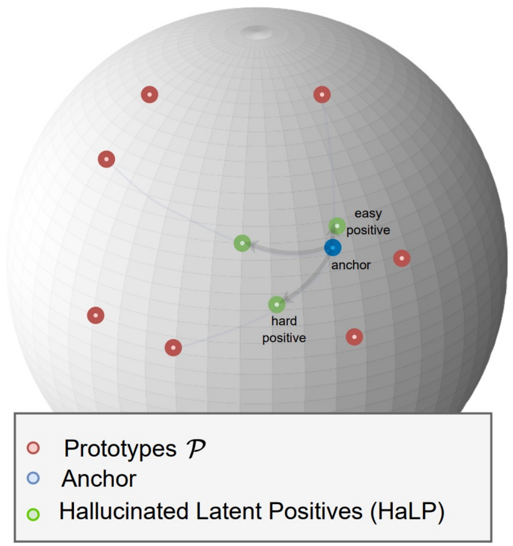




- We propose an objective function which can be used to generate positives of varying level of hardness
- Relaxations to the objective allow for closed form making the process very fast
- Final solution involves spherical linear interpolation of the anchor with a randomly chosen data prototype

Hallucinate new positives in the input space





- We propose an objective function which can be used to generate positives of varying level of hardness
- Relaxations to the objective allow for closed form making the process very fast
- Final solution involves spherical linear interpolation of the anchor with a randomly chosen data prototype

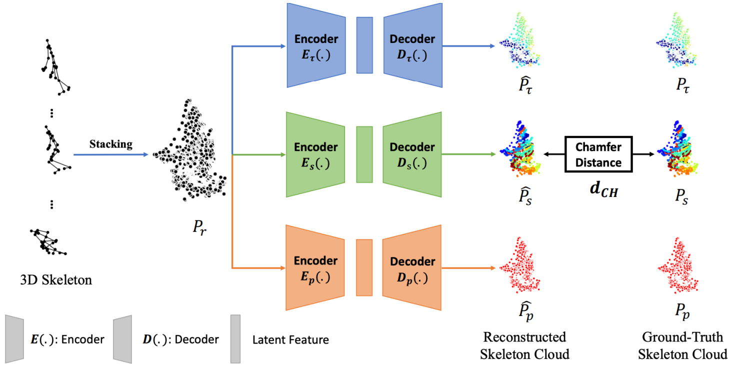
Why skeleton-based action recognition ?

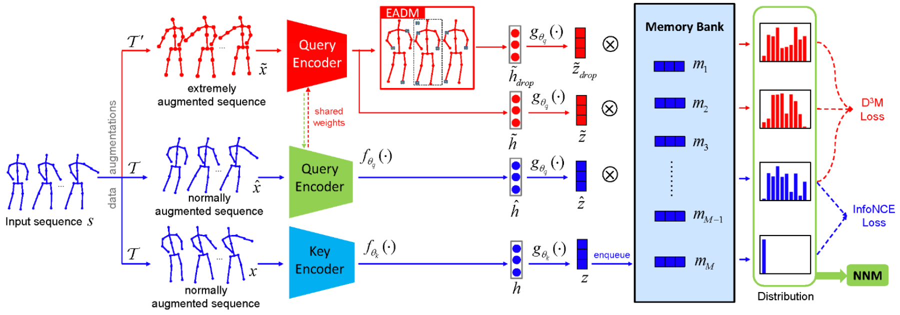
An example from Johansson's experiment

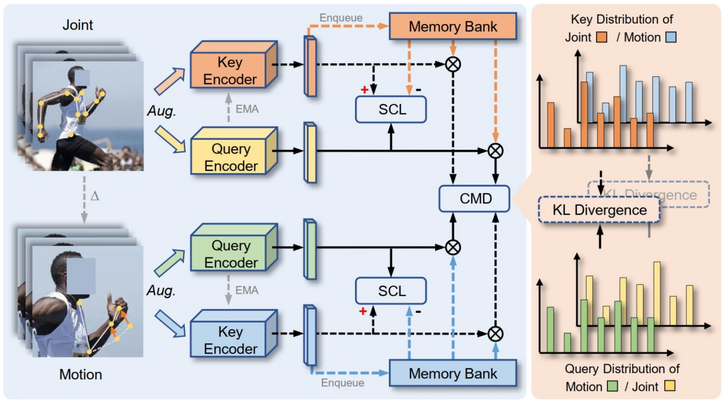
- Convey the action succinctly
- Reduce the impact of scene and object biases
- Reduced privacy concerns

Visual perception of biological motion and a model for its analysis, Johansson, Gunnar. 1973

Self-supervised skeleton-based action recognition







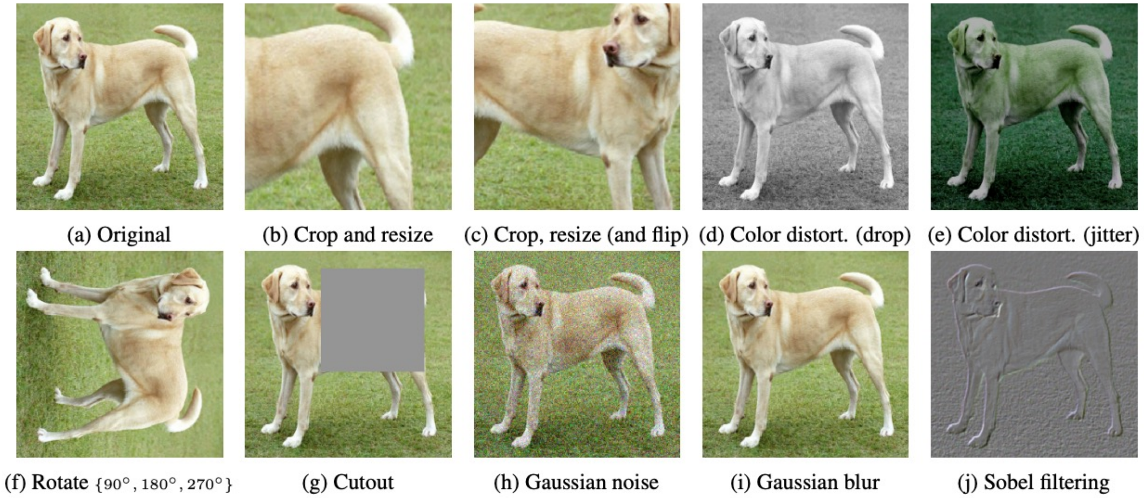
- Various pretext tasks proposed in the past : Skeleton coloring, masked modeling, contrastive learning
- Other research directions : encoders, augmentations, additional modalities

Skeleton cloud colorization for unsupervised 3d action representation learning, Yang et al. 2021

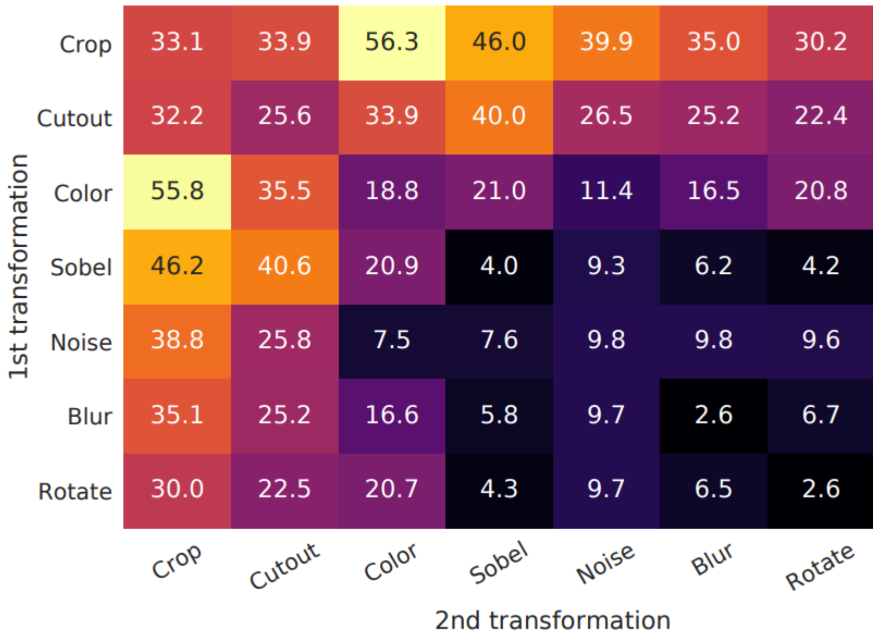
Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition, Guo et al. 2022

CMD: Self-supervised 3D Action Representation Learning with Cross-Modal Mutual Distillation, Mao et al. 2022

Data augmentations are critical



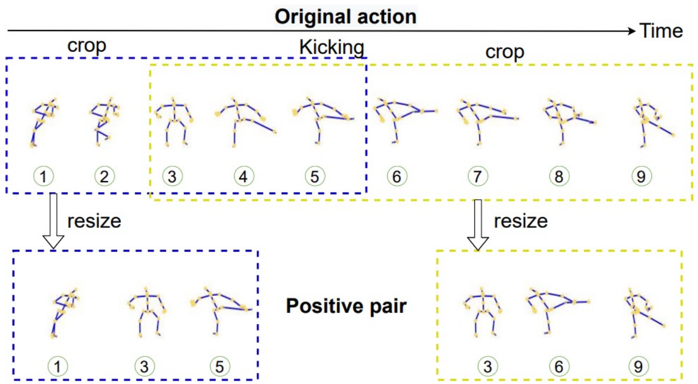
Examples of image data augmentations

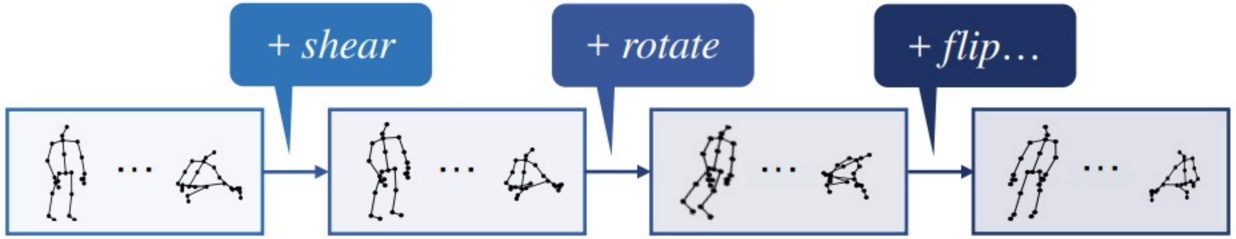


Composition of transformations is crucial

A simple framework for contrastive learning of visual representations, Chen et al. 2020

Augmentations for skeletons is hard



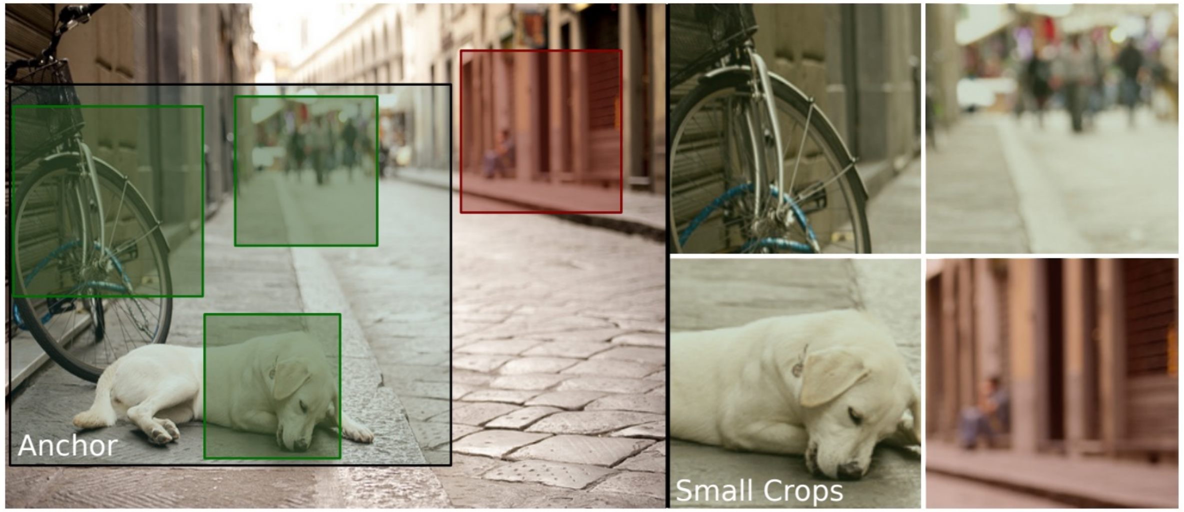


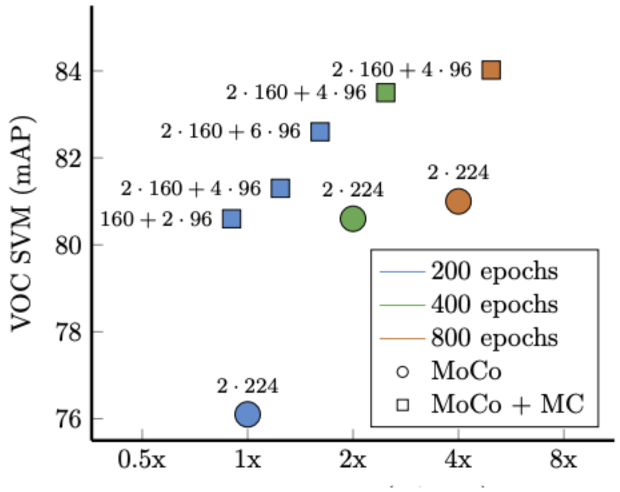
- Data augmentations require domain knowledge
- Crafting plausible augmentations for skeletons is challenging

Skeleton-Contrastive 3D Action Representation Learning, Thoker et al. 2021

Hierarchical Consistent Contrastive Learning for Skeleton-Based Action Recognition with Growing Augmentations, Zhang et al. 2022

Multi-view/Multi-crop strategy is helpful



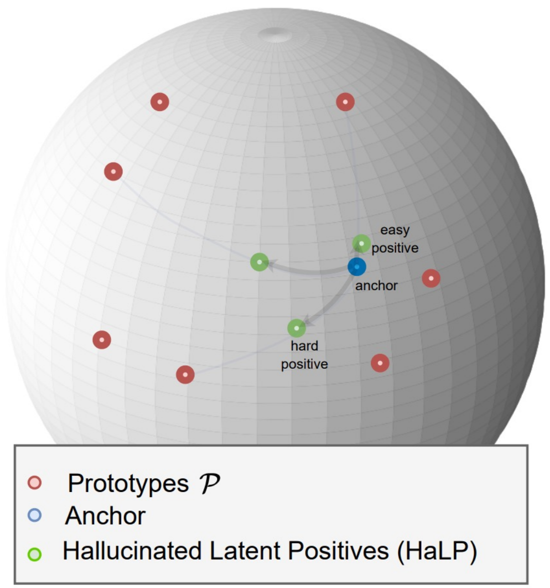


- Multi-view has been shown to be helpful but is expensive to train
- Difficulty in designing data augmentations for skeletons makes multiview more challenging

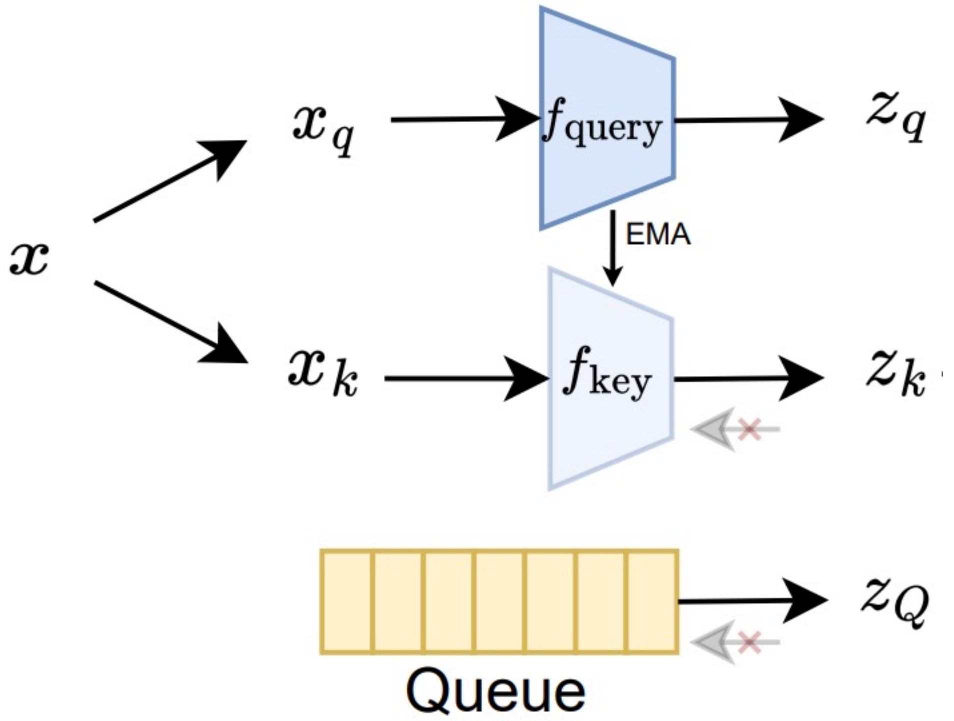
Revisiting Contrastive Methods for Unsupervised Learning of Visual Representations, Van Gansbeke et al. 2021

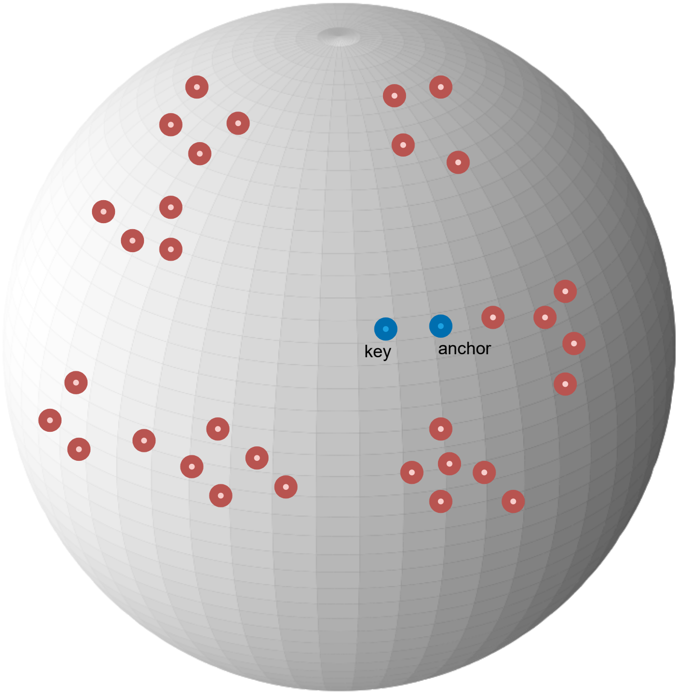
Hallucinating latent positives

- Does not require hand crafting new augmentations
- Generating multiple views is easy and inexpensive
- Can control for hardness and diversity



Our approach





Momentum contrast for unsupervised visual representation learning, He et al 2020

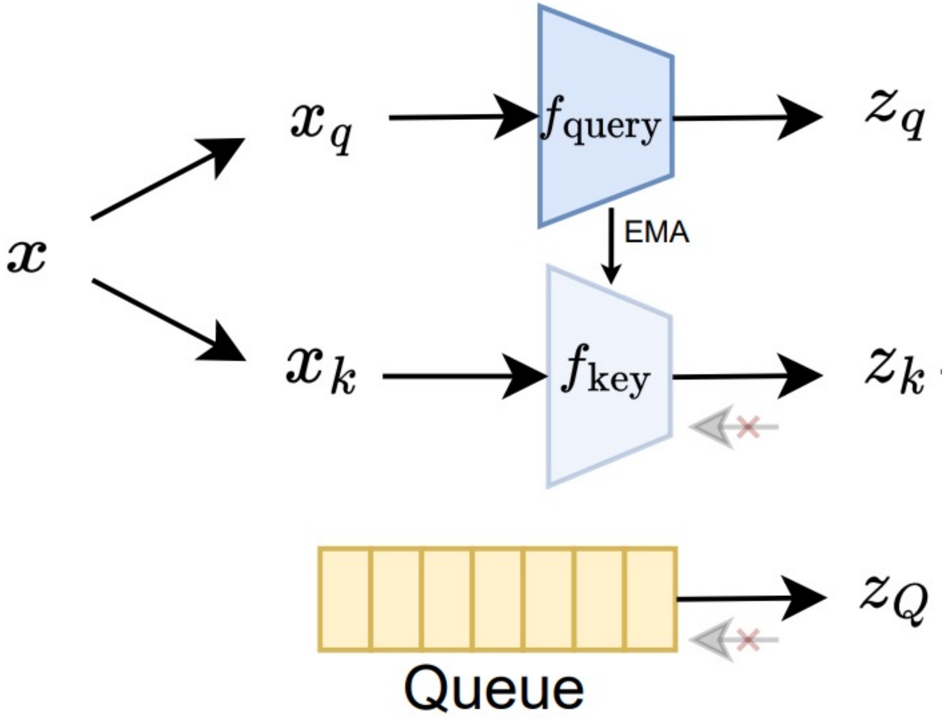
Our approach

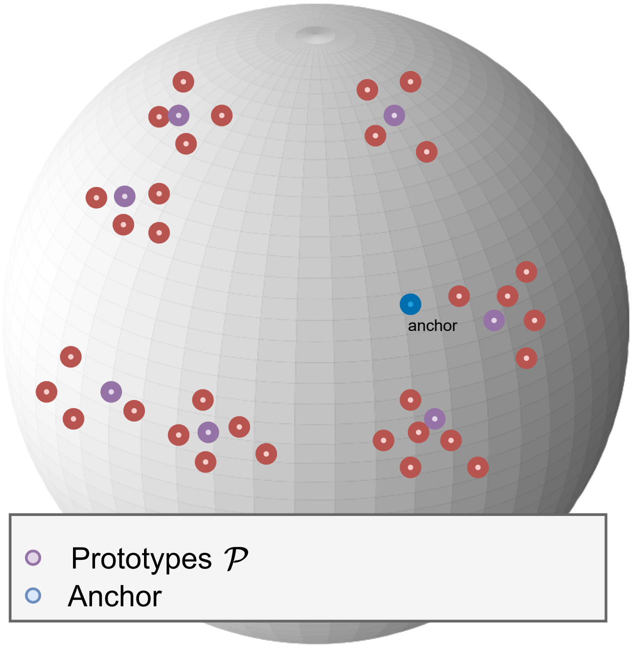
- Desiderata:
  - We want to generate positives of varying hardness which lie far from anchor positives
  - Have the same underlying class semantics

- Key intuition : We can explore the high dimensional space around the anchors to find locations that can be plausibly reached by the encoder

Our approach

- Clustering on hypersphere to extract prototypes
- Use key as an anchor





Our approach

- Formally, we define our objective as

$$
\begin{aligned}
& z^* = \arg\min_{z \in \mathbb{S}^{D-1}} \ \mathrm{sim}(z, P^*_{z_k}) \\
& \text{s.t.} \quad \mathrm{sim}(z, P^*_{z_k}) \geq \mathrm{sim}(z, P), \forall P \in \mathcal{P} \setminus \left\{ P^*_{z_k} \right\}
\end{aligned}
$$

Our approach

- Formally, we define our objective as

$$
\begin{aligned}
z^* &= \arg \min_{z \in \mathbb{S}^{D-1}} \operatorname{sim}(z, \boxed{P^*_{z_k}}) \\
\text{s.t.} \quad & \operatorname{sim}(z, P^*_{z_k}) \geq \operatorname{sim}(z, P), \forall P \in \mathcal{P} \setminus \{P^*_{z_k}\}
\end{aligned}
$$

Our approach

- Formally, we define our objective as

$$z^* = \arg\min_{z \in \mathbb{S}^{D-1}} \mathrm{sim}(z, P^*_{z_k})$$
$$\text{s.t.} \quad \mathrm{sim}(z, P^*_{z_k}) \geq \mathrm{sim}(z, P), \forall P \boxed{\in \mathcal{P} \setminus} \{P^*_{z_k}\}$$

Our approach

- Formally, we define our objective as

$$\boxed{z^* = \arg \min_{z \in \mathbb{S}^{D-1}} \ \mathrm{sim}(z, P^*_{z_k})}$$
$$\text{s.t.} \quad \mathrm{sim}(z, P^*_{z_k}) \geq \mathrm{sim}(z, P), \forall P \in \mathcal{P} \setminus \{P^*_{z_k}\}$$

- We want to generate hard positives which are far from the anchor but have the same closest prototype as the anchor.

Our approach

- Formally, we define our objective as

$$z^* = \arg \min_{z \in \mathbb{S}^{D-1}} \mathrm{sim}(z, P^*_{z_k})$$
$$\boxed{\text{s.t.} \quad \mathrm{sim}(z, P^*_{z_k}) \geq \mathrm{sim}(z, P), \forall P \in \mathcal{P} \setminus \{P^*_{z_k}\}}$$

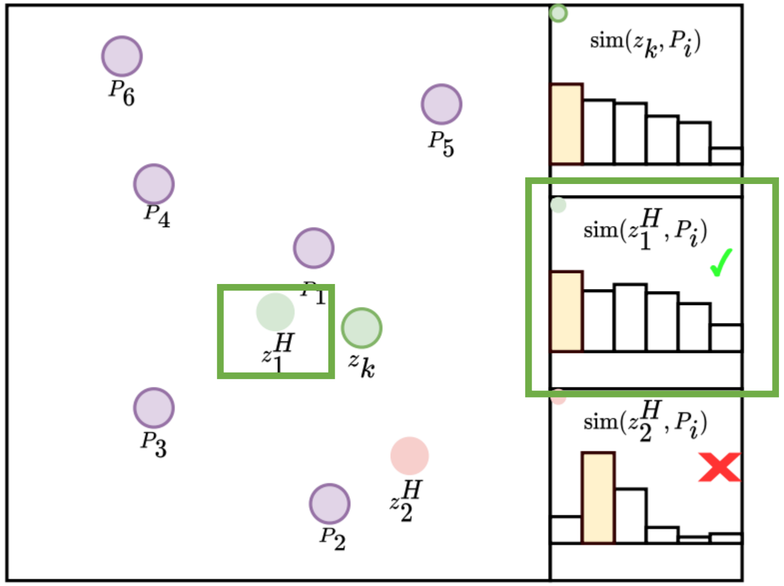
- We want to generate hard positives which are far from the anchor but have the same closest prototype as the anchor.

Our approach

- Formally, we define our objective as

$$z^* = \arg \min_{z \in \mathbb{S}^{D-1}} \ \mathrm{sim}(z, P^*_{z_k})$$
$$\text{s.t.} \quad \mathrm{sim}(z, P^*_{z_k}) \geq \mathrm{sim}(z, P), \forall P \in \mathcal{P} \setminus \{P^*_{z_k}\}$$

- We want to generate hard positives which are far from the anchor but have the same closest prototype as the anchor.

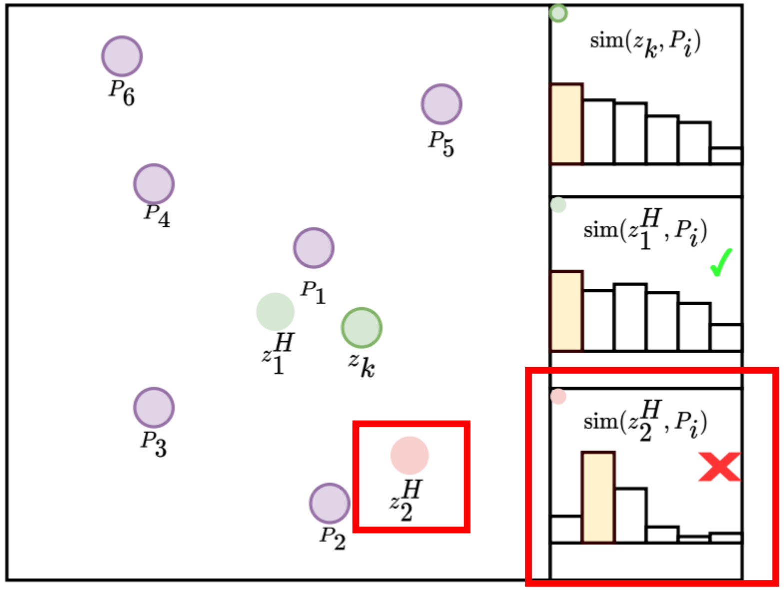


Our approach

- Formally, we define our objective as

$$z^* = \arg\min_{z \in \mathbb{S}^{D-1}} \operatorname{sim}(z, P^*_{z_k})$$
$$\text{s.t.} \quad \operatorname{sim}(z, P^*_{z_k}) \geq \operatorname{sim}(z, P), \forall P \in \mathcal{P} \setminus \{P^*_{z_k}\}$$

- We want to generate hard positives which are far from the anchor but have the same closest prototype as the anchor.

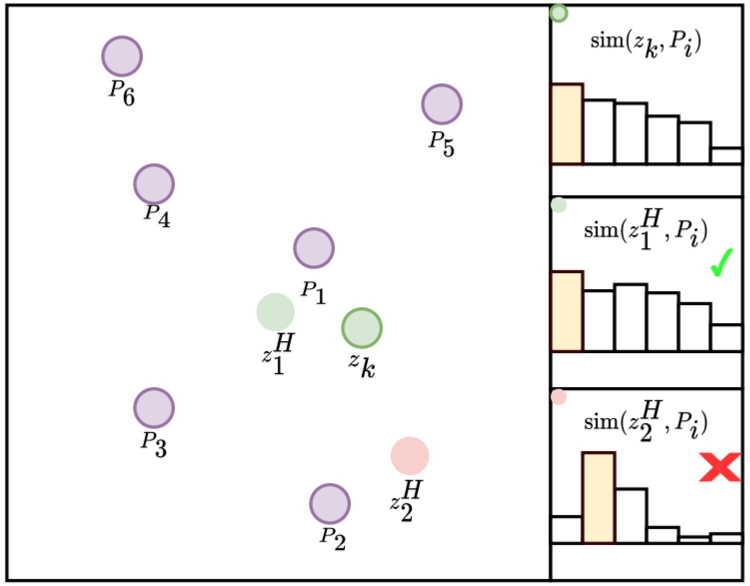


Our approach

- Formally, we define our objective as

$$z^* = \arg \min_{z \in \mathbb{S}^{D-1}} \ \mathrm{sim}(z, P^*_{z_k})$$
$$\text{s.t.} \quad \mathrm{sim}(z, P^*_{z_k}) \geq \mathrm{sim}(z, P), \forall P \in \mathcal{P} \setminus \{P^*_{z_k}\}$$

- We want to generate hard positives which are far from the anchor but have the same closest prototype as the anchor
- Expensive : requires iterative solver



Relaxation 1 : Restrict the search space

- Instead of searching in the whole space, we restrict the search of a new positive in a particular direction

$$z = \mathbf{proj}(z_k + d),$$

- We define the direction as that joining the anchor and a randomly selected prototype along the hypersphere

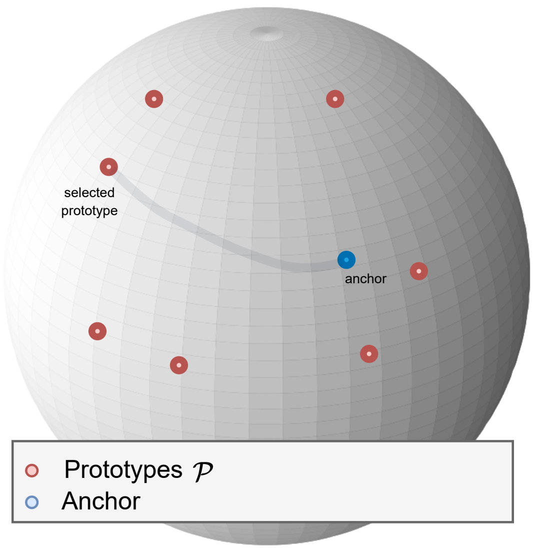
Relaxation 1 : Restrict the search space

- Instead of searching in the whole space, we restrict the search of a new positive in a particular direction

$$z = \mathrm{proj}(z_k + d),$$

- We define the direction as that joining the anchor and a randomly selected prototype along the hypersphere
- Manifold-Mixup along the geodesic

$$d(t, P_{\mathrm{sel}}, z_k) = \frac{\sin(1-t)\Omega}{\sin\Omega} z_k + \frac{\sin(t\Omega)}{\sin\Omega} P_{\mathrm{sel}} - z_k,$$
$$\text{where } t \in [0,1], \cos\Omega = {P_{\mathrm{sel}}}^\top z_k \text{ and } \Omega \in [0,\pi]$$

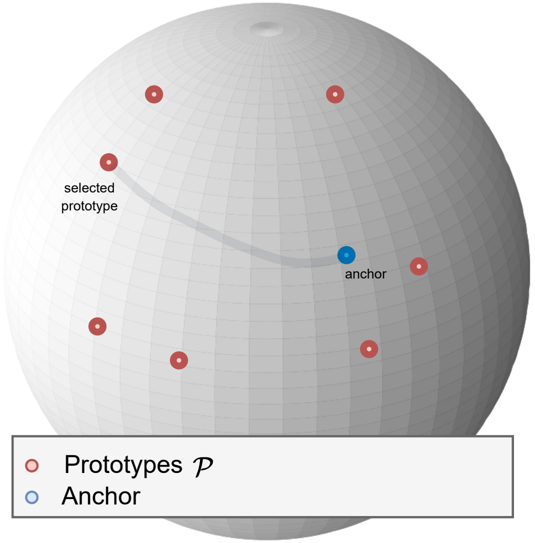


Relaxation 1 : Restrict the search space

- Instead of searching in the whole space, we restrict the search of a new positive in a particular direction

$$z = \mathrm{proj}(z_k + d),$$

- We define the direction as that joining the anchor and a randomly selected prototype along the hypersphere
- Manifold-Mixup along the geodesic



$$d(t, P_{\mathrm{sel}}, z_k) = \frac{\sin(1-t)\Omega}{\sin\Omega} z_k + \frac{\sin(t\Omega)}{\sin\Omega} P_{\mathrm{sel}} - z_k,$$
$$\text{where } t \in [0,1], \cos\Omega = {P_{\mathrm{sel}}}^\top z_k \text{ and } \Omega \in [0,\pi]$$

$$\boxed{\begin{aligned} t^* &= \arg\min_{t\in[0,1]} \mathrm{sim}(z, P^*_{z_k}), \text{ where} \\ z &= z_k + d(t, P_{\mathrm{sel}}, z_k) \\ \text{s.t.} \quad & \mathrm{sim}(z, P^*_{z_k}) \geq \mathrm{sim}(z, P_j), P_j \in \mathcal{P} \end{aligned}}$$

Relaxation 2

- Instead of solving the ranking objective for all prototypes, just solve it for closest and selected

$$t^* = \arg\min_{t \in [0,1]} \operatorname{sim}(z, P^*_{z_k}), \text{where}$$

$$z = z_k + d(t, P_{\text{sel}}, z_k)$$

$$\boxed{\text{s.t.} \quad \operatorname{sim}(z, P^*_{z_k}) \geq \operatorname{sim}(z, P_{\text{sel}}),}$$

Relaxation 2

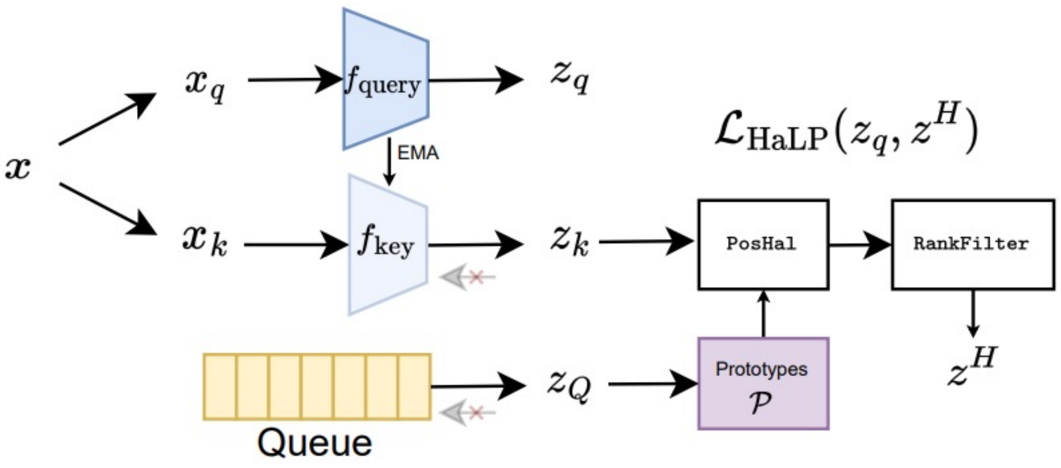
- Instead of solving the ranking objective for all prototypes, just solve it for closest and selected

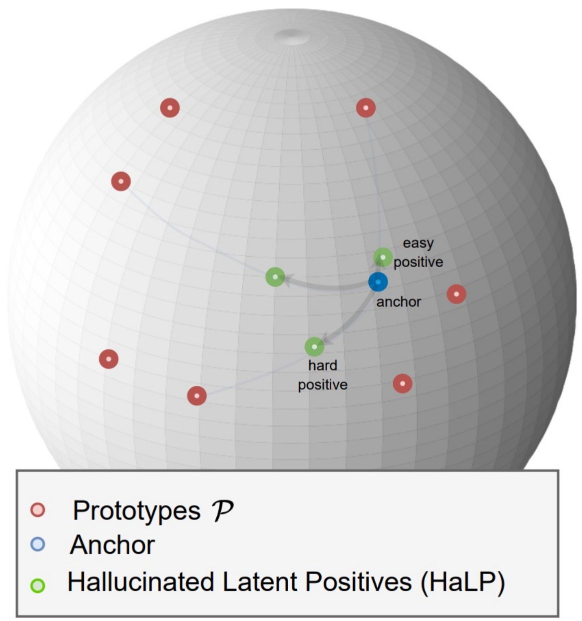
$$
\begin{aligned}
&t^* = \arg\min_{t\in[0,1]} \ \mathrm{sim}(z, P^*_{z_k}), \text{where} \\
&z = z_k + d(t, P_{\mathrm{sel}}, z_k) \\
&\text{s.t.} \quad \mathrm{sim}(z, P^*_{z_k}) \geq \mathrm{sim}(z, P_{\mathrm{sel}}),
\end{aligned}
$$

- Let's us derive a closed form solution

$$
\begin{aligned}
t^* &= \frac{1}{\Omega} \arctan\left(\frac{\sin\Omega}{\kappa + \cos\Omega}\right), \text{where} \\
\kappa &= \frac{1 - {P_{\mathrm{sel}}}^\top P^*_{z_k}}{{z_k}^\top (P^*_{z_k} - P_{\mathrm{sel}})},
\end{aligned}
$$

The final approach





$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CL}} + \mu \mathcal{L}_{\text{HaLP}} \quad \text{where}$$

$$\mathcal{L}_{\text{HaLP}} = -\frac{1}{G_{\text{filtered}}} \sum_{i}^{G_{\text{filtered}}} z_q^\top z_i^H / \tau$$

$$z_i^H \; = \; z_k + d(t_c, P_{\text{sel}}, z_k)$$

$$t_c \sim \mathtt{uniform}(0, \lambda t^*)$$

Key implementation details

- Datasets :
  - NTU-60
  - NTU-120
  - PKU-v2
- Encoder : BiGRU
- Works for both unimodal and multi-modal training
- Evaluation protocols :
  - Linear evaluation
  - kNN evaluation
  - Transfer learning
  - Semisupervised Learning

Results : Linear evaluation

Method	NTU-60		NTU-120		PKU-II
	x-sub	x-view	x-sub	x-set	x-sub
Additional training modalities or encoders					
ISC [50]	76.3	85.2	67.1	67.9	36.0
CrosSCLR-B [38]	77.3	85.1	67.1	68.6	41.9
CMD [38]	79.8	86.9	70.3	71.5	43.0
HaLP + CMD	**82.1**	**88.6**	**72.6**	**73.1**	**47.5**
Training using only joint					
LongT GAN [64]	39.1	48.1	-	-	26.0
MS^{2}L [37]	52.6	-	-	-	27.6
P&C [49]	50.7	76.3	42.7	41.7	25.5
AS-CAL [43]	58.5	64.8	48.6	49.2	-
H-Transformer [9]	69.3	72.8	-	-	-
SKT [61]	72.6	77.1	62.6	64.3	-
GL-Transformer [29]	76.3	83.8	66.0	68.7	-
SeBiReNet [41]	-	79.7	-	-	-
AimCLR [18]	74.3	79.7	-	-	-
Baseline	78.0	85.5	69.1	69.8	42.9
HaLP	**79.7**	**86.8**	**71.1**	**72.2**	**43.5**

Results : Linear evaluation

Method	NTU-60		NTU-120		PKU-II
	x-sub	x-view	x-sub	x-set	x-sub
Additional training modalities or encoders					
ISC [50]	76.3	85.2	67.1	67.9	36.0
CrosSCLR-B [38]	77.3	85.1	67.1	68.6	41.9
CMD [38]	79.8	86.9	70.3	71.5	43.0
HaLP + CMD	**82.1**	**88.6**	**72.6**	**73.1**	**47.5**
Training using only joint					
LongT GAN [64]	39.1	48.1	-	-	26.0
MS^2L [37]	52.6	-	-	-	27.6
P&C [49]	50.7	76.3	42.7	41.7	25.5
AS-CAL [43]	58.5	64.8	48.6	49.2	-
H-Transformer [9]	69.3	72.8	-	-	-
SKT [61]	72.6	77.1	62.6	64.3	-
GL-Transformer [29]	76.3	83.8	66.0	68.7	-
SeBiReNet [41]	-	79.7	-	-	-
AimCLR [18]	74.3	79.7	-	-	-
Baseline	78.0	85.5	69.1	69.8	42.9
HaLP	**79.7**	**86.8**	**71.1**	**72.2**	**43.5**

Results : Transfer learning and kNN evaluation

Method	To PKU-II	
	NTU-60	NTU-120
Additional training modalities or encoders		
ISC [50]	51.1	52.3
CrosSCLR-B	54.0	52.8
CMD	56.0	57.0
HaLP + CMD	**56.6**	**57.3**
Training using only joint		
LongT GAN [64]	44.8	-
MS^2L [37]	45.8	-
Baseline	53.3	53.4
HaLP	**54.8**	**55.4**

Transfer to PKU-II

Method	NTU-60		NTU-120	
	x-sub	x-view	x-sub	x-set
Additional training modalities or encoders				
ISC [50]	62.5	82.6	50.6	52.3
CrosSCLR-B	66.1	81.3	52.5	54.9
CMD	70.6	85.4	58.3	60.9
HaLP+CMD	**71.0**	**86.4**	**59.4**	**61.9**
Additional training modalities or encoders				
LongT GAN [64]	39.1	48.1	31.5	35.5
P&C [49]	50.7	76.3	39.5	41.8
Baseline	63.6	82.8	51.7	55.3
HaLP	**65.8**	**83.6**	**55.8**	**59.0**

kNN Evaluation

Analysis: Computational overheads

Method	Time/epoch	Train GPU memory	NTU-60 x-sub
Baseline	1x	1x	78.0
HaLP	1.13x	1x	**79.7**
CMD	3x	1.94x	79.8
HaLP+CMD	3.32x	1.94x	**82.1**

Use with alternative tasks and frameworks

	NCI-1	PROTEINS	DD	MUTAG
GraphCL	77.87 ± 0.41	74.39 ± 0.45	78.62 ± 0.40	86.80 ± 1.3
+HaLP	$\mathbf{78.88 \pm 0.41}$	$\mathbf{74.65 \pm 0.70}$	$\mathbf{79.20 \pm 0.60}$	$\mathbf{89.35 \pm 1.2}$

Graph Representation Learning

Approach	NTU-60 x-sub
CMD	79.8
CMD+HaLP	82.1
AimCLR	74.3
AimCLR + HaLP	75.2

HaLP with AimCLR

JUNE 18-22, 2023

CVPR

VANCOUVER, CANADA

HaLP: Hallucinating Latent Positives for Skeleton-based Self-Supervised Learning of Actions

Code

Questions ?

Poster Session

Anshul Shah
ashah95@jhu.edu

THU-AM-226