2021 ICCV OCTOBER 11-17 VIRTUAL

Improving Contrastive Learning by Visualizing Feature Transformation

Rui Zhu*, Bingchen Zhao*, Jingen Liu†, Zhenglong Sun, Chang Wen Chen

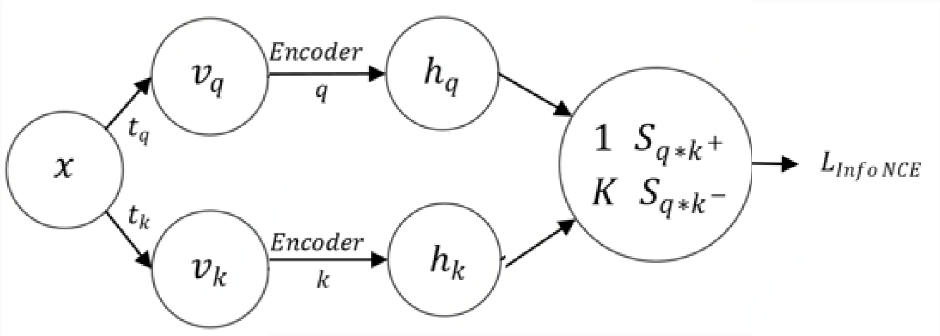
香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

JDAI Research

TONGJI UNIVERSITY 同濟大學 1907

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Pipeline of Contrastive Learning



Data Augmentation

- Generate views from one image by data augmentation

Encode

- Encode views into feature space

Pos/Neg pair score

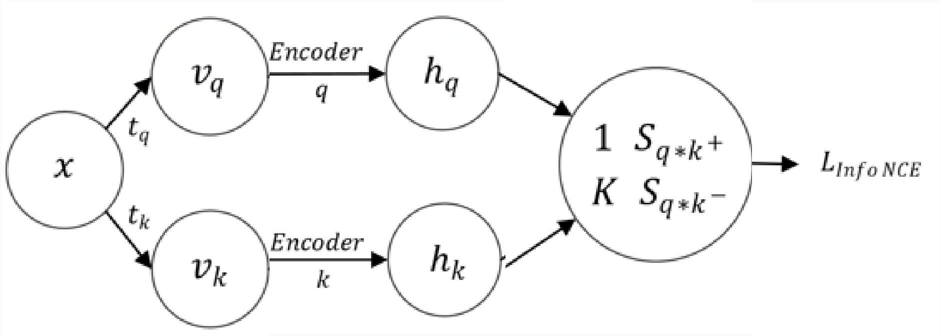
- Build positive and negative pair score

Contrast

- Contrast positive and negative in loss

Heavily rely on data augmentation!

Pipeline of Contrastive Learning



Data Augmentation

- Generate views from one image by data augmentation

Encode

- Encode views into feature space

Pos/Neg pair score

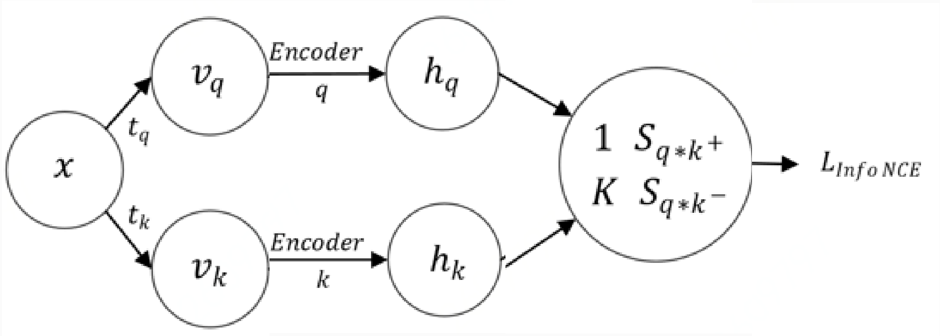
- Build positive and negative pair score

Contrast

- Contrast positive and negative in loss

Heavily rely on data augmentation!

Pipeline of Contrastive Learning



Data Augmentation

- Generate views from one image by data augmentation

Encode

- Encode views into feature space

Pos/Neg pair score

- Build positive and negative pair score

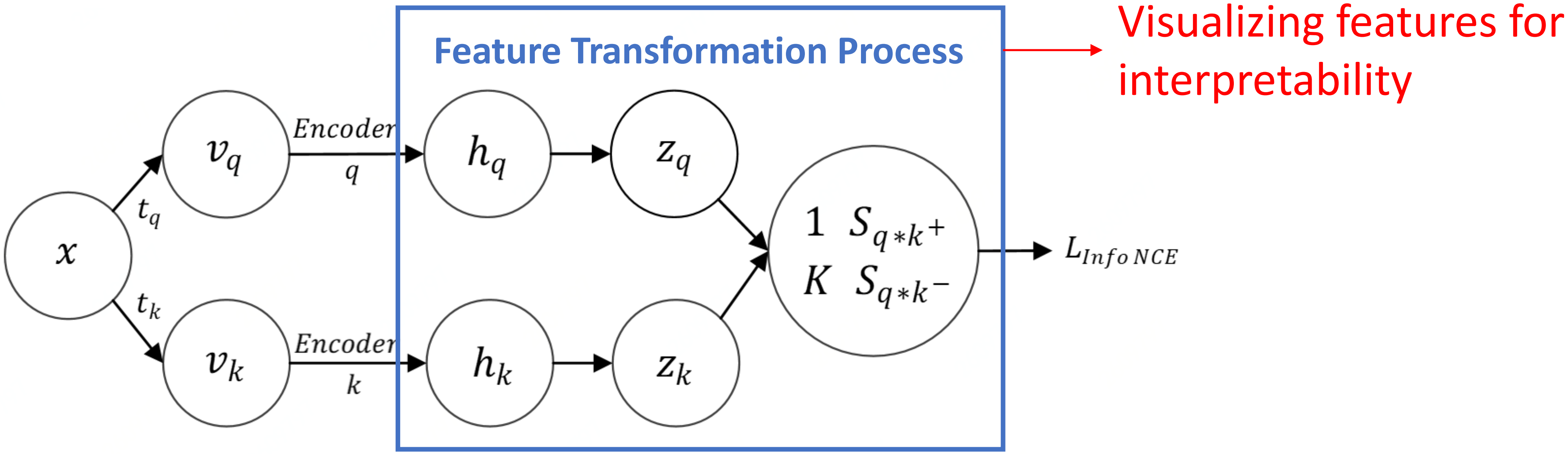
Contrast

- Contrast positive and negative in loss

How about Feature Transformation?

- Hard positives
- Diversified negatives

Feature Transformation



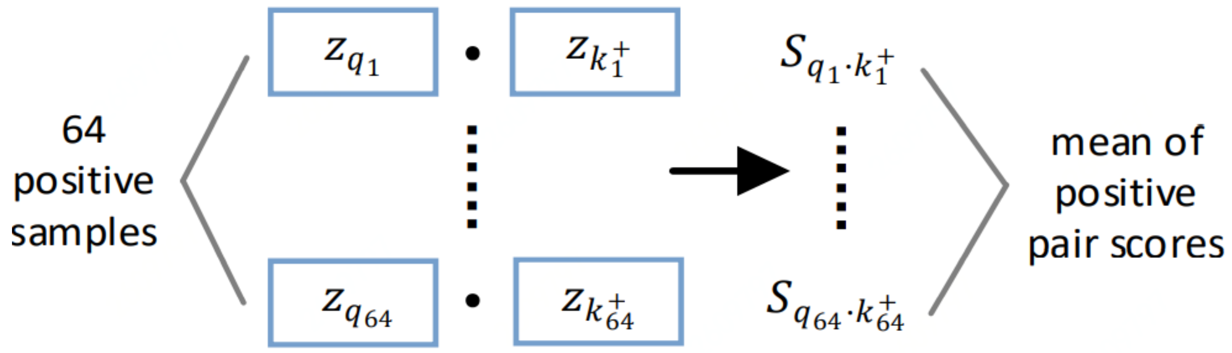
Unlike data augmentation, we propose Feature Transformation:

- Directly operate on feature embedding.
- Not based on human intuitive.
- Manipulate positive or negative pairs for different purpose.

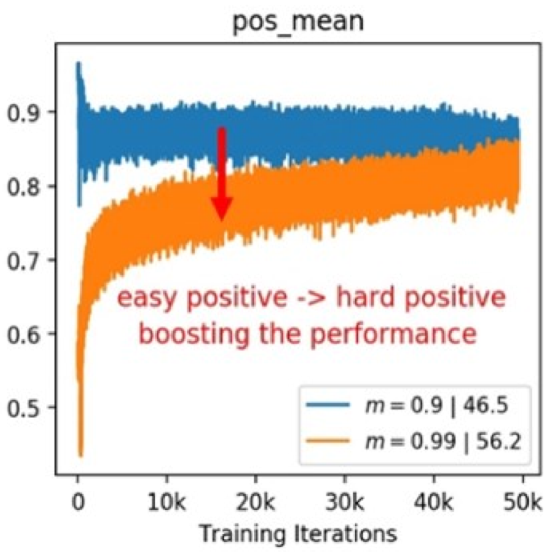
Visualizing Features or Pos/Neg Scores?

- Challenges of visualizing features:
  - *Costly to visualize high-dimensional features.*
  - *Needs large storage.*

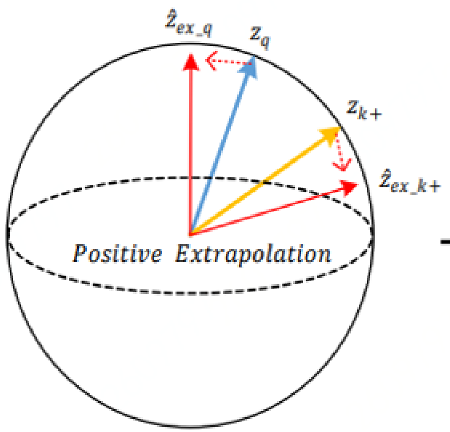
- Visualizing the statistics of pair score distribution is better:
  - *Positive/Negative Pair score → the minimum unit of contrastive loss.*
  - *Offline → no impact on training speed.*
  - *Negligible computation → being feasible for large scale dataset.*



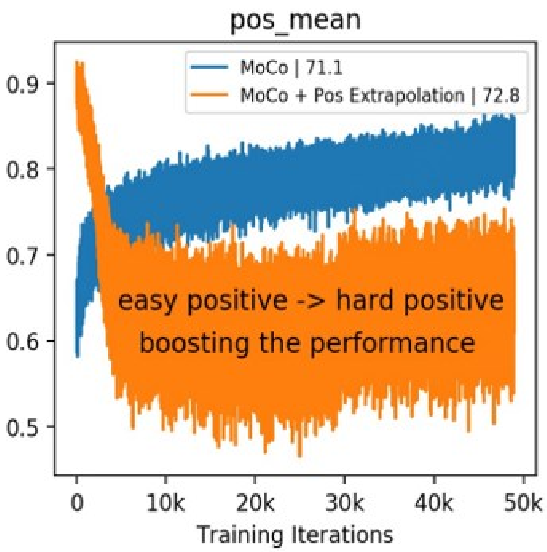
From Visualization to Feature Transformation



Observation



Proposed Feature Transformation



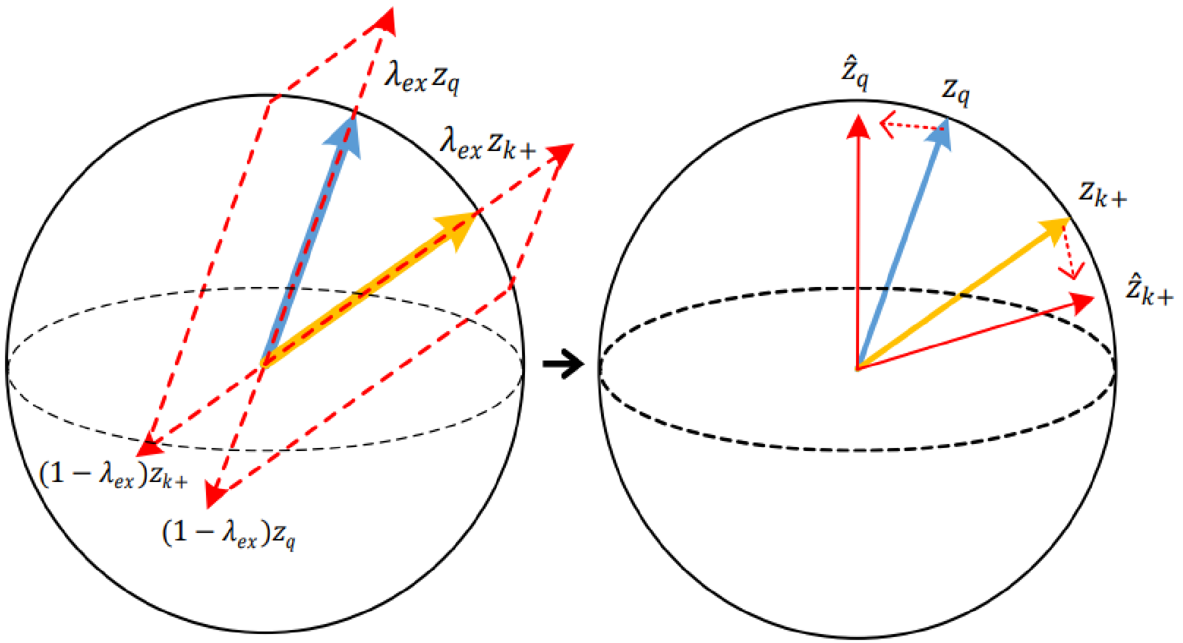
Performance Gain

- Observation: hard positive → higher transfer accuracy.
- Feature Transformation : hard positives for more view invariance.
- Explain the impact of model parameter by visualization tools.
- Trace back the training process by visualization tools.

Contributions

- ➢ Propose Feature Transformation to enhance contrastive learning:
  - *Extrapolate positive pairs → hard positives → to learn view invariance for model.*
  - *Interpolate negative samples → diversified negatives → to learn discriminative representations*
- ➢ Design a practical visualization tool → to trace back analyze training process.
- ➢ Empirically analyze the efficacy of Feature Transformation.
- ➢ Extensive experiments and good results on down-stream tasks.

Feature Transformation: Positive Extrapolation



Increase view variance of positive pair:

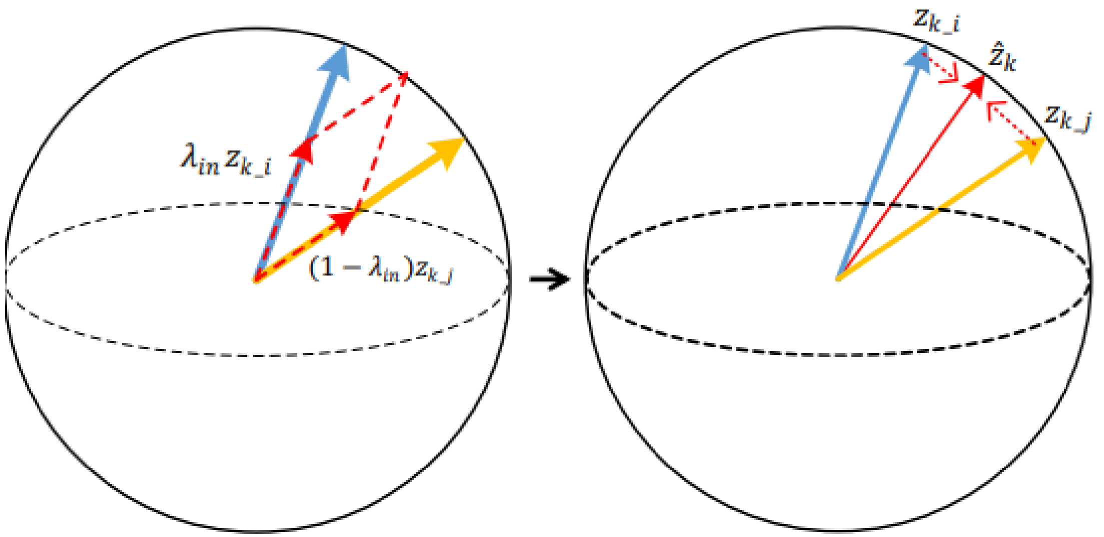
- Extrapolation pushes away positive pair
- A minor direction change to convey a larger view variance
- Transfer easy positives to hard positives.

What if the positive interpolation?

- Obvious performance drops
- The view variance of positive pairs ↓

Method	α_{ex}	pos interpolation/extrapolation
MoCo	0.2	69.1 / 71.6
(baseline: 71.1)	2.0	67.4 / 72.8

Feature Transformation: Negative Interpolation



Increase the diversity of negative examples:

- Randomly interpolating two features in queue.
- Contrast with more new negatives in each training step.
- Original queue → discrete distribution of negatives.
- Fill in the incomplete distribution, leading to a more discriminative model.

Extending queue or Negative Feature Transformation?

- Original queue (even doubled) << Negative FT queue.
- Negative FT queue + Original queue ≈ Negative FT queue.
- Negative FT provides sufficient diversified negatives.

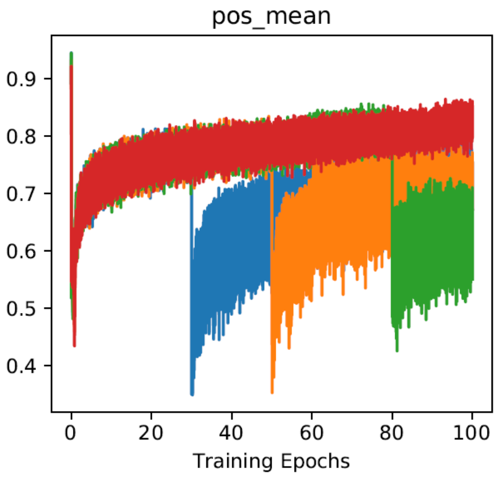
Method	α_{in}	Z_n	queue size	Acc
moco+ original queue	-	Z_{neg}	K	71.10
moco+ original queue	-	Z_{neg}	$2K$	71.40
moco+ Neg FT queue	1.6	$\hat{Z}_{neg}$	K	74.64
moco+ Neg FT+original	1.6	$\tilde{Z}_{neg}$	$2K$	74.73

Discussion: When to add Feature transformation?

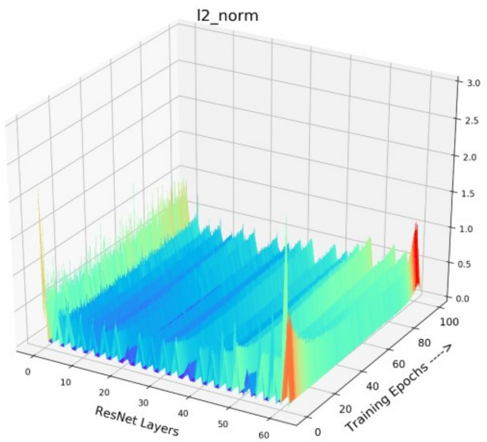
Starting Feature Transformation in the various training stage:

- Consistently boosts the accuracy.
- Starting earlier improves more.
- Providing hard positives when inserted.
- Bringing a greater gradient for training.
- Plug-and-play

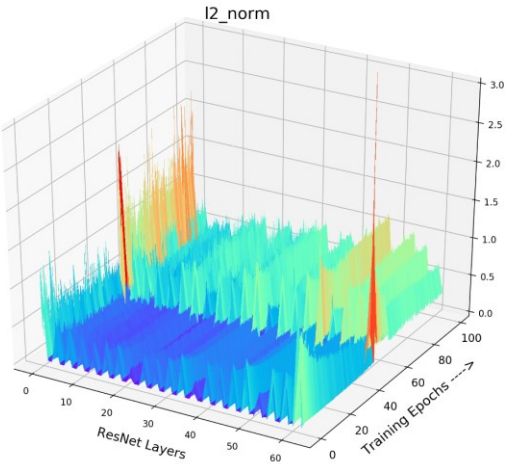
FT begin epoch	0	2	30	50	80	-
Res18 acc (%)	62.6	**63.3**	62.9	61.8	59.2	56.2
Res50 acc (%)	**76.9**	76.4	75.9	74.0	72.2	71.1



Mean of positive scores



Baseline MoCo gradient landscape



Adding FT in 50th epoch

Discussion: Could the gains of FT vanish if training longer?

Method	Pre-train Epochs	Acc %
MoCo-V2 → MoCo-V2 + FT	200	75.6 → 78.3, 2.7%↑
(on ImageNet-100)	500	80.7 → 81.5, 0.8%↑

- Longer training weakens the improvement from Feature Transformation.
- More epochs → contrast more positive and negative pairs.
- Fast convergence by providing diversified and discriminative pairs.

Ablation studies on ImageNet-100:

Method	MoCov1	MoCov2	simCLR	Infomin	swav	SimSiam
baseline*	71.10	75.61	74.32	81.9	82.1	77.1
+pos FT	72.80	76.22	75.80	-	-	77.8
+neg FT	74.64	77.12	76.71	-	-	
+both	76.87	78.33	78.25	83.2	83.2	
+both$_{dim}$	**77.21**	**79.21**	**78.81**	-	-	

- Positive and negative Feature Transformation are complementary.
- Generic and robust for various contrastive models.
- Boosts the MoCo-V1, MoCo-V2 and SIMCLR.

Results on ImageNet-1K and Transfer to Fine-grained Dataset:

pre-train	IN-1k	inat-18	CUB200	FGVC-aircraft
supervised	76.1	66.1	81.9*	82.6*
mocov1[14]	60.6	65.6	82.8*	83.5*
mocov1+ours	61.9	67.3	83.2	84.0
mocov2[7]	67.5	66.8*	82.9*	83.6*
mocov2+ours	**69.6**	**67.7**	**83.1**	**84.1**
mocov2+MoCHi[20]	68.0	-	-	-
mocov2+UnMix[38]	68.6	-	-	-

- Improves MoCo-V1 and MoCo-V2 by 1.3% and 2.1% on Imagenet-1K.
- Larger performance gain than mixup based methods, e.g., UnMix[1] and MoCHi[2] respectively.
- Better transfer performance on iNaturalist2018.
- Consistent improvement on CUB-200 and FGVC-aircraft.

[1] Shen, Z., Liu, Z., Liu, Z., Savvides, M., Darrell, T., & Xing, E. Un-mix: Rethinking image mixtures for unsupervised visual representation learning. arXiv:2003.05438.
[2] Kalantidis, Y., Sariyildiz, M. B., Pion, N., Weinzaepfel, P., & Larlus, D. Hard negative mixing for contrastive learning. NeurIPS 2020.

Transfer Performance on Object Detection Dataset:

pre-train	IN-1k Top-1	Faster [35] R50-C4 VOC AP	AP_{50}	AP_{75}	Mask R-CNN [15] R50-C4 COCO AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
random init*	-	33.8	60.2	33.1	26.4	44.0	27.8	29.3	46.9	30.8
supervised*	76.1	53.5	81.3	58.8	38.2	58.2	41.2	33.3	54.7	35.2
infomin*	70.1	57.6	82.7	64.6	39.0	58.5	42.0	34.1	55.2	36.3
mocoV1[14]	60.6	55.9	81.5	62.6	38.5	58.3	41.6	33.6	54.8	35.6
mocoV1+ours	61.9	56.1	82.0	62.0	39.0	58.7	42.1	34.1	55.1	36.0
mocoV2[7]	67.5	57.0	82.4	63.6	39.0	58.6	41.9	34.2	55.4	36.2
mocoV2+ours	**69.6**	**58.1**	**83.3**	65.1	**39.5**	**59.2**	42.1	**34.6**	55.6	36.5
mocoV2+mochi[20]	68.0	57.1	82.7	64.1	39.4	59.0	42.7	34.5	55.7	36.7
DetCo[53]	68.6	57.8	82.6	64.2	39.4	59.2	42.3	34.4	55.7	36.6
InsLoc[55]	-	57.9	82.9	65.3	39.5	59.1	**42.7**	34.5	56.0	36.8

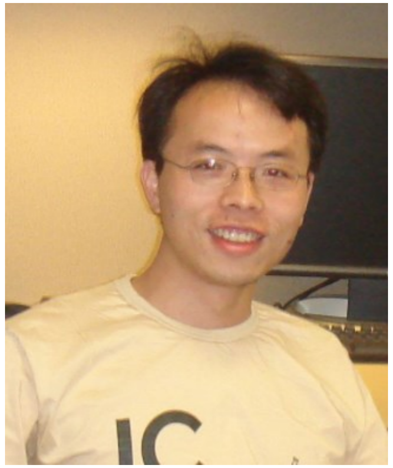
- Strongly improves the transfer accuracy on PASCAL VOC and MSCOCO.
- Less task-biased and generic:

  Beats some detection-oriented methods (DetCo[1] and InsLoc[2]).

[1] Xie, E., Ding, J., Wang, W., Zhan, X., Xu, H., Li, Z., & Luo, P. Detco: Unsupervised contrastive learning for object detection. ICCV 2021.
[2] Yang, C., Wu, Z., Zhou, B., & Lin, S. Instance localization for self-supervised detection pretraining. CVPR 2021.

Thanks for Listening!





Rui Zhu

The Chinese University of Hong Kong, Shenzhen

JD AI Research, Beijing

Bingchen Zhao

Tongji University, Shanghai

Jingen Liu

JD AI Research, Mountain View

Zhenglong Sun

The Chinese University of Hong Kong, Shenzhen

Chang Wen Chen

The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China

Codes at Github!

https://github.com/DTennant/CL-Visualizing-Feature-Transformation