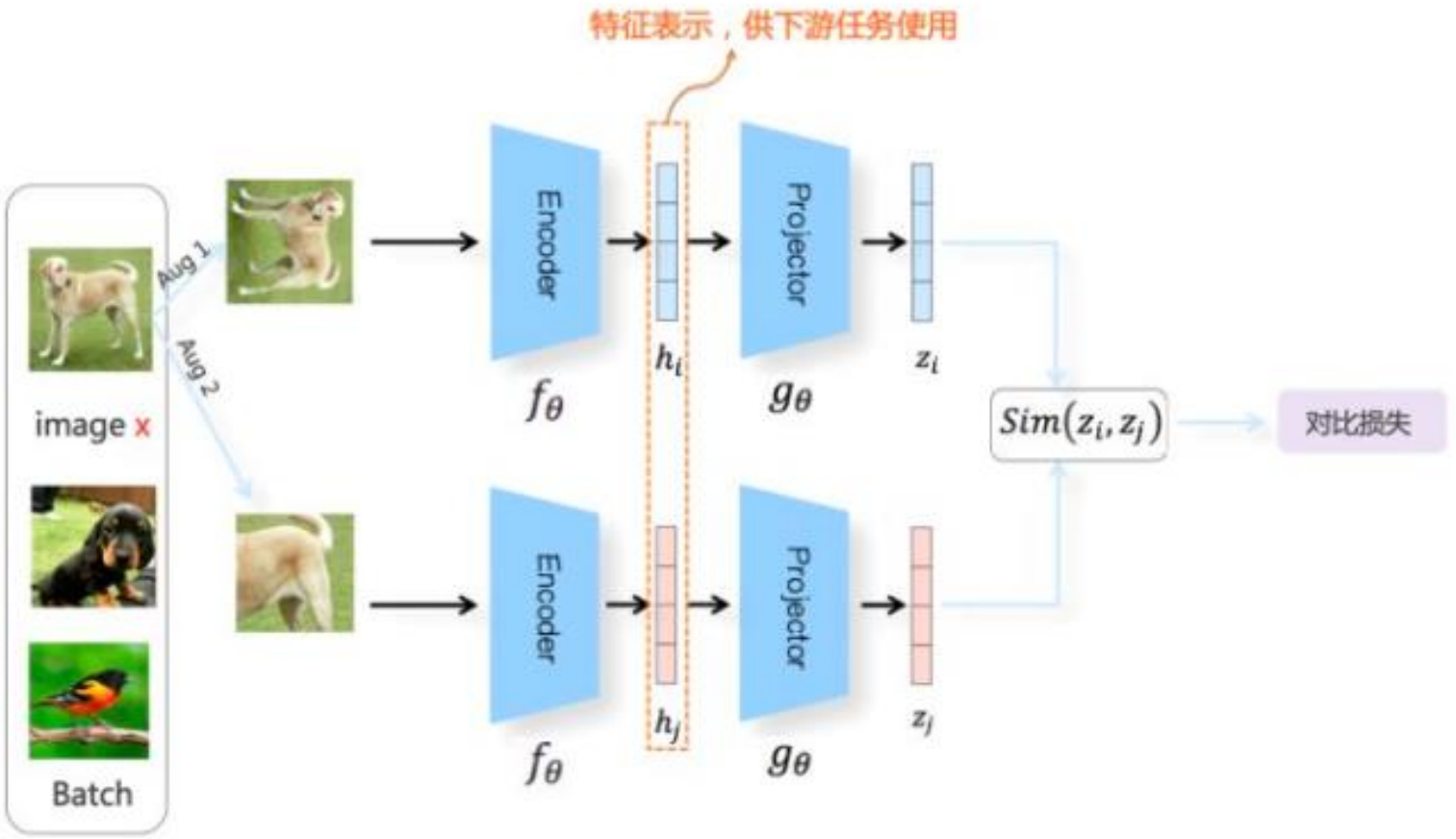
南京航空航天大學

Nanjing University of Aeronautics and Astronautics

Two articles on Contrastive Learning

Contrastive learning



$$L_i = -log(exp(S(z_i, z_i^+) \ / \ \tau) / \sum_{(j=0)}^{K} exp(S(z_i, z_j) \ / \ \tau))$$

Motion-aware Contrastive Video Representation Learning via Foreground-background Merging

Shuangrui Ding[1*] Maomao Li[2] Tianyu Yang[2] Rui Qian[3]
Haohang Xu[1] Qingyi Chen[4] Jue Wang[2] Hongkai Xiong[1†]
[1]Shanghai Jiao Tong University [2]Tencent AI Lab
[3]The Chinese University of Hong Kong [4]University of Michigan
{dsr1212, xuhaohang, xionghongkai}@sjtu.edu.cn tianyu-yang@outlook.com
{limaomao07, arphid}@gmail.com qr021@ie.cuhk.edu.hk chenqy@umich.edu

CVPR 2022

Motivation

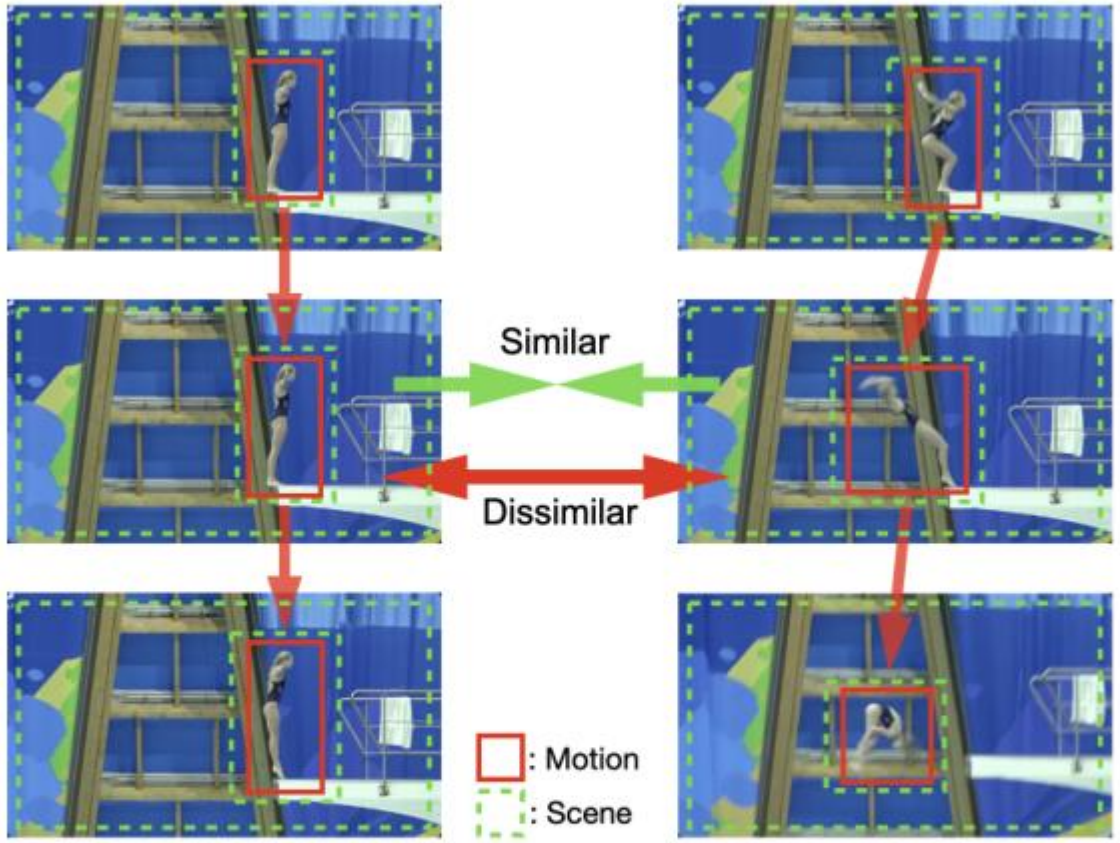


Figure 2. Illustration on a diving sequence. The green dashed box represents the scene and the red box means motion area. The two clips have the same background but distinct motions. Drawing such positive pairs closer inclines the model towards static bias.

Method

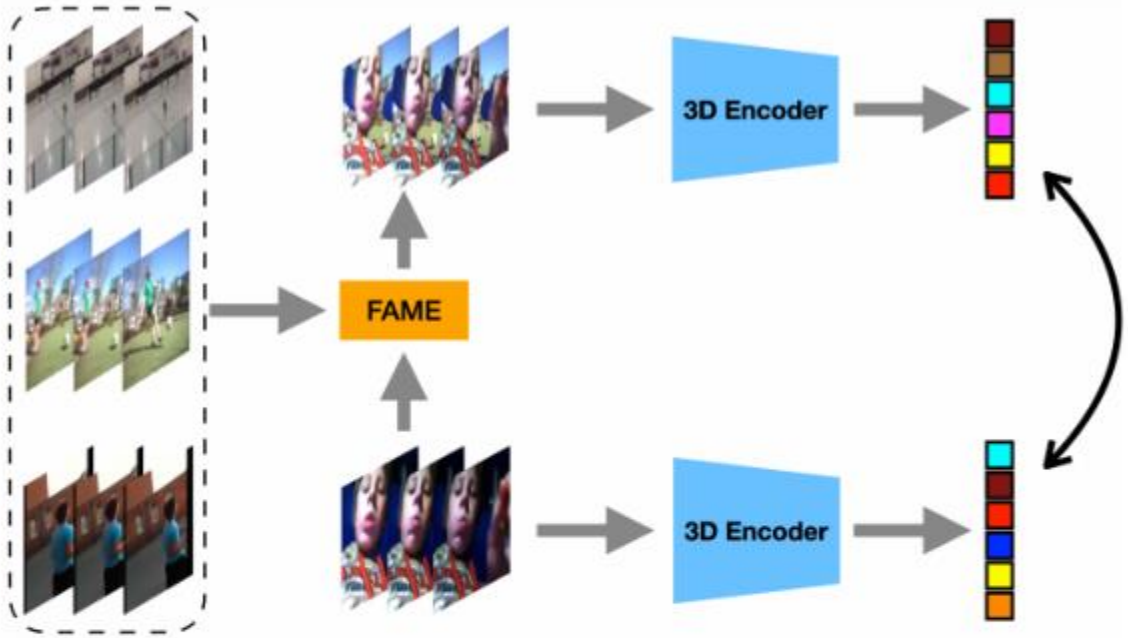
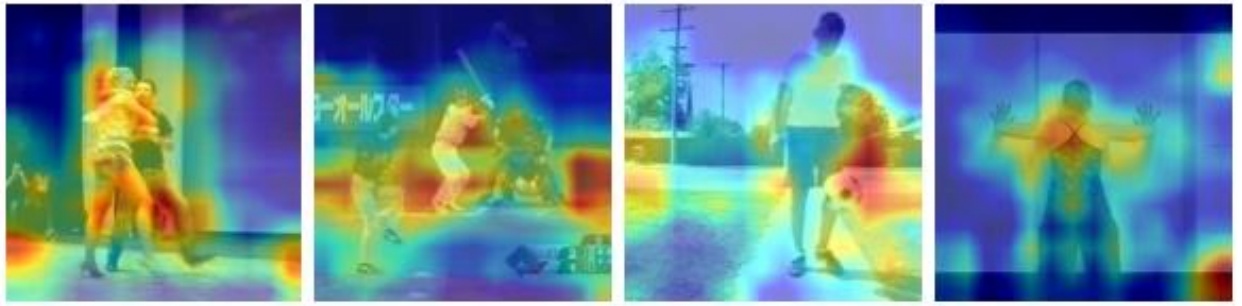
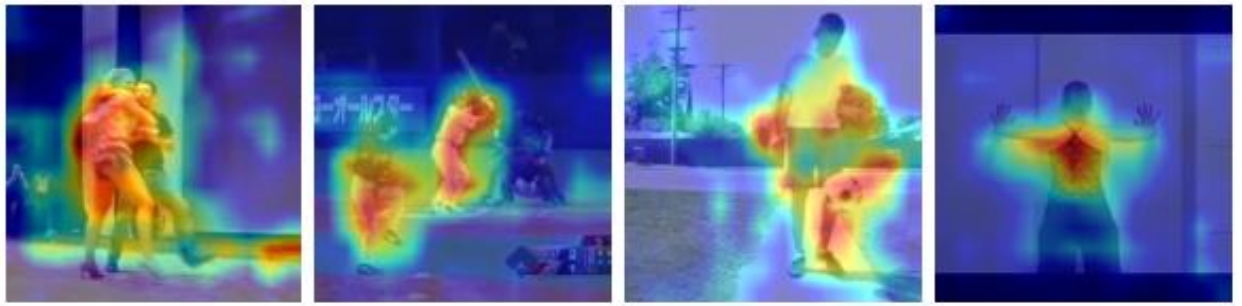


Figure 4. The contrastive learning framework with the proposed FAME. We first randomly sample two clips from a video and use FAME to generate new clips by composing the original foreground onto various backgrounds from other videos. Then, we feed the augmented clips into the existing contrastive learning scheme and perform self-supervised pretraining.

Experiments



(a) Results of vanilla contrastive learning.



(b) Results of our approach FAME.

Figure 1. Class-agnostic activation map [3] visualization of important areas. The heatmap indicates how much the pretrained model attends to the region. Compared to the conventional approach, our method mitigates the background bias significantly.

Experiments

Action Recognition: A video clip contains a clear action,input a video, and output its action classification

Method	Backbone	Pretrain Dataset	Frames	Res.	Freeze	UCF101	HMDB51
CBT [49]	S3D	Kinetics-600	16	112	✓	54.0	29.5
CCL [33]	R3D-18	Kinetics-400	16	112	✓	52.1	27.8
MemDPC [22]	R3D-34	Kinetics-400	40	224	✓	54.1	30.5
RSPNet [6]	R3D-18	Kinetics-400	16	112	✓	61.8	**42.8**
MLRep [43]	R3D-18	Kinetics-400	16	112	✓	63.2	33.4
FAME (Ours)	R(2+1)D	Kinetics-400	16	112	✓	**72.2**	42.2
VCP [38]	R(2+1)D	UCF101	16	112	✗	66.3	32.2
PRP [64]	R(2+1)D	UCF101	16	112	✗	72.1	35.0
TempTrans [29]	R(2+1)D	UCF101	16	112	✗	81.6	46.4
3DRotNet [30]	R3D-18	Kinetics-400	16	112	✗	62.9	33.7
Spatio-Temp [55]	C3D	Kinetics-400	16	112	✗	61.2	33.4
Pace Prediction [56]	R(2+1)D	Kinetics-400	16	112	✗	77.1	36.6
SpeedNet [4]	S3D-G	Kinetics-400	64	224	✗	81.1	48.8
VideoMoCo [41]	R(2+1)D	Kinetics-400	32	112	✗	78.7	49.2
RSPNet [6]	R(2+1)D	Kinetics-400	16	112	✗	81.1	44.6
MLRep [43]	R3D-18	Kinetics-400	16	112	✗	79.1	47.6
ASCNet [26]	R3D-18	Kinetics-400	16	112	✗	80.5	52.3
SRTC [69]	R(2+1)D	Kinetics-400	16	112	✗	82.0	51.2
FAME (ours)	R(2+1)D	Kinetics-400	16	112	✗	**84.8**	**53.5**
DSM [53]	I3D	Kinetics-400	16	224	✗	74.8	52.5
BE [54]	I3D	Kinetics-400	16	224	✗	86.8	55.4
FAME (ours)	I3D	Kinetics-400	16	224	✗	**88.6**	**61.1**

Table 5. Comparison with the existing self-supervised video representation learning methods for action recognition on UCF101 and HMDB51. To compare fairly, we list each work's setting, including backbone architecture used, pretrain dataset and spatial-temporal resolution. Freeze (tick) indicates linear probe, and no freeze (cross) means finetune.

Targeted Supervised Contrastive Learning for Long-Tailed Recognition

Tianhong Li[1,*] Peng Cao[1,*] Yuan Yuan[1] Lijie Fan[1] Yuzhe Yang[1]
Rogerio Feris[2] Piotr Indyk[1] Dina Katabi[1]

[1]MIT CSAIL, [2]MIT-IBM Watson AI Lab

CVPR 2022

Related introduction(ICLR 2021, Exploring balanced feature spaces for representation learning.)

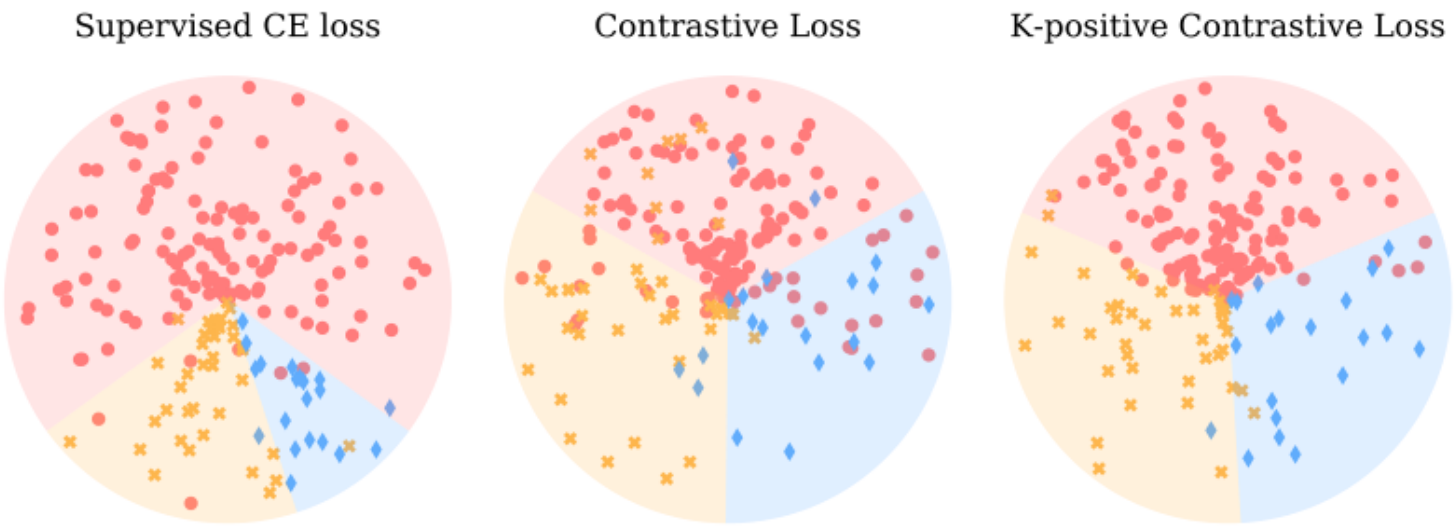


Figure 1: *Feature spaces learned with different losses given an imbalanced dataset.* The supervised cross-entropy (CE) learns a space biased to the dominant class. The space learned by unsupervised contrastive loss is balanced but less semantically discriminative. Our proposed k-positive contrastive loss learns a balanced and discriminative feature space. The shadow area ( ) indicates the decision boundary of each class.

$$\mathcal{L}_{\mathrm{CE}} = \frac{1}{N}\sum_{i=1}^{N} -\log p_{y_i} \qquad \mathcal{L}_{\mathrm{CL}} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{\exp(v_i \cdot v_i^+/\tau)}{\exp(v_i \cdot v_i^+/\tau) + \sum_{v_i^- \in V^-} \exp(v_i \cdot v_i^-/\tau)}$$

$$\mathcal{L}_{\mathrm{KCL}} = \frac{1}{N(k+1)}\sum_{i=1}^{N} \sum_{v_j^+ \in \{\tilde{v}_i\} \cup V_{i,k}^+} -\log \frac{\exp(v_i \cdot v_j^+/\tau)}{\exp(v_i \cdot \tilde{v}_i/\tau) + \sum_{v_j \in V_i} \exp(v_i \cdot v_j/\tau)}$$

Related introduction(PMLR2021, Dissecting supervised constrastive learning.)

Definition 3 (ρ-**Sphere-inscribed regular simplex**). *Let* $h, K \in \mathbb{N}$ *with* $K \leq h+1$. *We say that* $\zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ *form the vertices of a regular simplex inscribed in the hypersphere of radius* $\rho > 0$, *if and only if the following conditions hold:*

(S1) $\sum_{i \in [K]} \zeta_i = 0$

(S2) $\|\zeta_i\| = \rho$ *for* $i \in [K]$

(S3) $\exists d \in \mathbb{R} : d = \langle \zeta_i, \zeta_j \rangle$ *for* $1 \leq i < j \leq K$

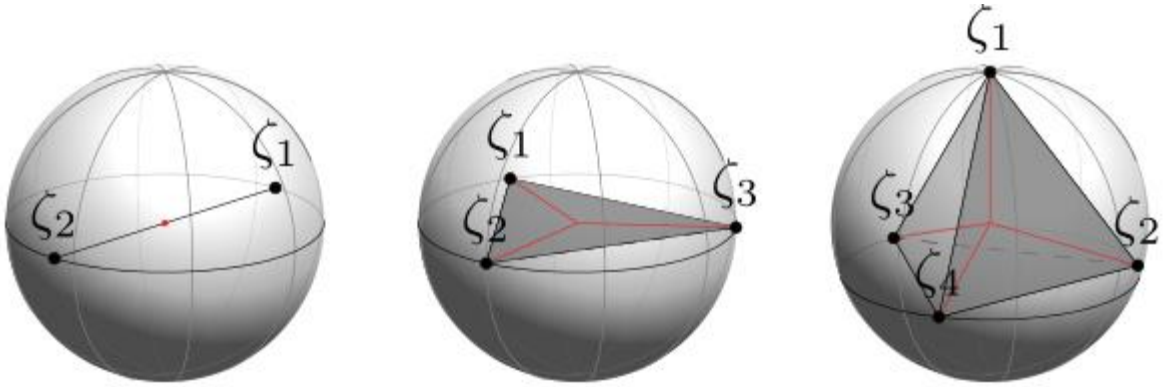


Figure 3: Regular simplices inscribed in $\mathbb{S}^2$.

Motivation

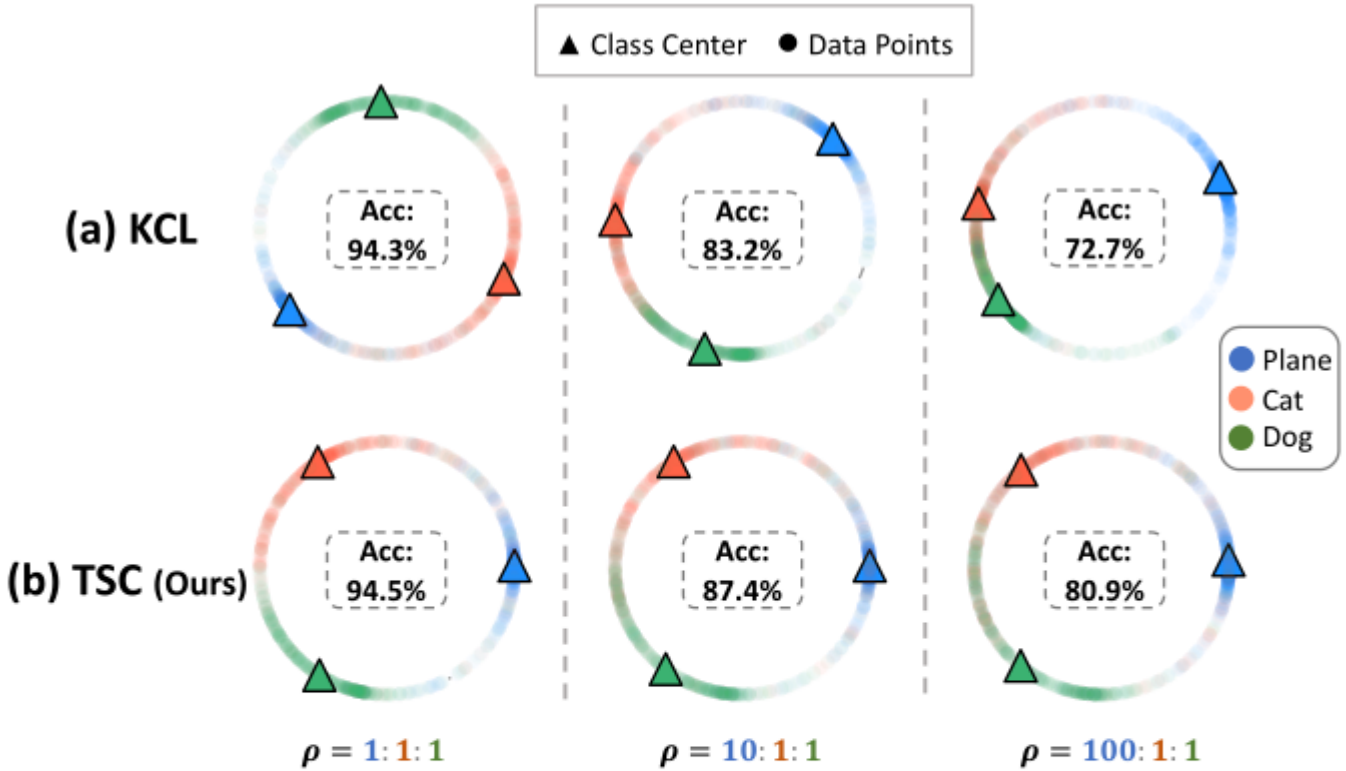


Figure 1. Test data feature distribution of (a) k-positive contrastive learning (KCL) and (b) TSC for three classes of CIFAR10 (plane, cat, dog), for different training data imbalance ratios ρ. With high imbalance ratio, class centers learned by KCL exhibit poor uniformity while class centers learned by TSC are still uniformly distributed and thus TSC achieves better performance (where Acc refers to Accuracy on test data).

Overview of TSC

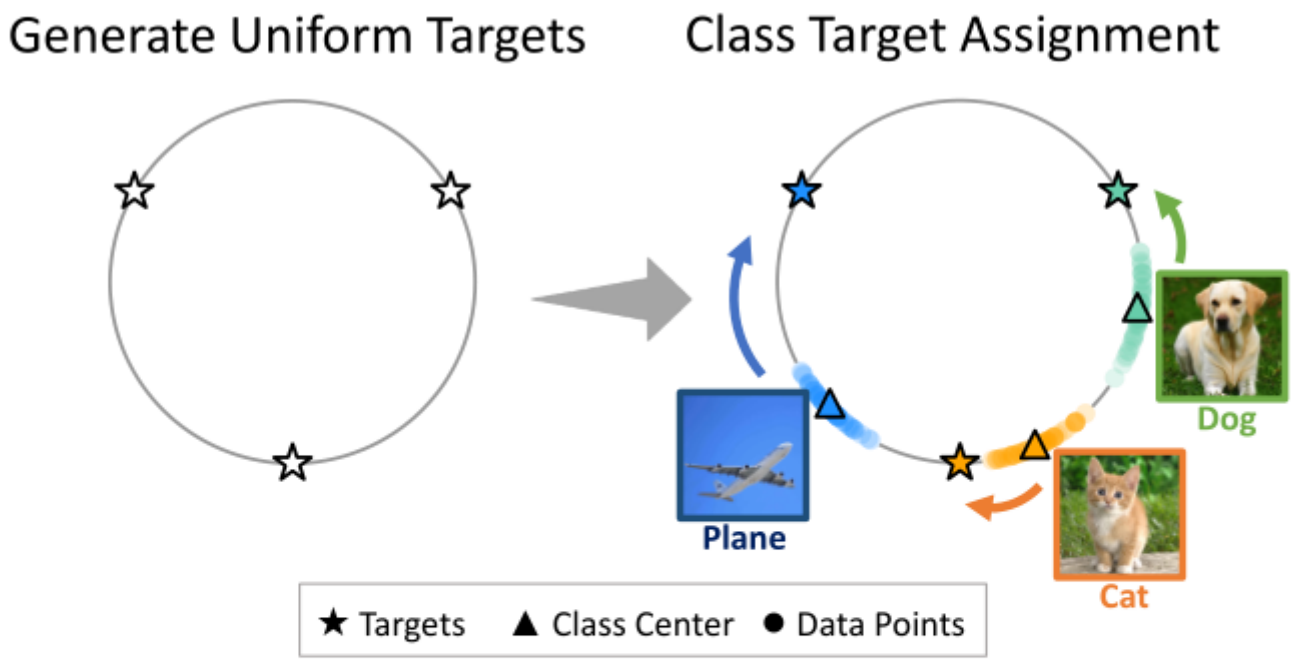


Figure 2. Illustration of TSC. It first computes the optimal targets for the class centers on the hypersphere. Then, during training, in each iteration, each target is assigned to the nearest class, and a targeted contrastive learning loss is designed to encourage the samples from each class to move to the assigned target position.

Method(Target Generation)

unit hypersphere $\mathcal{S}^{d-1} = \{u \in \mathbb{R}^d : ||u|| = 1\}$

target positions of C classes $\{t_i^*\}_{i=1}^C$,

d < (C− 1), computing the vertices of a regular simplex becomes very hard

$$\mathcal{L}_u(\{t_i\}_{i=1}^C) = \frac{1}{C}\sum_{i=1}^{C}\log\sum_{j=1}^{C} e^{t_i^T \cdot t_j/\tau}.$$

the minimum of Lu after gradient descent will be equal to its analytical minimum when d≥ (C− 1)

Method(Matching-Training Scheme)

- **Class-Target Assignment**

minimizes the distance between the target positions and the normalized class centers

$$\{\sigma_i^*\}_i = \underset{\{\sigma_i\}_i}{\arg\min} \frac{1}{C} \sum_{i=1}^{C} ||t_{\sigma_i} - c_i||$$

- **Targeted Supervised Contrastive Loss**

$$\mathcal{L}_{TSC} = -\frac{1}{N} \sum_{i=1}^{N} \Big(\frac{1}{k+1} \sum_{v_j^+ \in \tilde{V}_{i,k}^+} \log \frac{e^{v_i^T \cdot v_j^+ / \tau}}{\sum_{v_j \in \tilde{V}_i \cup U} e^{v_i^T \cdot v_j / \tau}} + \lambda \log \frac{e^{v_i^T \cdot t^*_{\sigma^*_{y_i}} / \tau}}{\sum_{v_j \in \tilde{V}_i \cup U} e^{v_i^T \cdot v_j / \tau}} \Big)$$

$$V_i = \{v_n\}_{n=1}^{N} \backslash \{v_i\}$$

$$\tilde{V}_i = \{\tilde{v}_i\} \cup V_i$$

$$\tilde{V}_{i,k}^+ = \{\tilde{v}_i\} \cup V_{i,k}^+$$

U is the set of pre-compute targets $U = \{t_i^*\}_{i=1}^{C}$.

Experiments

Table 1. Top-1 accuracy (%) of ResNet-32 on long-tailed CIFAR-10 and CIFAR-100. TSC consistently improves on past imbalanced learning techniques and achieves the best performance. Previous SOTA results for each imbalance ratio are colored with gray. We report the accuracy of our re-implemented KCL (†) since they do not report their performance on CIFAR in [18].

Dataset	CIFAR-10-LT			CIFAR-100-LT		
Imbalance Ratio (ρ)	100	50	10	100	50	10
CE	70.4	74.8	86.4	38.3	43.9	55.7
CB-CE [9]	72.4	78.1	86.8	38.6	44.6	57.1
Focal [27]	70.4	76.7	86.7	38.4	44.3	55.8
CB-Focal [9]	74.6	79.3	87.1	39.6	45.2	58.0
CE-DRW [4]	75.1	78.9	86.4	40.5	44.7	56.2
CE-DRS [4]	74.5	78.6	86.3	40.4	44.5	56.1
LDAM [4]	73.4	76.8	87.0	39.6	45.0	56.9
LDAM-DRW [4]	77.0	80.9	88.2	42.0	46.2	58.7
M2m-ERM [23]	78.3	-	87.9	42.9	-	58.2
M2m-LDAM [23]	79.1	-	87.5	43.5	-	57.6
KCL† [18]	77.6	81.7	88.0	42.8	46.3	57.6
TSC	**79.7**	**82.9**	**88.7**	**43.8**	**47.4**	**59.0**

Table 2. TSC outperforms previous state-of-the-art single-model methods on ImageNet-LT. Previous SOTA results of each class split (many, medium, few, all) are colored with gray. Please note that the KCL accuracy for each class split reported in [18] does not match the reported accuracy on all classes (61.8*0.385+49.4*0.479+30.9*0.136=51.658 which cannot be rounded to 51.5), indicating that their reported results may have a typo. Therefore, we also report the result of our re-implemented KCL (denoted with †), which achieves similar accuracy on all classes but slightly different accuracy on each split.

Methods	Many	Medium	Few	All
OLTR [28]	35.8	32.3	21.5	32.2
τ-norm [19]	56.6	44.2	27.4	46.7
cRT [19]	58.8	44.0	26.1	47.3
LWS [19]	57.1	45.2	29.3	47.7
FCL [18]	61.4	47.0	28.2	49.8
KCL [18]	61.8	49.4	30.9	51.5
KCL †	62.4	49.0	29.5	51.5
TSC	**63.5**	**49.7**	**30.4**	**52.4**

Experiments

Table 3. TSC outperforms previous state-of-the-art single-model methods on challenging **iNaturalist 2018** [37] dataset, which contains 8142 classes. Previous SOTA results for each class split (many, medium, few, all) are colored with gray.

Methods	Many	Medium	Few	All
CE	72.2	63.0	57.2	61.7
CB-Focal	-	-	-	61.1
OLTR [28]	59.0	64.1	64.9	63.9
LDAM + DRW [4]	-	-	-	64.6
cRT [19]	69.0	66.0	63.2	65.2
τ-norm [19]	65.6	65.3	65.9	65.6
LWS [19]	65.0	66.3	65.5	65.9
KCL [18]	-	-	-	68.6
TSC	**72.6**	**70.6**	**67.8**	**69.7**

Thanks