Supplementary material
Similarity Contrastive Estimation for Self-Supervised Soft Contrastive Learning

A. Pseudo-Code of SCE

```
# dataloader: loader of batches of size bsz
# epochs: number of epochs
# T1: weak distribution of data augmentations
# T2: strong distribution of data augmentations
# f1, g1: online encoder and projector
# f2, g2: momentum encoder and projector
# queue: memory buffer
# tau: online temperature
# tau_m: momentum temperature
# lambda_: coefficient between contrastive and relational aspects

for i in range(epochs):
  for x in dataloader:
    x1, x2 = T1(x), T2(x)
    z1, z2 = g1(f1(x1)), g2(f2(x2))

    stop_grad(z2)

    sim2_pos = zeros(bsz)
    sim2_neg = einsum("nc,kc->nk", z2, queue)
    sim2 = cat([sim2_pos, sim2_neg]) / tau_m
    s2 = softmax(sim2)
    w2 = lambda_ * one_hot(sim2_pos, bsz+1) + (1 - lambda_) * s2


    sim1_pos = einsum("nc,nc->n", z1, z2)
    sim1_neg = einsum("nc,kc->nk", z1, queue)
    sim1 = cat([sim1_pos, sim1_neg]) / tau
    p1 = softmax(sim1)

    loss = cross_entropy(p1, w2)
    loss.backward()

    update(f1.params)
    update(g1.params)
    momentum_update(f2.params, f1.params)
    momentum_update(g2.params, g1.params)
    fifo_update(queue, z2)
```

Algorithm 1: Pseudo-Code of SCE in a pytorch style

B. Proof Proposition 1. in Sec. 3.2

Proposition. *L_{SCE} defined as*

$$L_{SCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N} w^2_{ik}\log\left(p^1_{ik}\right),$$

can be written as:

$$L_{SCE} = \lambda \cdot L_{InfoNCE} + \mu \cdot L_{ReSSL} + \eta \cdot L_{ceil},$$

with $\mu = \eta = 1 - \lambda$ and

$$L_{Ceil} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\sum_{j=1}^{N}\mathbb{1}_{i\neq j}\cdot\exp(\mathbf{z_i^1}\cdot\mathbf{z_j^2}/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{z_i^1}\cdot\mathbf{z_j^2}/\tau)}\right).$$

Proof. Recall that:

$$\begin{aligned}
p^1_{ik} &= \frac{\exp(\mathbf{z_i^1}\cdot\mathbf{z_k^2}/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{z_i^1}\cdot\mathbf{z_j^2}/\tau)},\\
s^2_{ik} &= \frac{\mathbb{1}_{i\neq k}\cdot\exp(\mathbf{z_i^2}\cdot\mathbf{z_k^2}/\tau_m)}{\sum_{j=1}^{N}\mathbb{1}_{i\neq j}\cdot\exp(\mathbf{z_i^2}\cdot\mathbf{z_j^2}/\tau_m)},\\
w^2_{ik} &= \lambda\cdot\mathbb{1}_{i=k} + (1-\lambda)\cdot s^2_{ik}.
\end{aligned}$$

We decompose the second loss over k in the definition of L_{SCE} to make the proof:

$$\begin{aligned}
L_{SCE} &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N} w^2_{ik}\log\left(p^1_{ik}\right)\\
&= -\frac{1}{N}\sum_{i=1}^{N}\left[w^2_{ii}\log\left(p^1_{ii}\right) + \sum_{\substack{k=1\\k\neq i}}^{N} w^2_{ik} log\left(p^1_{ik}\right)\right]\\
&= \underbrace{-\frac{1}{N}\sum_{i=1}^{N} w^2_{ii}\left(p^1_{ii}\right)}_{(1)} \underbrace{-\frac{1}{N}\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N} w^2_{ik}\log\left(p^1_{ik}\right)}_{(2)}.
\end{aligned}$$

First we rewrite (1) to retrieve the $L_{InfoNCE}$ loss.

$$\begin{aligned}
(1) &= -\frac{1}{N}\sum_{i=1}^{N} w^2_{ii} log\left(p^1_{ii}\right)\\
&= -\frac{1}{N}\sum_{i=1}^{N}\lambda\cdot\log\left(p^1_{ii}\right)\\
&= -\lambda\cdot\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp(\mathbf{z_i^1}\cdot\mathbf{z_i^2}/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{z_i^1}\cdot\mathbf{z_j^2}/\tau)}\right)\\
&= \lambda\cdot L_{InfoNCE}.
\end{aligned}$$

Now we rewrite (2) to retrieve the L_{ReSSL} and L_{Ceil} losses.

$$\begin{aligned}
(2) &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N} w^2_{ik}\log\left(p^1_{ik}\right)\\
&= -\frac{1}{N}\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N}(1-\lambda)\cdot s^2_{ik}\cdot\log\left(p^1_{ik}\right)\\
&= -(1-\lambda)\cdot\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N} s^2_{ik}\cdot\log\left(p^1_{ik}\right)\\
&= -(1-\lambda)\cdot\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}\left[s^2_{ik}\cdot\log\left(\frac{\exp(\mathbf{z_i^1}\cdot\mathbf{z_k^2}/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{z_i^1}\cdot\mathbf{z_j^2}/\tau)}\right)\right]
\end{aligned}$$

$$= -(1-\lambda) \cdot \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{N} \Bigg[s_{ik}^2 \cdot \Bigg(\log\left(\exp(\mathbf{z_i^1} \cdot \mathbf{z_k^2}/\tau)\right) - \log\left(\sum_{j=1}^{N} \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)\right) \Bigg) \Bigg]$$

$$= -(1-\lambda) \cdot \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{N} \Bigg[s_{ik}^2 \cdot \Bigg(\log\left(\exp(\mathbf{z_i^1} \cdot \mathbf{z_k^2}/\tau)\right) - \log\left(\sum_{j=1}^{N} \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)\right) + \log\left(\sum_{j=1}^{N} \mathbb{1}_{i \neq j} \cdot \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)\right) - \log\left(\sum_{j=1}^{N} \mathbb{1}_{i \neq j} \cdot \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)\right) \Bigg) \Bigg]$$

$$= -(1-\lambda) \cdot \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{N} \Bigg[s_{ik}^2 \cdot \Bigg(\log\left(\frac{\exp(\mathbf{z_i^1} \cdot \mathbf{z_k^2}/\tau)}{\sum_{j=1}^{N} \mathbb{1}_{i \neq j} \cdot \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)}\right) + \log\left(\frac{\sum_{j=1}^{N} \mathbb{1}_{i \neq j} \cdot \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)}\right) \Bigg) \Bigg]$$

$$= -(1-\lambda) \cdot \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{N} \left[s_{ik}^2 \cdot \log\left(\frac{\exp(\mathbf{z_i^1} \cdot \mathbf{z_k^2}/\tau)}{\sum_{j=1}^{N} \mathbb{1}_{i \neq j} \cdot \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)}\right) \right] - (1-\lambda) \cdot \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{N} \left[s_{ik}^2 \cdot \log\left(\frac{\sum_{j=1}^{N} \mathbb{1}_{i \neq j} \cdot \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)}\right) \right].$$

Because $s_{ii}^2 = 0$ and $\mathbf{s_i^2}$ is a probability distribution, we have:

$$\sum_{k=1}^{N} s_{ik}^2 \cdot \log\left(\frac{\exp(\mathbf{z_i^1} \cdot \mathbf{z_k^2}/\tau)}{\sum_{j=1}^{N} \mathbb{1}_{i \neq j} \cdot \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)}\right) = \sum_{\substack{k=1 \\ k \neq i}}^{N} s_{ik}^2 \cdot \log\left(\frac{\mathbb{1}_{i \neq k} \cdot \exp(\mathbf{z_i^1} \cdot \mathbf{z_k^2}/\tau)}{\sum_{j=1}^{N} \mathbb{1}_{i \neq j} \cdot \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)}\right),$$

$$\sum_{k=1}^{N} s_{ik}^2 \cdot \log\left(\frac{\sum_{j=1}^{N} \mathbb{1}_{i \neq j} \cdot \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)}\right) = \log\left(\frac{\sum_{j=1}^{N} \mathbb{1}_{i \neq j} \cdot \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)}\right).$$

Then:

$$(2) = -(1-\lambda) \cdot \frac{1}{N} \sum_{i=1}^{N} \sum_{\substack{k=1 \\ k \neq i}}^{N} \left[s_{ik}^2 \cdot \log\left(\frac{\mathbb{1}_{i \neq k} \cdot \exp(\mathbf{z_i^1} \cdot \mathbf{z_k^2}/\tau)}{\sum_{j=1}^{N} \mathbb{1}_{i \neq j} \cdot \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)}\right) \right] - (1-\lambda) \cdot \frac{1}{N} \sum_{i=1}^{N} \left[\log\left(\frac{\sum_{j=1}^{N} \mathbb{1}_{i \neq j} \cdot \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)}\right) \right]$$
$$= (1-\lambda) \cdot L_{ReSSL} + (1-\lambda) \cdot L_{Ceil}.$$

□

C. Classes to construct ImageNet100

To build the ImageNet100 dataset, we used the classes shared by the CMC [1] authors in the supplementary material of their publication. We also share these classes in Tab. 1.

100 selected classes from ImageNet			
n02869837	n01749939	n02488291	n02107142
n13037406	n02091831	n04517823	n04589890
n03062245	n01773797	n01735189	n07831146
n07753275	n03085013	n04485082	n02105505
n01983481	n02788148	n03530642	n04435653
n02086910	n02859443	n13040303	n03594734
n02085620	n02099849	n01558993	n04493381
n02109047	n04111531	n02877765	n04429376
n02009229	n01978455	n02106550	n01820546
n01692333	n07714571	n02974003	n02114855
n03785016	n03764736	n03775546	n02087046
n07836838	n04099969	n04592741	n03891251
n02701002	n03379051	n02259212	n07715103
n03947888	n04026417	n02326432	n03637318
n01980166	n02113799	n02086240	n03903868
n02483362	n04127249	n02089973	n03017168
n02093428	n02804414	n02396427	n04418357
n02172182	n01729322	n02113978	n03787032
n02089867	n02119022	n03777754	n04238763
n02231487	n03032252	n02138441	n02104029
n03837869	n03494278	n04136333	n03794056
n03492542	n02018207	n04067472	n03930630
n03584829	n02123045	n04229816	n02100583
n03642806	n04336792	n03259280	n02116738
n02108089	n03424325	n01855672	n02090622

Table 1: The 100 classes selected from ImageNet to construct ImageNet100.

D. Data augmentations details for evaluation protocol

The data augmentations used for the evaluation protocol are:

- **training set for large datasets**: random crop to the resolution 224×224 and a random horizontal flip with a probability of 0.5.
- **training set for small and medium datasets**: random crop to the dataset resolution with a padding of 4 for small datasets and a random horizontal flip with a probability of 0.5.
- **validation set for large datasets**: resize to the resolution 256×256 and center crop to the resolution 224×224.
- **validation set for small and medium datasets**: resize to the dataset resolution.

E. Implementation details for pretraining small and medium datasets

Implementation details for small and medium datasets. We use the ResNet-18 encoder and pretrain for

Dataset	τ	$\tau_m = 0.03$	$\tau_m = 0.04$	$\tau_m = 0.05$	$\tau_m = 0.06$	$\tau_m = 0.07$	$\tau_m = 0.08$	$\tau_m = 0.09$	$\tau_m = 0.1$
CIFAR10	0.1	89.93	90.03	90.06	90.20	90.16	90.06	89.67	88.97
CIFAR10	0.2	89.98	90.12	90.12	90.05	90.13	90.09	90.22	**90.34**
CIFAR100	0.1	64.49	64.90	65.19	65.33	65.27	**65.45**	64.89	63.87
CIFAR100	0.2	63.71	63.74	63.89	64.05	64.24	64.23	64.10	64.30
STL10	0.1	89.34	**89.94**	89.87	89.84	89.72	89.52	88.99	88.41
STL10	0.2	88.4	88.23	88.4	88.35	87.54	88.32	88.80	88.59
Tiny-IN	0.1	50.23	51.12	51.41	51.66	**51.90**	51.58	51.37	50.46
Tiny-IN	0.2	48.56	48.85	48.35	48.98	49.06	49.15	49.66	49.64

Table 2: Effect of varying the temperature parameters τ_m and τ on the Top-1 accuracy on small and medium datasets.

200 epochs. Because the images are smaller, and ResNet is suitable for larger images, typically 224×224, we follow guidance from SimCLR and replace the first 7×7 Conv of stride 2 with a 3×3 Conv of stride 1. We also remove the first pooling layer. The strong data augmentation distribution applied is: random resized crop, color distortion with a strength of 0.5, gray scale with a probability of 0.2, gaussian blur with probability of 0.5, and horizontal flip with probability of 0.5. The weak data augmentation distribution is composed of a random resized crop and a random horizontal flip with the same parameters as the strong data augmentation distribution.

We use 2 GPUs for a total batch size of 256. The memory buffer size is set to 4,096 for small datasets and 16,384 for medium datasets. The projector is a 2 fully connected layer network with a hidden dimension of 512 and an output dimension of 256. A batch normalization is applied after the hidden layer. The SGD optimizer is used during training with a momentum of 0.9 and a weight decay of $5e^{-4}$. A linear warmup is applied during 5 epochs to reach the initial learning rate of 0.06. The learning rate is scaled using the linear scaling rule: $lr = initial\_learning\_rate * batch\_size/256$ and then follows the cosine decay scheduler without restart. The momentum value to update the momentum network is 0.99 for small datasets and 0.996 for medium datasets.

F. Temperature influence on small and medium datasets

We made a temperature search on CIFAR10, CIFAR100, STL10 and Tiny-ImageNet by varying τ in $\{0.1, 0.2\}$ and τ_m in $\{0.03, ..., 0.10\}$. The results are in Tab. 2. As for ImageNet100, we need a sharper distribution on the output of the momentum encoder. Unlike ReSSL [2], SCE do not collapse when $\tau_m \rightarrow \tau$ thanks to the contrastive aspect. For our baselines comparison in Sec. 4.2, we use the best temperatures found for each dataset.

References

[1] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *16th European Conference on Computer Vision*, pages 776–794, 2020.

[2] Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Ressl: Relational self-supervised learning with weak augmentation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems*, pages 2543–2555, 2021.