Non-Contrastive Learning Meets Language-Image Pre-Training

Jinghao Zhou Li Dong Zhe Gan Lijuan Wang Furu Wei
Microsoft

A. Methodology Comparison

In this section, we compare different non-contrastive objectives derived from cross-entropy loss with different techniques to avoid model collapse, *i.e.*, to ensure both the *sharpness* and *smoothness* of the representation. We first showcase a general formulation to achieve this end in Algorithm 1, and substantiate different self-supervised techniques proposed in the vision field to our setting. We tuned hyper-parameters for each item to suit best for language-image pre-training. We follow the proposition of non-collapsing representations [2, 3], where

$$||\nabla \overline{H}_\theta(\boldsymbol{p}_i)|| + ||\nabla H_\theta(\overline{\boldsymbol{p}})|| > 0 \quad (1)$$

$\forall i \in [B]$ if the network parameters θ lead to collapsing representation, *i.e.*, $\boldsymbol{p}_j = \boldsymbol{p}_k, \forall j, k \in [B]$. $\boldsymbol{p}_i \in \mathbb{R}^K$ is a K-dimensional probability distribution for i$^{\text{th}}$ instance. In both cases where $\boldsymbol{p}_i = \frac{1}{K}\mathbf{1}_K$ or $\boldsymbol{p}_i \neq \frac{1}{K}\mathbf{1}_K$, the above equation holds such that the unified objectives is immune to collapsing representations.

Experiments are conducted with ViT-B/16 on CC12M for 25 epochs, the results of which are listed in Tab. A1. For Centering, we use additional running variance and scale parameters to uniform the target probability, which yields more stable training compared to the original [8] with a running mean only. For objectives with an explicit function $\text{uniform}(\cdot)$, the asymmetry of the forward pass of the target and predicted probability incurs an inconsistency between training and evaluation, which leads to poor zero-shot performance. While the descent linear probing accuracy of these objectives suggests capabilities to learn representation, they lose a unique advantage for zero-shot recognition. MSN [2] uses the mean entropy maximization regularizer to ease the usage of explicit uniform function, which is essentially the HE term. Comparatively, the joint optimization of EH and HE does not rely on different temperatures for target and predicted probability, creating perfect symmetry between the two branches, and is observed to also improve the training stability (Nan in row 5 *vs.* 37.0% in row 7).

Algorithm 1: PyTorch-like pseudo-code of the unified objective for non-contrastive pre-training.

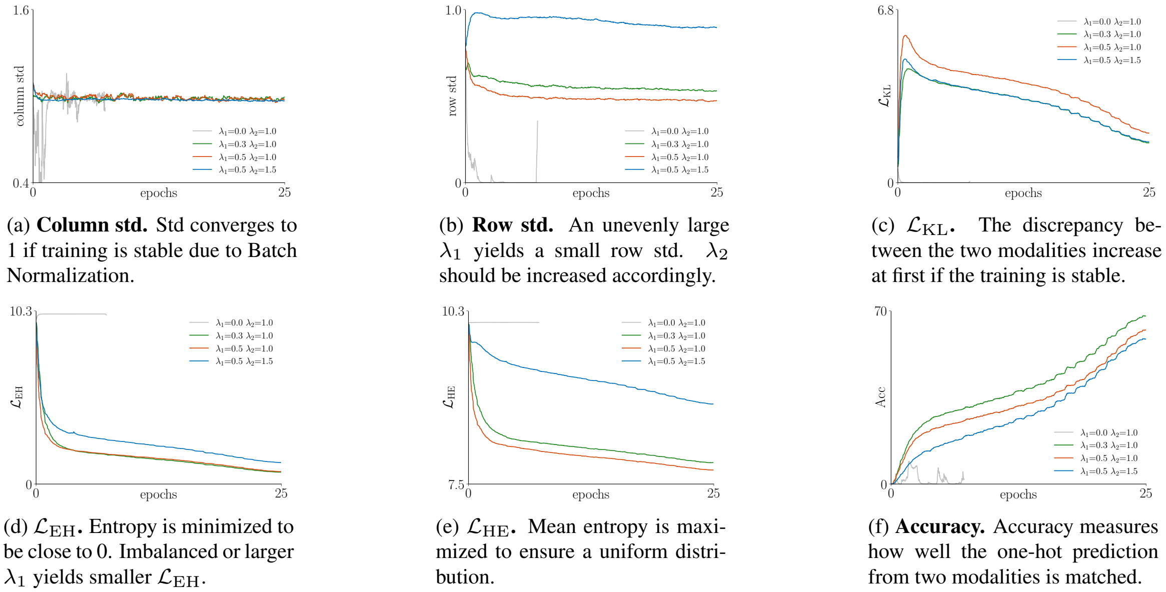
```
// $\boldsymbol{f}_I, \boldsymbol{f}_T$:  image & text projection
// $\boldsymbol{p}_\cdot, \boldsymbol{q}_\cdot$:  target & predicted probability
def $\mathcal{L}_{\text{unified}}$ ($\boldsymbol{f}_I$, $\boldsymbol{f}_T$, $\tau_s$, $\tau$, $\lambda_1$, $\lambda_2$):
    $\boldsymbol{p}_I$ = uniform(softmax($\boldsymbol{f}_I$ / $\tau_s$, dim=1))
    $\boldsymbol{p}_T$ = uniform(softmax($\boldsymbol{f}_T$ / $\tau_s$, dim=1))
    $\boldsymbol{q}_I$ = softmax($\boldsymbol{f}_I$ / $\tau$, dim=1)
    $\boldsymbol{q}_T$ = softmax($\boldsymbol{f}_T$ / $\tau$, dim=1)
    $\mathcal{L}_{\text{CE}}$ = - ($\boldsymbol{p}_I \cdot$ log($\boldsymbol{q}_T$) + $\boldsymbol{p}_T \cdot$ log($\boldsymbol{q}_I$)).sum(dim=1)
            .mean(dim=0)
    $\mathcal{L}_{\text{EH}}$ = - ($\boldsymbol{p}_I \cdot$ log($\boldsymbol{p}_I$) + $\boldsymbol{p}_T \cdot$ log($\boldsymbol{p}_T$)).sum(dim=1)
            .mean(dim=0)
    $\overline{\boldsymbol{p}}_I, \overline{\boldsymbol{p}}_T$ = $\boldsymbol{p}_I$.mean(dim=0), $\boldsymbol{p}_T$.mean(dim=0)
    $\mathcal{L}_{\text{HE}}$ = - ($\overline{\boldsymbol{p}}_I \cdot$ log($\overline{\boldsymbol{p}}_I$) + $\overline{\boldsymbol{p}}_T \cdot$ log($\overline{\boldsymbol{p}}_T$)).sum(dim=1)
    $\mathcal{L}$ = $\mathcal{L}_{\text{CE}}$ + $\lambda_1 \cdot \mathcal{L}_{\text{EH}}$ - $\lambda_2 \cdot \mathcal{L}_{\text{HE}}$
    return $\mathcal{L}$ / 2
```

Method	Sharpness		Smoothness		ZS	LN
	τ/τ_s	λ_1	$\text{uniform}(\cdot)$	λ_2		
CE	1	0	-	0	Nan	
SwAV [7]	1 / 0.25	0	Sinkhorn	1*	27.9	70.3
SCSF [1]	1 / 0.5	0	Batch-Softmax	0	26.5	70.0
DINO [8]	1 / 0.7	0	Centering†	0	22.9	69.1
MSN [2]	1 / 0.8	0	-	1	Nan	
	1 / 0.7	0	-	1	37.4	70.8
nCLIP	1 / 0.8	0.2	-	1.2	37.4	70.0
	1	0.5	-	1.5	37.5	71.0

Table A1. **Comparison over different non-contrastive objectives.** CE denotes vanilla cross-entropy. ZS and LN denote top-1 zero-shot and linear probing accuracy. * The original SwAV [7] implementation sets λ_2 as 0 while it is necessary under our setup to set λ_2 as 1 to avoid collapsing solutions. † Centering is performed in practice before (instead of after) softmax.

B. Additional Ablations

We provide additional ablation studies in this section. Experiments are conducted with ViT-B/16 on CC12M for 25 epochs by default.



(a) **Column std.** Std converges to 1 if training is stable due to Batch Normalization.

(b) **Row std.** An unevenly large λ_1 yields a small row std. λ_2 should be increased accordingly.

(c) $\mathcal{L}_{\mathrm{KL}}$**.** The discrepancy between the two modalities increase at first if the training is stable.

(d) $\mathcal{L}_{\mathrm{EH}}$**.** Entropy is minimized to be close to 0. Imbalanced or larger λ_1 yields smaller $\mathcal{L}_{\mathrm{EH}}$.

(e) $\mathcal{L}_{\mathrm{HE}}$**.** Mean entropy is maximized to ensure a uniform distribution.

(f) **Accuracy.** Accuracy measures how well the one-hot prediction from two modalities is matched.

Figure B1. **Entropy regularizer.** We showcase the training statistics with different λ_1 and λ_2. Setting $\lambda_1 > 0$ is the key to avoiding collapsing solutions. $\lambda_1 = 0.5$ and $\lambda_2 = 1.5$ leads to stable training and optimal downstream performance.

B.1. Entropy Regularizer

In this section, we testify claims made on Sec. 4.4 using pre-training statistics as clues. We monitor the loss scale of $\mathcal{L}_{\mathrm{KL}}$(=$\mathcal{L}_{\mathrm{CE}}$-$\mathcal{L}_{\mathrm{EH}}$), $\mathcal{L}_{\mathrm{EH}}$, and $\mathcal{L}_{\mathrm{HE}}$, as well as the standard deviation (std) of the row and the column of probability matrix $\boldsymbol{P}^{\mathrm{T}} = [\boldsymbol{p}_1^{\mathrm{T}}, \boldsymbol{p}_2^{\mathrm{T}}, ..., \boldsymbol{p}_B^{\mathrm{T}}] \in \mathbb{R}^{B\times K}$. Note that $\mathcal{L}_{\mathrm{CE}}+\lambda_1\cdot\mathcal{L}_{\mathrm{EH}}-\lambda_2\cdot\mathcal{L}_{\mathrm{HE}} = \mathcal{L}_{\mathrm{KL}}+(1+\lambda_1)\cdot\mathcal{L}_{\mathrm{EH}}-\lambda_2\cdot\mathcal{L}_{\mathrm{HE}}$. With the results and statistics given in Tab. 9a and Fig. B1, we present the following discussions:

1) $\boldsymbol{\lambda_1 = 0, \lambda_2 = 1}$ (gray). When using only $\mathcal{L}_{\mathrm{HE}}$, the model fails to converge. The row std decrease to 0 and column std oscillates drastically. $\mathcal{L}_{\mathrm{EH}}$ and $\mathcal{L}_{\mathrm{HE}}$ do not decrease, while $\mathcal{L}_{\mathrm{KL}}$ remains 0. This indicates that the distributions are collapsing to a *constant uniform distribution*.

2) $\boldsymbol{\lambda_1 > 0, \lambda_2 = 1}$ (green & orange). If $\lambda_1 > 0$ throughout training, the column std will be less oscillating at the beginning. Both the row std and $\mathcal{L}_{\mathrm{KL}}$ will increase rapidly. This mitigates the instability at the start. However, the row std will be undesirably low after some iterations, this is due to *dimensional collapse*, where several dimensions will not be favored by any instance.

3) $\boldsymbol{\lambda_1 = \lambda_2 - 1 > 0}$ (blue). This couples $\mathcal{L}_{\mathrm{EH}}$ and $\mathcal{L}_{\mathrm{HE}}$ into one term $\lambda_2 \cdot (\mathcal{L}_{\mathrm{EH}} - \mathcal{L}_{\mathrm{HE}})$, optimizing less on $\mathcal{L}_{\mathrm{KL}}$ while more on non-collapsing condition, which stabilizes the training at the beginning when two modalities can be drastically different. $\mathcal{L}_{\mathrm{KL}}$ arises more rapidly at the beginning. Increasing λ_2 eases the dimensional collapse at the latter part of the training.

B.2. Meeting in One Latent Space

We experiment with another idea where contrastive and non-contrastive objectives meet in one shared latent space instead of two separate latent spaces in a multi-task manner. Specifically, we introduce negative samples to the non-contrastive objective by replacing Eq. (2) to the matrix multiplication form with InfoNCE. Let $\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_B]$ and $\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, ..., \boldsymbol{q}_B]$. We consider the following CE term:

$$\begin{aligned}
\mathcal{L}_{\tilde{\mathrm{CE}}} &= \mathrm{InfoNCE}((\boldsymbol{P}^{\mathrm{T}}\log(\boldsymbol{Q}) + \log(\boldsymbol{P})^{\mathrm{T}}\boldsymbol{Q})/\sigma) \\
&= -\frac{1}{B}\sum_{i\in B}\log\frac{\exp((\boldsymbol{p}_i^{\mathrm{T}}\log(\boldsymbol{q}_i) + \log(\boldsymbol{p}_i)^{\mathrm{T}}\boldsymbol{q}_i)/\sigma)}{\sum_j \exp((\boldsymbol{p}_i^{\mathrm{T}}\log(\boldsymbol{q}_j) + \log(\boldsymbol{p}_i)^{\mathrm{T}}\boldsymbol{q}_j)/\sigma)} \\
&= -\frac{1}{B}\sum_{i\in B}((\boldsymbol{p}_i^{\mathrm{T}}\log(\boldsymbol{q}_i) + \log(\boldsymbol{p}_i)^{\mathrm{T}}\boldsymbol{q}_i)/\sigma - \Phi(Z_{i,\cdot})) \\
&= \mathcal{L}_{\mathrm{CE}}/\sigma - \frac{1}{B}\sum_{i\in B}\Phi(Z_{i,\cdot}). \qquad (2)
\end{aligned}$$

Φ denotes log-sum-exponential $\log\sum_j \exp(x_j)$ and is an approximation of $\max(\cdot)$ function with $\sigma \to 0$. Compared to Eq. (2) that only computes losses over the diagonal terms of the distance matrix, we now also optimize over the off-diagonal terms. This ensures unmatched negative image-text pairs are distracted under the distance metric of cross-entropy, while matched positive pairs are attracted. The final objective is

$$\mathcal{L}_{\mathrm{yCLIP}} = \mathcal{L}_{\tilde{\mathrm{CE}}} + \lambda_1 \cdot \mathcal{L}_{\mathrm{EH}} - \lambda_2 \cdot \mathcal{L}_{\mathrm{HE}} \qquad (3)$$

We would like to see whether the two objectives conflict and can be optimized in on latent space. As shown in

Method	λ_1	λ_2	Acc_C	Acc_nC	ZS	LN
CLIP	-	-	88.3	-	36.8	68.5
nCLIP	-	-	-	59.0	37.5	71.0
yCLIP	0	0	81.6	0.1	38.9	69.9
	0.5	1.5	83.2	0.1	40.2	70.3
	3.0	4.0	66.3	29.3	29.4	68.9
xCLIP	0.5	1.5	89.2	60.2	42.4	72.2

Table B2. **Meeting strategy of CLIP and nCLIP.** They are met in one latent space (yCLIP) via a hybrid objective or separate latent spaces (xCLIP) via multi-tasking of vanilla objectives. Acc_C denotes the accuracy of CLIP to discriminate positive samples across the batch. Acc_nC denotes the accuracy of nCLIP to predict the same one-hot assignment across the channel.

Tab. B2, when $\lambda_1 = \lambda_2 = 0$, $\mathcal{L}_{\text{yCLIP}}$ reduces to $\mathcal{L}_{\text{CLIP}}$ with sole difference on distance metric and incurs a 2.1% and 1.4% performance gain on ZS and LN, respectively. When $\lambda_1 = \lambda_2 - 1 > 0$, $\mathcal{L}_{\text{yCLIP}}$ behaves more similar to $\mathcal{L}_{\text{nCLIP}}$ with scaling and an additional negative term. Setting $\lambda_1 = 0.5,\ \lambda_1 = 1.5$ is insufficient to avoids model collapse as implied by Acc_nC. While increasing the weights of regularizers and setting $\lambda_1 = 3,\ \lambda_1 = 4$ yields seemingly balanced Acc_C and Acc_nC, we observe only a 29.4% zero-shot accuracy with the pre-trained model, validating the contradiction of two objectives in one shared latent space. While a bespoke design may exist, we opt for simple multi-tasking of CLIP and nCLIP in separate spaces.

C. Additional Implementation

Zero-shot classification. We follow the same setup as [28], with prompt engineering for each of the 27 evaluation datasets, including Food-101 [6], CIFAR-10 [23], CIFAR-100 [23], Birdsnap [5], SUN397 [35], Stanford Cars [22], FGVC Aircraft [25], Pascal VOC 2007 Classification [15], Describable Textures [12], Oxford-IIIT Pets [27], Caltech-101 [16], Oxford Flowers 102 [26], MNIST [24], Facial Emotion Recognition 2013 [18], STL-10 [13], EuroSAT [19], RESISC45 [11], GTSRB [30], KITTI [17], Country211 [28], PatchCamelyon [33], UCF101 [29], Kinetics700 [9], CLEVR Counts [20], Hateful Memes [21], Rendered SST2 [28], and ImageNet [14]. The final text embedding is ensembled by averaging all text embeddings with different prompts.

Fine-tuning & semi-supervised learning. We use a training recipe from [4], with a layer-wise learning rate decay rate of 0.65, a weight decay of 0.05, a drop path rate of 0.1, a total epoch of 100, and DeepSpeed. We use `[CLS]` token for classification. We disable relative positional embedding and layer scaling. We sweep over four learning rates $\{3e^{-3}, 4e^{-3}, 5e^{-3}, 6e^{-3}\}$ for all models. For semi-supervised learning with partial data, we find that keeping CLIP's projection head yields better performance, especially with 1% of data. Specifically, the `[CLS]` token is further forwarded to CLIP's image projection head and is classified via a cosine classifier with the temperature learned during pre-training. we initialize the weight of the classifier as the text embedding of ImageNet labels. We do not observe evident gain when fine-tuning on full data.

Model	Data	Supervision	$\mathcal{J}$
random	-	-	25.7
DeiT [31]	ImageNet	class	30.4
MSN [2]	ImageNet	self	38.6
TWIST [34]	ImageNet	self	44.1
iBOT [37]	ImageNet	self	44.1
DINO [8]	ImageNet	self	44.7
MoCoV3 [10]	ImageNet	self	**45.9**
CLIP	IT35M	text	41.2
nCLIP	IT35M	text	43.7
xCLIP	IT35M	text	41.9

Table D3. **Masking probing.** $\mathcal{J}$ denotes Jaccard similarity between predictions and the ground-truth. Models with ViT-B/16 are listed.

Model	Data	mAcc	mIoU
CLIP	IT35M	42.9	24.8
xCLIP	IT35M	**53.1**	**38.4**

Table D4. **Unsupervised segmentation with GroupViT.** Models based on ViT-S are evaluated.

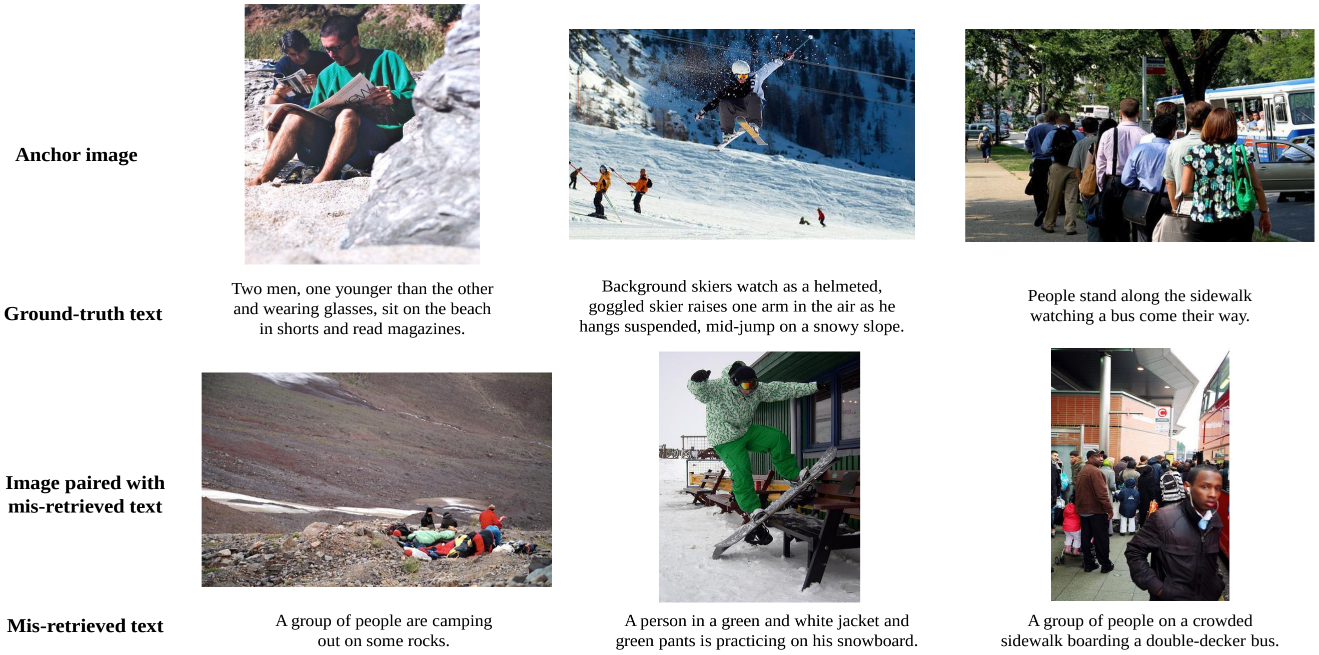
D. Additional Results

D.1. Mask Probing

The results are shown in Tab. D3. For the top panel, we showcase the performance of a range of self-supervised models pre-trained with ImageNet-1K. MoCoV3 [10] achieves a $\mathcal{J}$ of 45.9 points while the supervised baseline [31] lags behind with only a $\mathcal{J}$ of 30.4 points. For the bottom panel, we showcase text-supervised models pre-trained with different objectives. We note that text supervision can also derive explicit boundary scene layout, with all models achieving decent results. Among them, nCLIP performs the best with 43.9 points which is on par with advanced self-supervised models.

D.2. Unsupervised Segmentation with GroupViT

We consider small-size GroupViT [36] under the setting of unsupervised semantic segmentation with PASCAL VOC 2012 [15] dataset. We strictly follow the original setup to train CLIP and a similar recipe to train nCLIP, with only one difference initializing learning rate as $5e^{-4}$. We do not use the multi-label loss as in [36] for simplicity. The results are shown in Tab. D4. xCLIP achieves 38.4 points of



Anchor image

Ground-truth text

Two men, one younger than the other and wearing glasses, sit on the beach in shorts and read magazines.

Background skiers watch as a helmeted, goggled skier raises one arm in the air as he hangs suspended, mid-jump on a snowy slope.

People stand along the sidewalk watching a bus come their way.

Image paired with mis-retrieved text

Mis-retrieved text

A group of people are camping out on some rocks.

A person in a green and white jacket and green pants is practicing on his snowboard.

A group of people on a crowded sidewalk boarding a double-decker bus.

Figure D2. **Failure case analysis** of nCLIP on zero-shot image-to-text retrieval.

mIoU and 53.1 points of mAcc, which is sufficiently better than CLIP baseline, indicating the non-contrastive objective help to learn better object boundaries and scene layouts.

E. Visualization

E.1. Failure Cases of nCLIP in Retrieval

We demonstrate several failure cases of nCLIP (that is correctly retrieved by CLIP) in zero-shot image-to-text retrieval. As shown in Fig. D2, nCLIP tends to mis-retrieve texts containing similar visual elements (*e.g.*, *rocks* in the first column). Their paired images usually contain similar objects, conceptions, and scene layouts. Empirically, nCLIP tends to overlook fine-grained features, *i.e.*, color, attribute, or number but remains generally good sense in predicting high-level semantics. In these senses, the non-contrastive objective serves better as an appending regularizer instead of one single term when downstream tasks solicits direct fine-grained projections as in zero-shot retrieval.

F. Hyper-Parameters

See Tab. F5 for pre-training hyper-parameters.

Hyper-parameter	Value
batch size	4096
training epochs	32
learning rate	$1e^{-3}$
learning rate end	$2e^{-6}$
learning rate scheduler	cosine decay
weight decay	0.2
warm-up epochs	3
optimizer	AdamW
Adam β_1	0.9
Adam β_2	0.98
Adam ϵ	$1e^{-6}$
λ_1	0.5
λ_2	1.5
head arch	4096 - 32768

Table F5. **Pre-training hyper-parameters.**

G. Class Representation

To validate the non-contrastive term helps learning better semantics-meaningful representation, we visualize the t-SNE [32] of ImageNet-1K classes over validation set in Figs. G3 to G6. Specifically, we showcase both the text's embedding and an average of images' embedding for each of the 1000 classes. We use cosine distance and CE as pre-computed metrics for contrastive and non-contrastive objectives, respectively. We run t-SNE with a perplexity of 20 and a learning rate of 200 for 5000 iterations. The non-contrastive objective comparatively leads to more semantic-meaningful clusters. For example, animals with different species are better separated and visually-similar objects are better attracted.

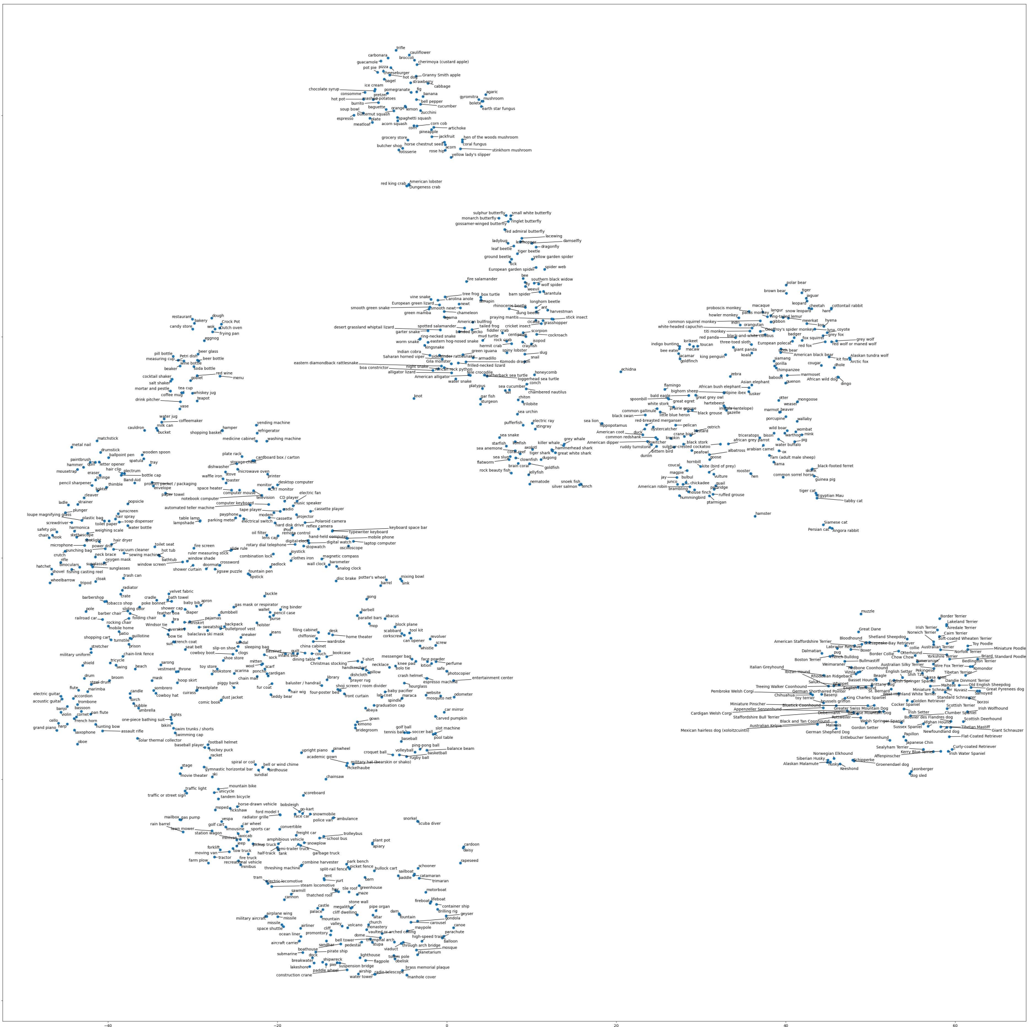


Figure G3. **t-SNE visualization** on image embeddings of CLIP.

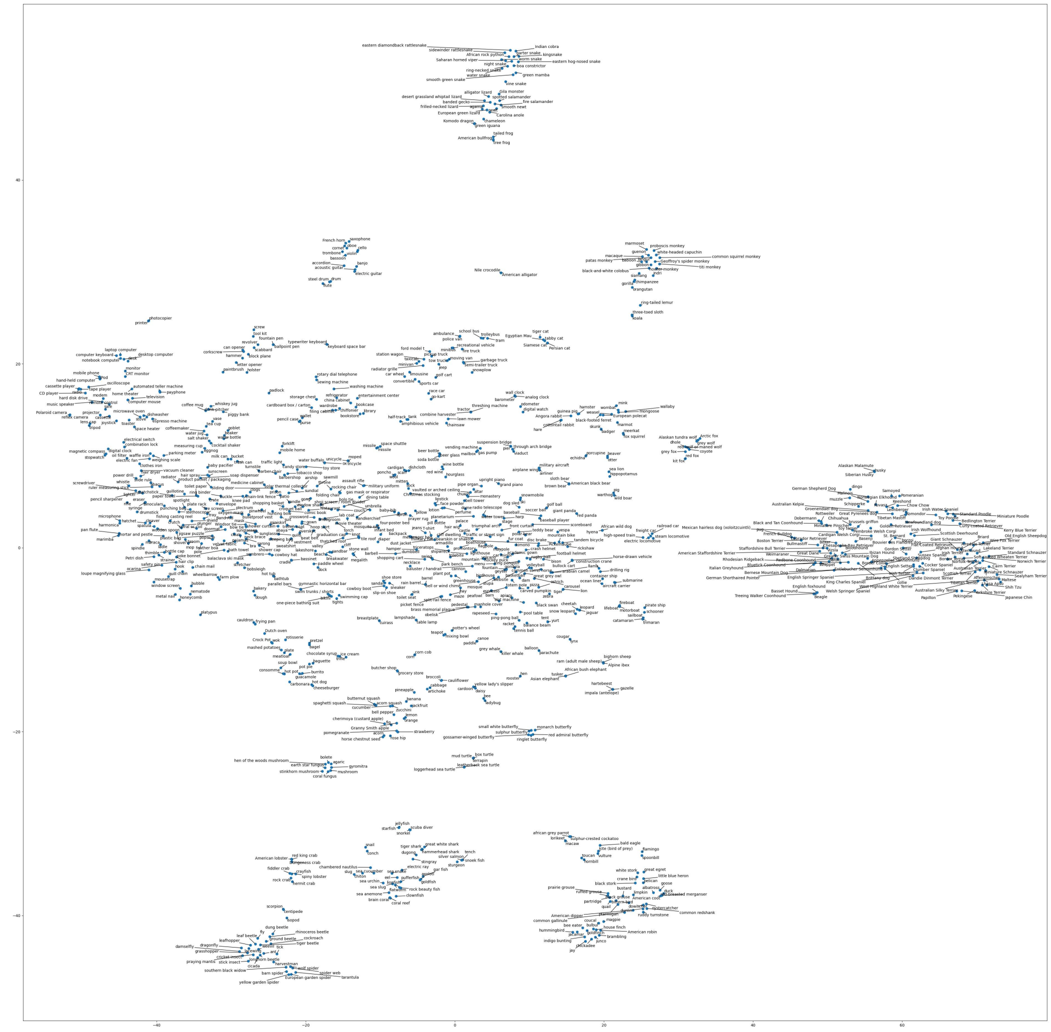


Figure G4. **t-SNE visualization** on image embeddings of nCLIP.

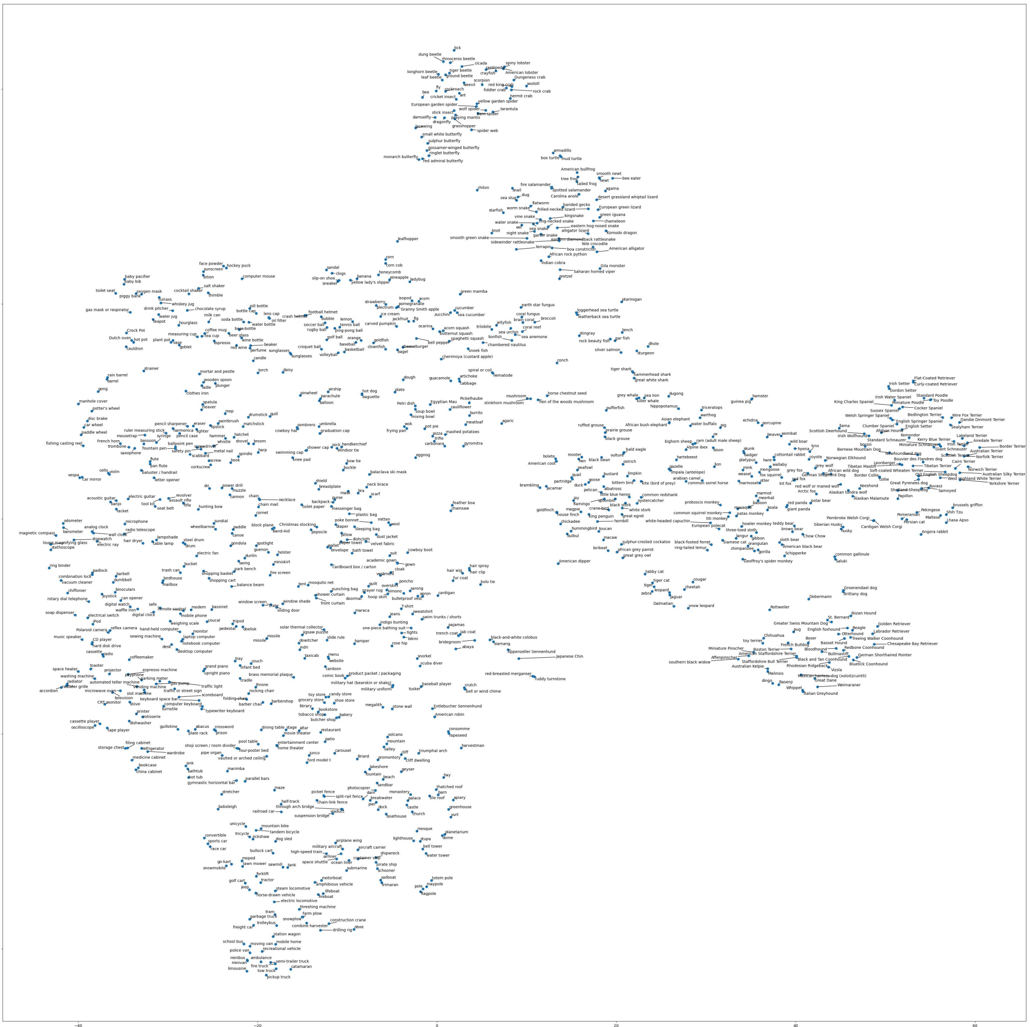


Figure G5. **t-SNE visualization** on text embeddings of CLIP.

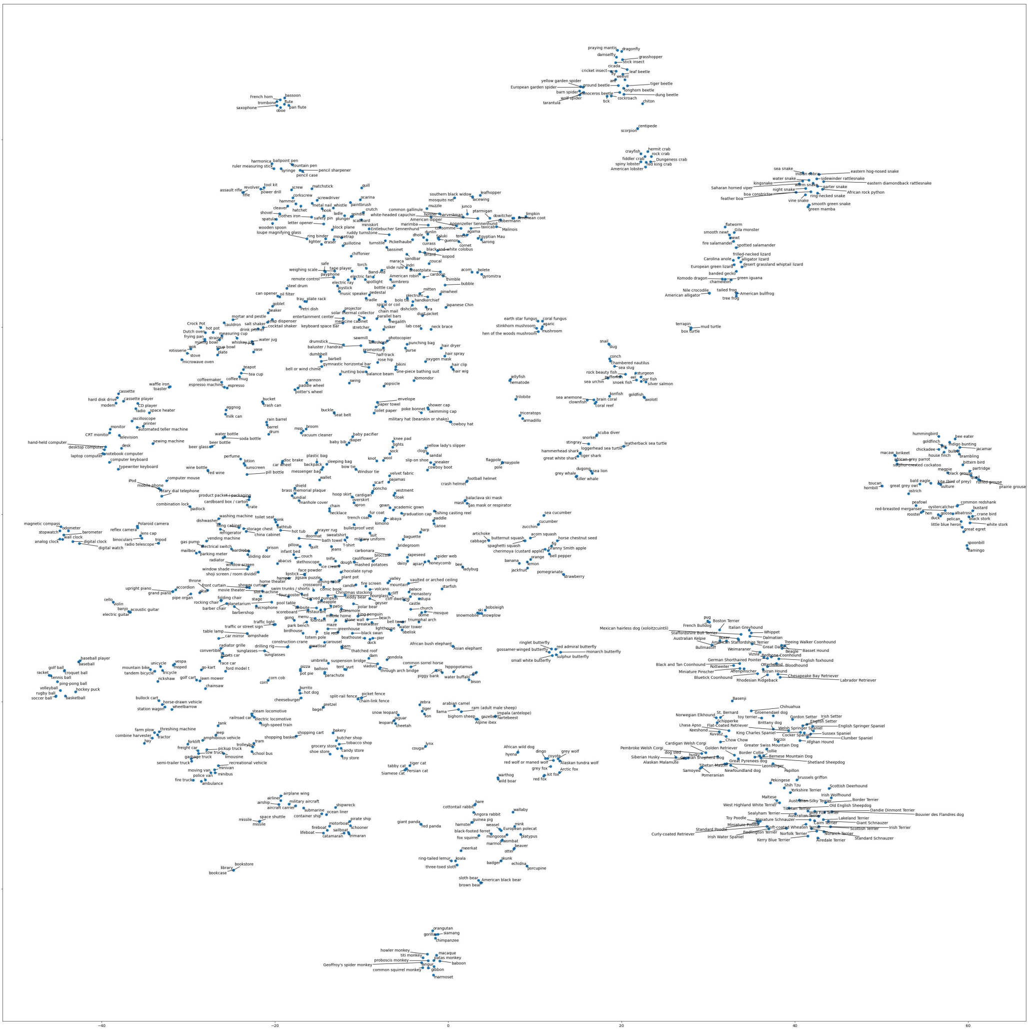


Figure G6. **t-SNE visualization** on text embeddings of nCLIP.

References

[1] Elad Amrani and Alex Bronstein. Self-supervised classification network. In *ECCV*, 2022. 1

[2] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *ECCV*, 2022. 1, 3

[3] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In *ICCV*, 2021. 1

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *ICLR*, 2022. 3

[5] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *CVPR*, 2014. 3

[6] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 3

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 1

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1, 3

[9] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 3

[10] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 3

[11] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE*, 2017. 3

[12] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 3

[13] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *ICAIL*, 2011. 3

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

[15] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 3

[16] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *TPAMI*, 2006. 3

[17] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 3

[18] Panagiotis Giannopoulos, Isidoros Perikos, and Ioannis Hatzilygeroudis. Deep learning approaches for facial emotion recognition: A case study on fer-2013. *Advances in hybridization of intelligent methods*, 2018. 3

[19] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 3

[20] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 3

[21] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeuIPS*, 2020. 3

[22] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, 2013. 3

[23] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. *Citeseer*, 2009. 3

[24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 1998. 3

[25] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 3

[26] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 3

[27] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 3

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[29] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3

[30] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *IJCNN*, 2011. 3

[31] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 3

[32] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 4

[33] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *MICCAI*, 2018. 3

[34] Feng Wang, Tao Kong, Rufeng Zhang, Huaping Liu, and Hang Li. Self-supervised learning by estimating twin class distributions. *arXiv preprint arXiv:2110.07402*, 2021. 3

[35] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 3

[36] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, 2022. 3

[37] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *ICLR*, 2022. 3