Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm

ICLR 2022

Yangguang Li*[1], **Feng Liang*[2]**, Lichen Zhao*[1], Yufeng Cui [1], Wanli Ouyang[3], Jing Shao [1], Fengwei Yu [1], Junjie Yan [1]

* indicates equal contributions

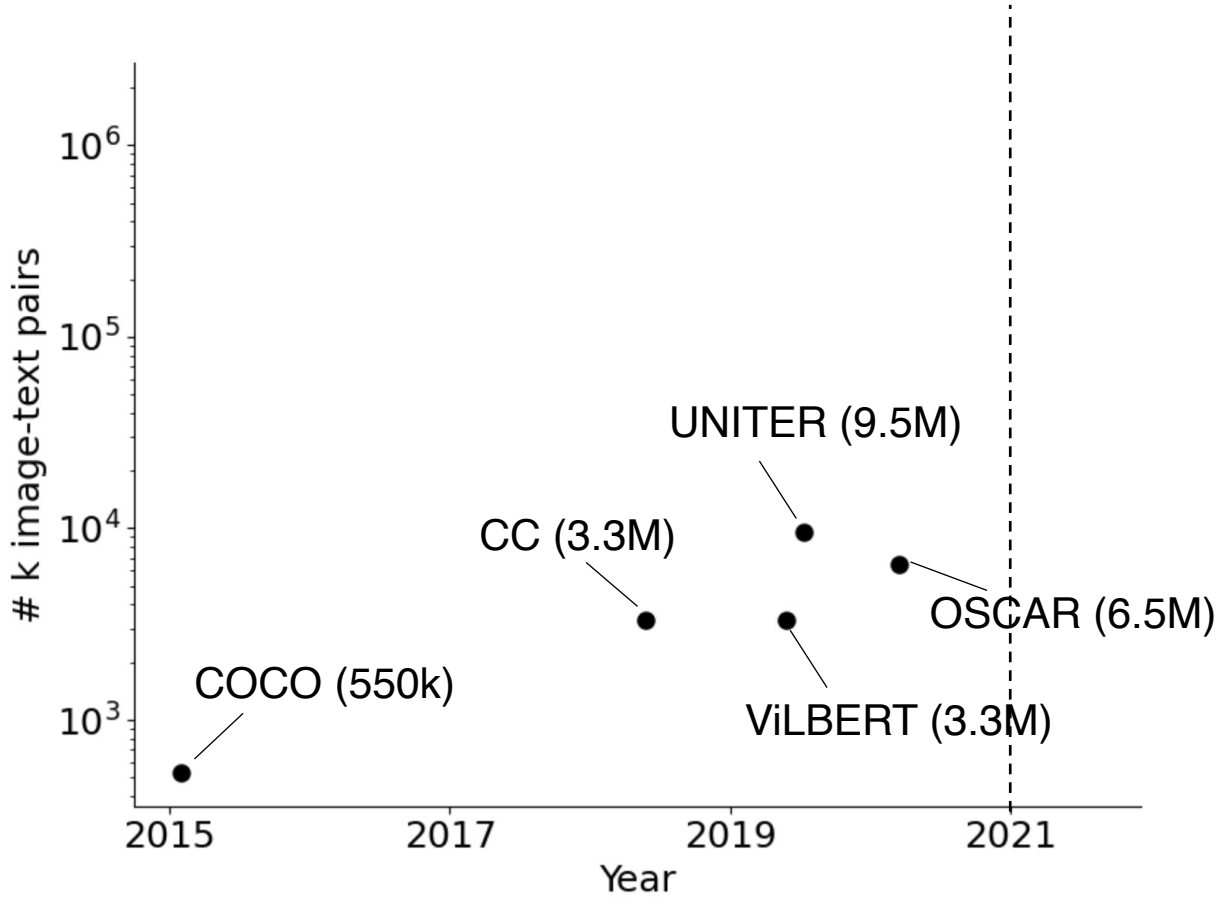
[1]SenseTime Research

[2]UT Austin

[3]The University of Sydney

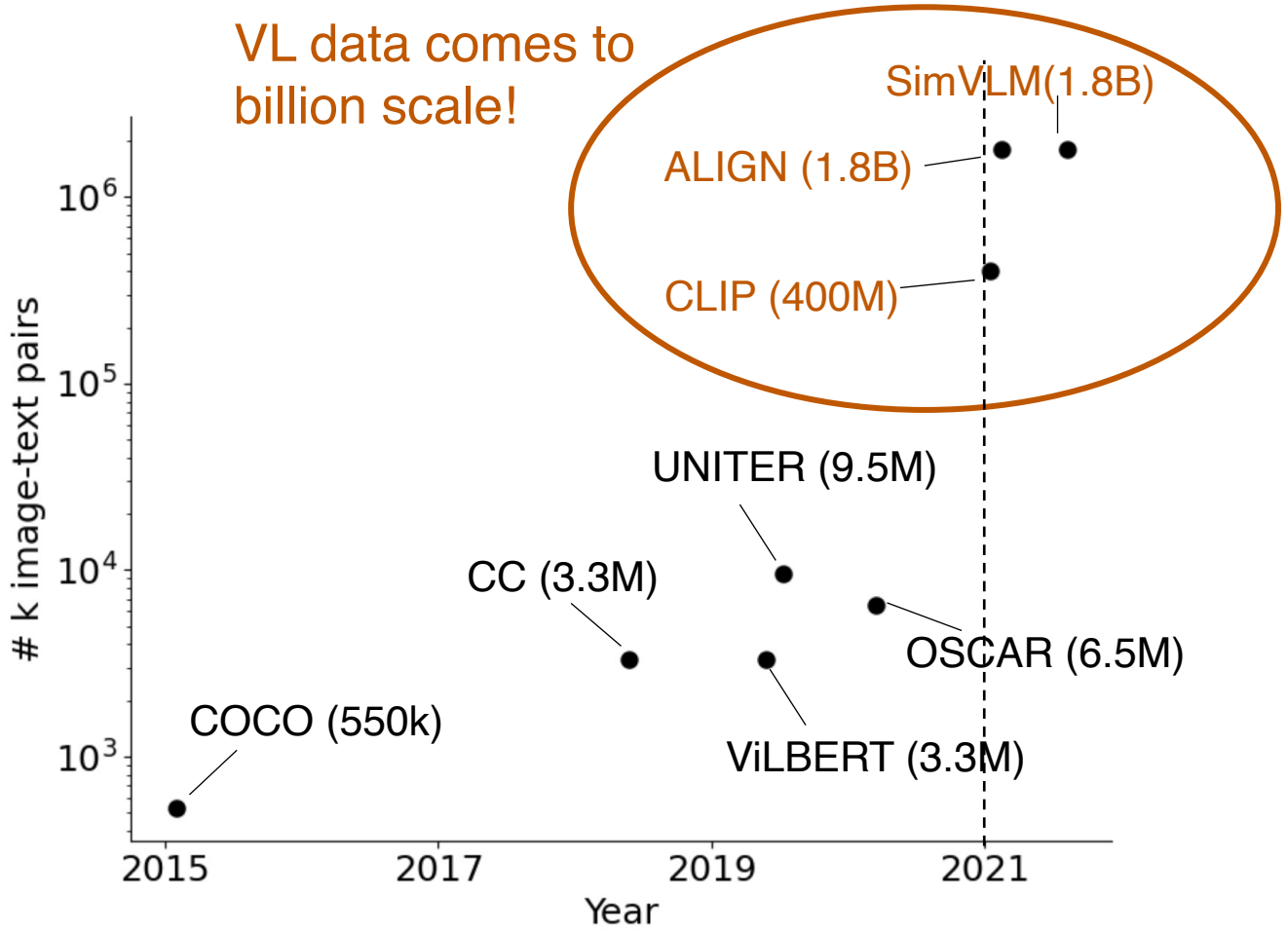
Scaling-up is all you need

Vision-Language Data Size: before 2021, million-scale



Scaling-up is all you need

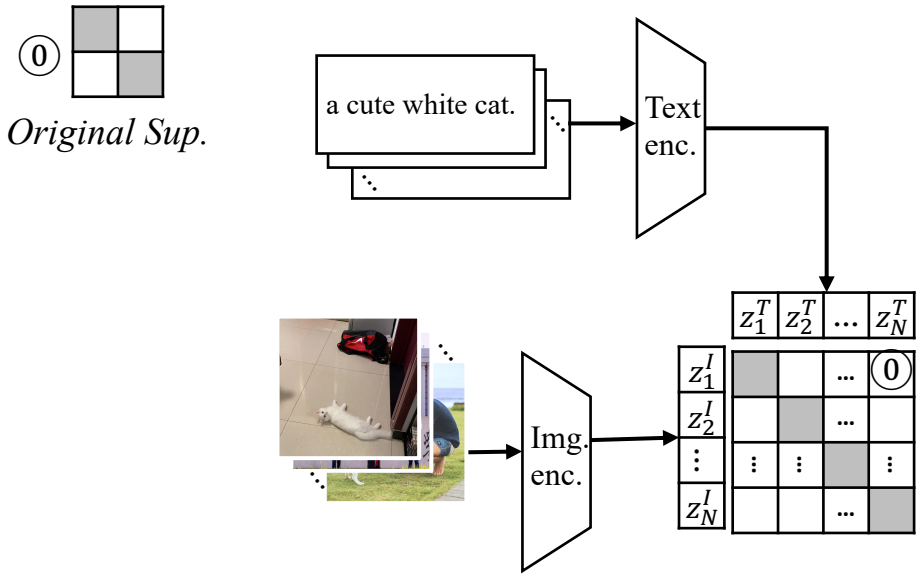
Vision-Language Data Size: after 2021, billion-scale!



Scaling-up is costly

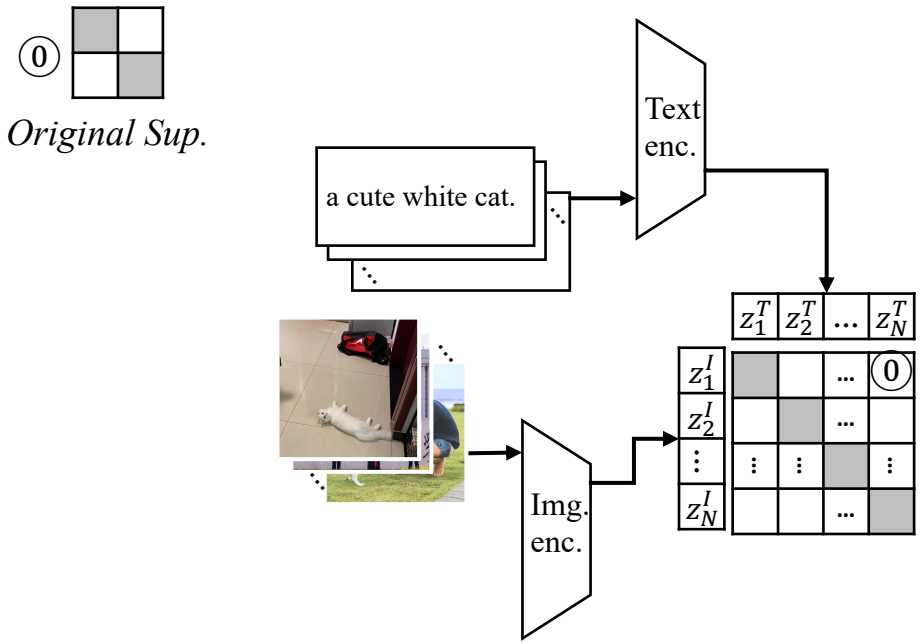
- Motivation: CLIP & ALIGN require huge storage and computing resources, which is not affordable for most laboratories and companies.
- Question: can we training the vision-language model more efficiently?
- Solution: We dig inside into the intrinsic supervision within the image-text data.

CLIP recap



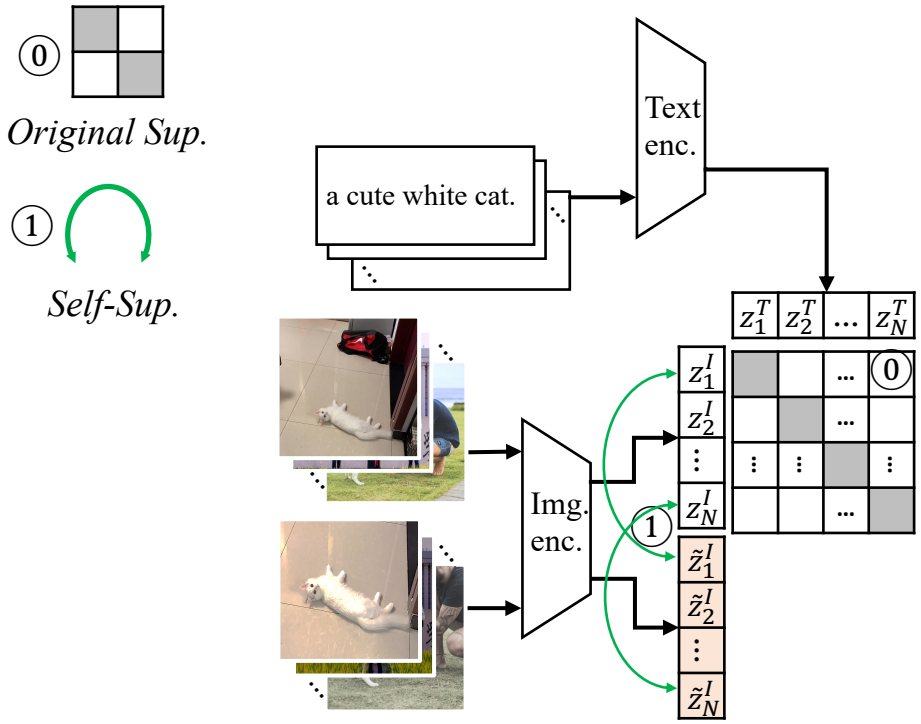
$$L_I = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(\boldsymbol{z}_i^I, \boldsymbol{z}_i^T)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\boldsymbol{z}_i^I, \boldsymbol{z}_j^T)/\tau)}$$

Missing: self-supervision ①



Missing: self-supervision ①

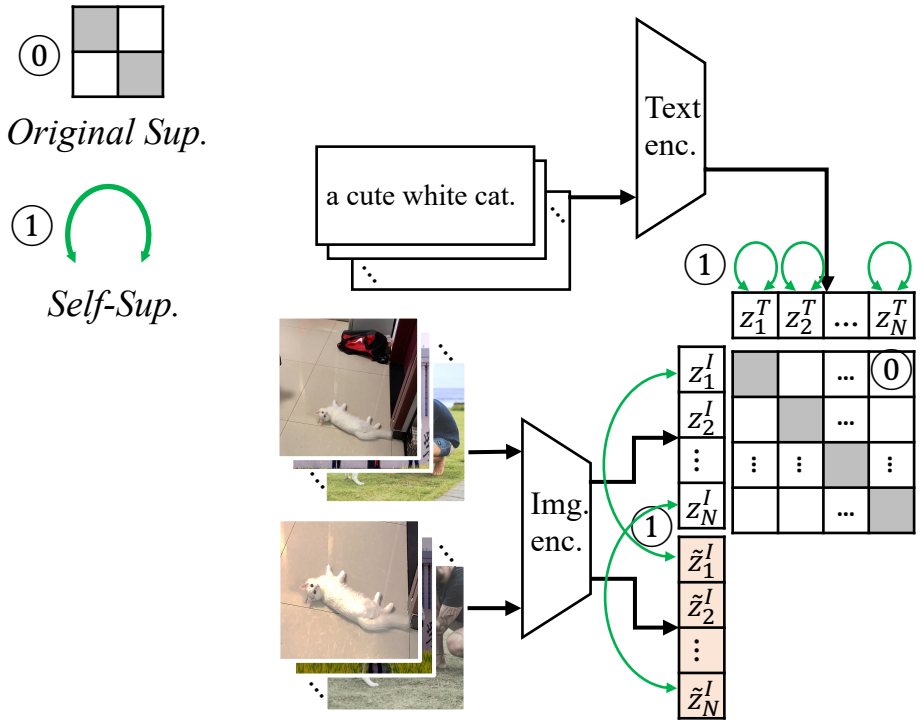
There underlies rich structural information within each modality itself.

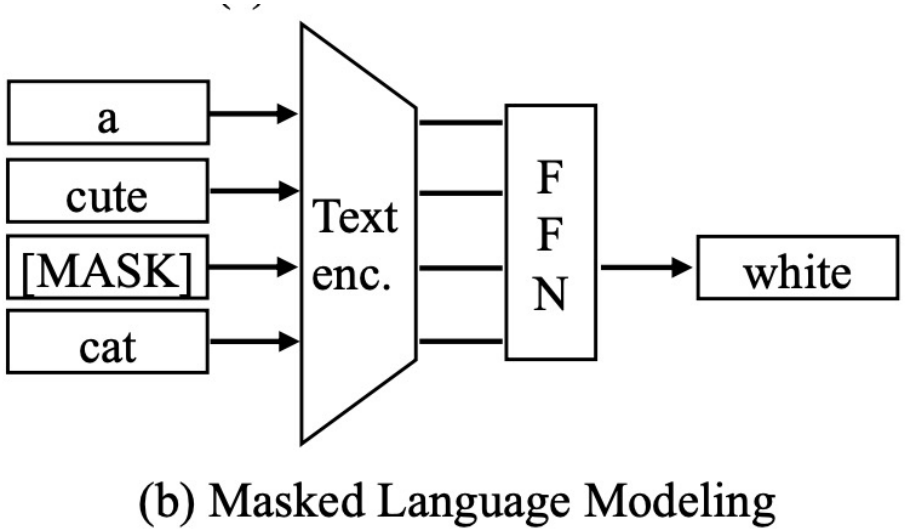


Self-supervised for image: we adopt the simple yet effective SimSiam *[X Chen et al.]* in this paper.

Missing: self-supervision ①

There underlies rich structural information within each modality itself.



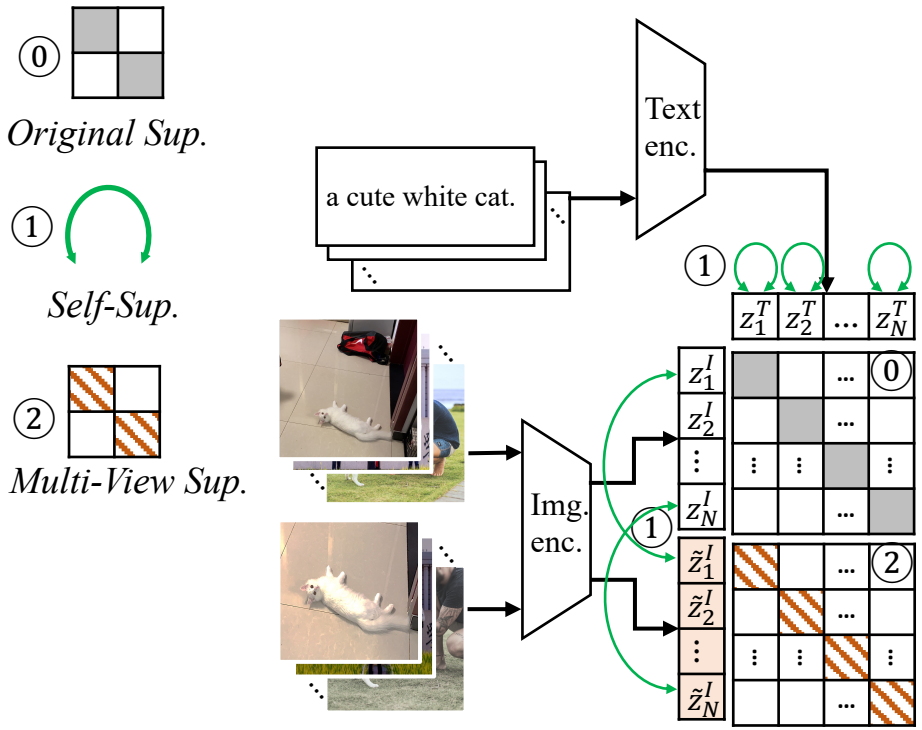


(b) Masked Language Modeling

Self-supervised for text : we adopt the most widely used Masked Language Modeling *[J Devlin et al.]* in this paper.

Missing: multi-view supervision ②

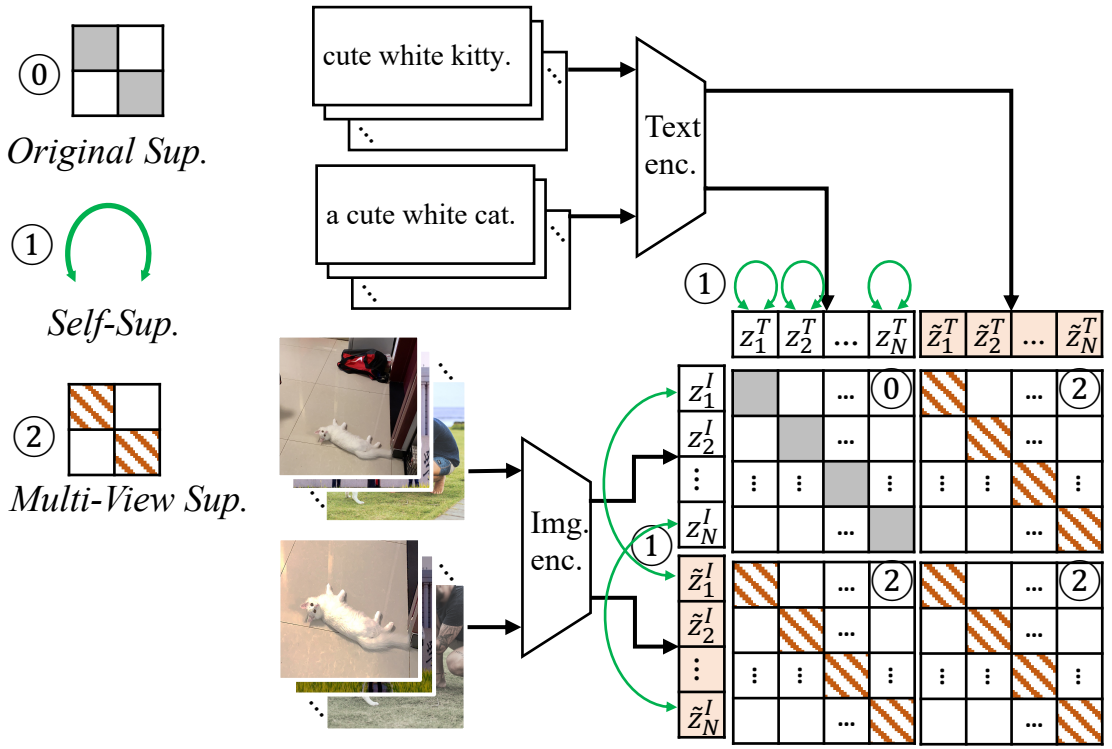
To learn more invariant cross-modal information from multi-view.



One text is paired with two image views, resulting 1x auxiliary supervision. The implementation is compatible with the original CL loss.

Missing: multi-view supervision ②

To learn more invariant cross-modal information from multi-view.



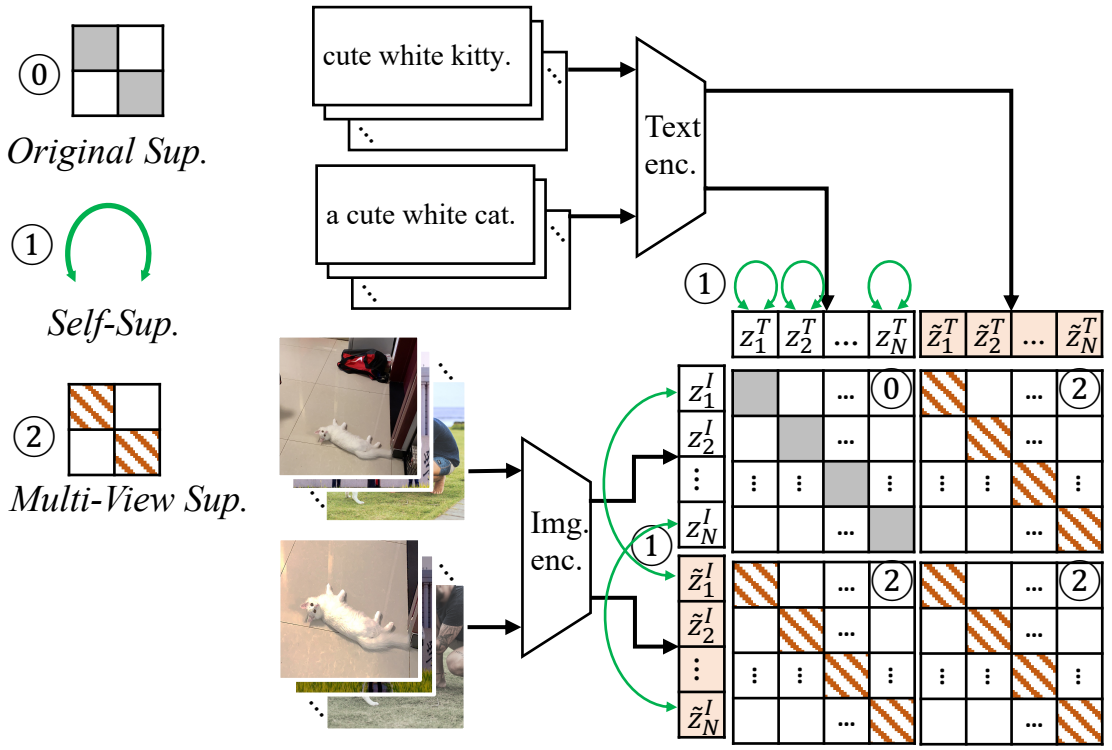
Augmentations of EDA:

- Synonym Replacement (SR):
- Random Insertion (RI):
- Random Swap (RS):
- Random Deletion (RD):

We can further extend the multi-view to texts. The text augmentation is a text classification augmentation EDA *[J Wei et al.].* It does not change the intellectural meaning of a sentence.

Missing: nearest-neighbor supervision ③

One image is very likely to have other caption within the dataset.

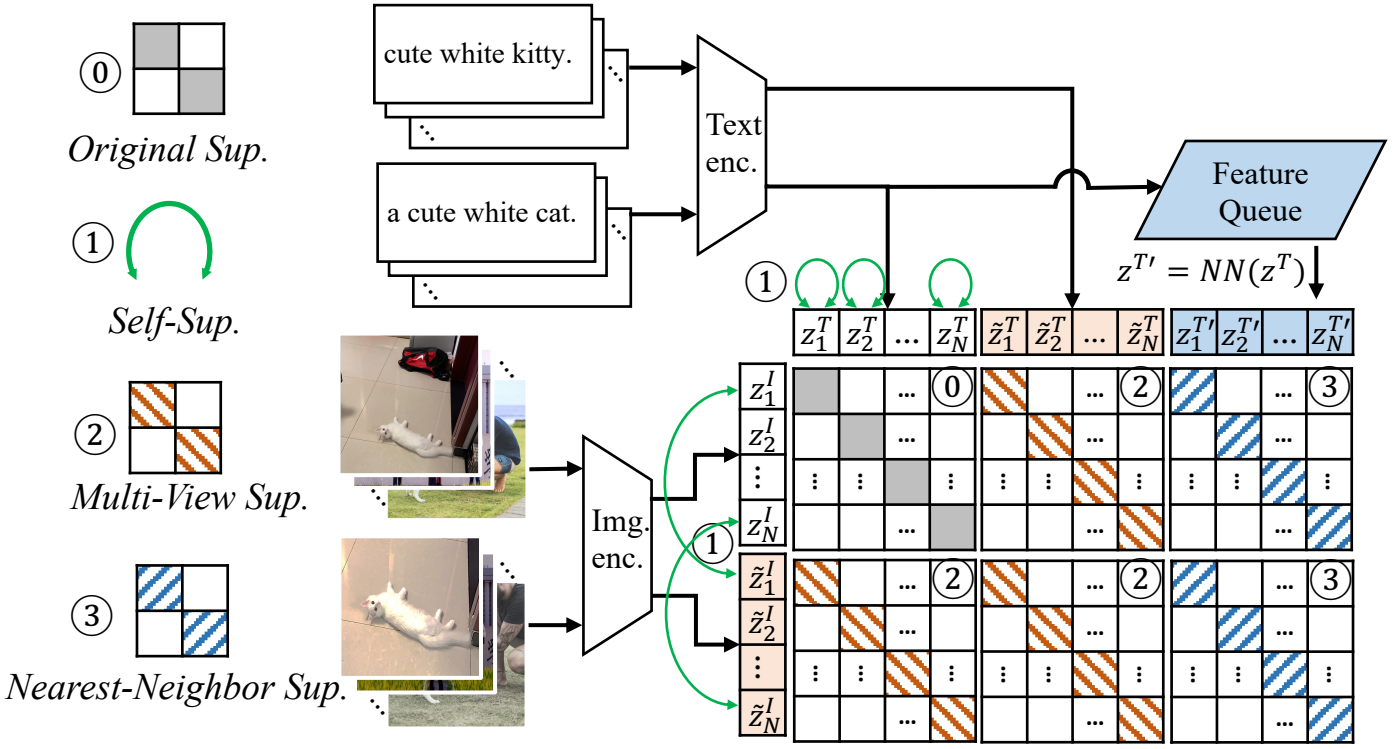


The image-text pairs are collected from the Internet.

Figure 3: Examples of Nearest Neighbor from Conceptual Captions dataset.

Missing: nearest-neighbor supervision ③

One image is very likely to have other caption within the dataset.



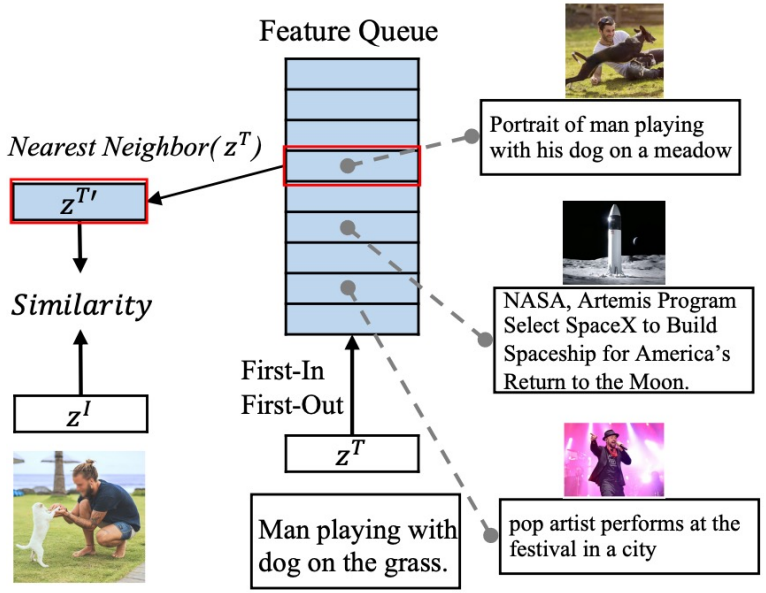
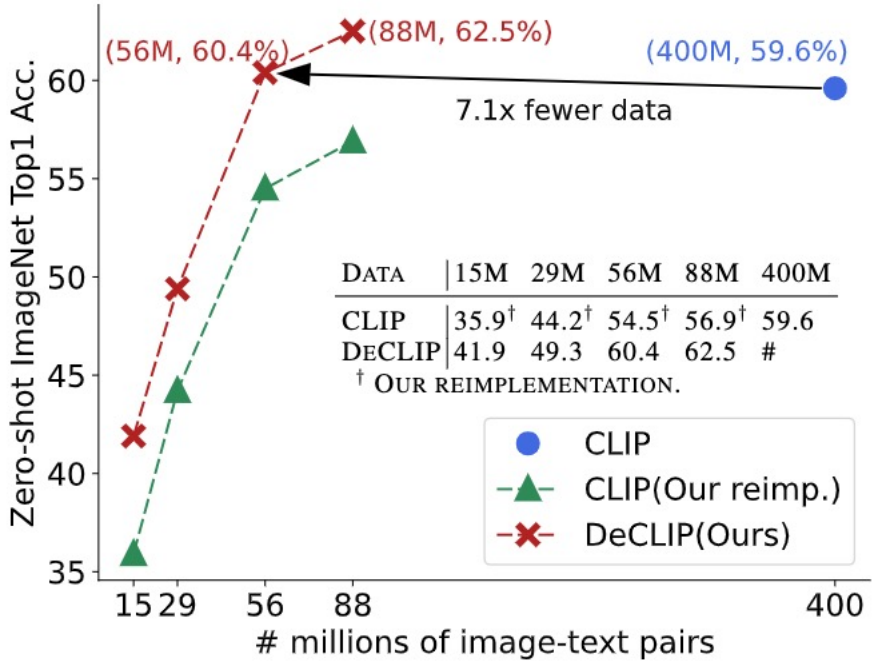


Figure 6: Nearest-Neighbor Supervision. $z^{T'}$ is the NN of feature z^T in the embedding space. $z^{T'}$ will serve as an additional objective for z^I. We use the feature-level nearest neighbor for the text descriptions as the supervision.

The distance between two features could be measured by a simple Euclidean distance. We maintain a FIFO queue Q to simulate the whole data distribution.

Data efficient CLIP(DeCLIP)

Aggregating these three additional supervision leads to our Data efficient CLIP (DeCLIP).



DATA	15M	29M	56M	88M	400M
CLIP	35.9†	44.2†	54.5†	56.9†	59.6
DECLIP	41.9	49.3	60.4	62.5	#

† OUR REIMPLEMENTATION.

Figure 1: Zero-shot performance of CLIP-ResNet50 and our DeCLIP-ResNet50 when using different amounts of data. (88M, 62.5%) denotes the use of 88M data with top-1 accuracy 62.5% on the ImageNet-1K validation dataset. Our model has much better data efficiency.

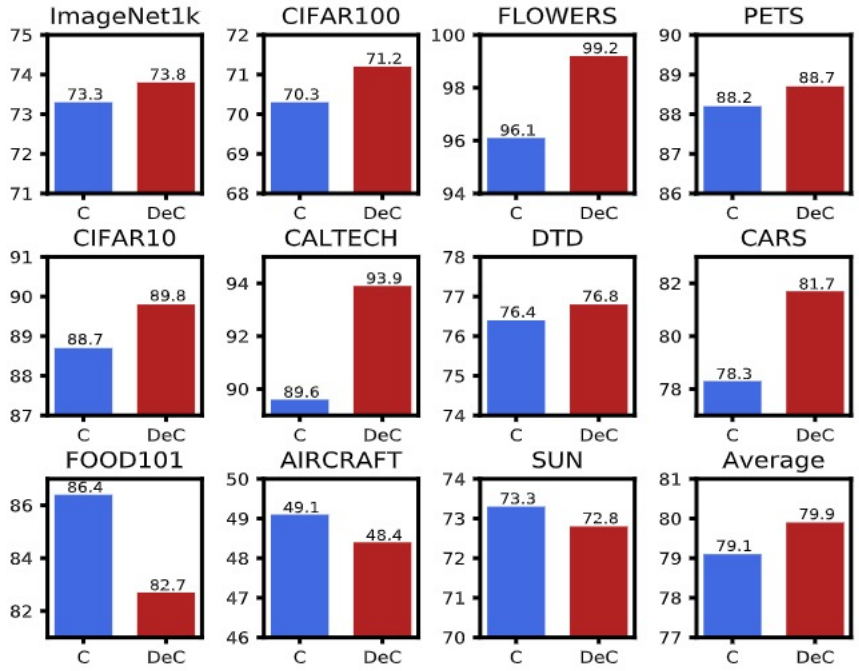


Figure 2: Transfer the DeCLIP-ResNet50 (abbr. as DeC) and CLIP-ResNet50 (abbr. as C) to 11 downstream visual datasets using linear probe verification. Our DeCLIP achieves better results in 8 out of 11 datasets.

Ablation on additional supervision

Table 4: Ablation on additional supervision. SS/MVS/NNS denotes Self-Supervision, Multi-View Supervision and Nearest-Neighbor Supervision, respectively.

CLIP	MVS	SS	NNS	ZERO-SHOT
✓	×	×	×	20.6
✓	✓	×	×	24.8(↑ +4.2)
✓	✓	✓	×	25.4(↑ +4.8)
✓	✓	✓	✓	**27.2(↑ +6.6)**

We Use DeCLIP-ResNet50 on a smaller CC3M dataset for ablation study.

Multi-View Supervision(MVS) boostes amazing 4.2% improvement over the original CLIP.

Self Supervision(SS) might bring more improvements with proper dedicated SSL methods.

NNS further brings 1.8% improvement on the high basis of SS.

Ablation on training cost

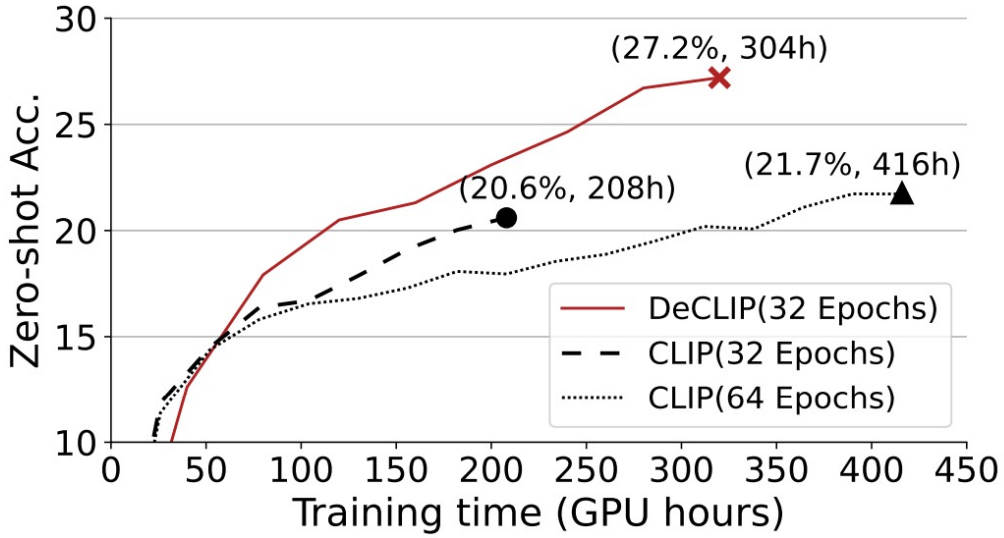


Figure 7: Ablation on pre-training cost on CC3M dataset. The proposed method performs better with less training time.

Since we need to encode twice for each image-text pair, one DeCLIP iteration equals 1.5× CLIP iteration.

Training CLIP longer doesn't bring much gain. It reveals that our framework can extract richer and more representative features through our proposed supervision.

Thank you!

Training codes, models available at:
https://github.com/Sense-GVT/DeCLIP