# SILC: Improving Vision Language Pretraining with Self-Distillation

Muhammad Ferjad Naeem [1⋆] Yongqin Xian [2*] Xiaohua Zhai [3*]
Lukas Hoyer [1,2○] Luc Van Gool [1] Federico Tombari [2,4]
[1] ETH Zurich [2] Google [3] Google Deepmind [4] TU Munich

## Abstract

*Image-Text pretraining on web-scale image caption datasets has become the default recipe for open vocabulary classification and retrieval models thanks to the success of CLIP and its variants. Several works have also used CLIP features for dense prediction tasks and have shown the emergence of open-set abilities. However, the contrastive objective used by these models only focuses on image-text alignment and does not incentivise image feature learning for dense prediction tasks. In this work, we introduce SILC, a novel framework for vision language pretraining. SILC improves image-text contrastive learning with the simple addition of local-to-global correspondence learning by self-distillation. We show that distilling local image features from an exponential moving average (EMA) teacher model significantly improves model performance on dense predictions tasks like segmentation, while also providing improvements on image-level tasks such as classification and retrieval. SILC models sets a new state of the art for zero-shot classification, few shot classification, image and text retrieval and zero-shot segmentation.*

## 1. Introduction.

Recent advancements in self-supervised learning [1, 2, 6, 10] and weakly supervised learning on web data [7, 12, 15] has spearheaded the development of foundational language [4, 11] and vision-language models [7, 12, 15]. These methods get around the long term challenge of obtaining large labelled dataset by developing self-supervision objectives. Developing open vocabulary computer vision models that can reason beyond a pre-determined set of classes has been a long-term challenge. The introduction of web image-text datasets and the progress in compute have enabled significant advances in this field. Popularized by CLIP [12], contrastive pretraining utilizes large datasets with paired image and text from the web and trains a vision-language model (VLM) to embed them to a shared latent space. Since these models are trained on a wide set of concepts, the learned VLM allows for open vocabulary inference [12]. However, developing open vocabulary dense prediction models for segmentation and detection is still an open challenge, since internet-scale datasets do not have dense pixel-level labels. Several works have found that incorporating VLMs in segmentation and detection models can unlock some open vocabulary abilities [3, 5, 8, 13, 14]. Since CLIP is not trained for these tasks, these methods get around its limitations by tuning the learned model with some dense prediction labelled dataset. However, since the contrastive pretraining objective does not explicitly encourage learning good local features for dense prediction tasks, these methods are limited by the VLM's intrinsic performance [10] as we also show later in our experiments.

In the self-supervised literature, enforcing local-to-global consistency by self-distillation has emerged as a powerful pretraining objective [1, 10, 16] to learn vision backbones that are competitive on classification as well as dense prediction tasks, e.g. segmentation and detection. However, these backbones can not directly be used for zero-shot or open vocabulary inference as they do not contain any notion of class or language in the model. In this work, we propose SILC, which combines the advantages of these two branches and unifies image-text contrastive pretraining and local-to-global consistency learning. SILC utilises a web image-text dataset to learn one model that improves VLM performance on existing classification and retrieval tasks while especially improving performance on zero-shot and open vocabulary segmentation, open vocabulary detection, captioning and Visual Question Answering (VQA).

Our contributions are as follows: 1. We propose a novel training framework for VLMs that pairs contrastive pretraining on image-text data with self-distillation on web images. 2. While conceptually very simple, we show that by learning stronger visual features with better local understanding, SILC models offer consistent improvements on multitude of computer vision tasks. These improvements are especially apparent on zero-shot segmentation. 3. We contribute a new foundation model that sets a new state of the art on zero-shot classification, few-shot classification, image-to-text and text-to-image retrieval and zero-shot semantic segmentation.

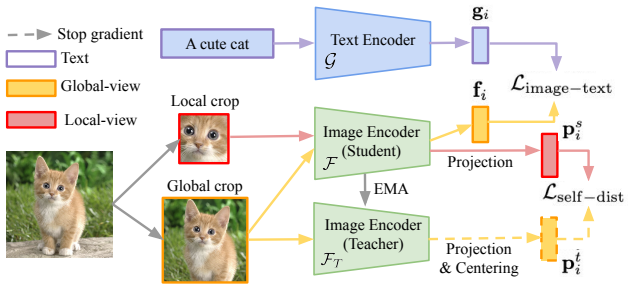


Figure 1. **SILC** is a two-tower transformer based VLM. The first component of our training objective uses a global view of an image covering a large area and its paired caption to optimise a batch-wise contrastive loss for images and texts. The second component of our training objective enforces local-to-global consistency by self-distillation between the main model (the student) and an Exponential Moving Average (EMA)-based teacher. This local-to-global correspondence additionally allows the model to learn good visual features.

## 2. Method.

SILC builds on the contrastive pretraining framework of CLIP [12] and SigLIP [15]. SILC consists of a two-tower transformer model with a shared embedding space. We utilize a web-scale paired image-text dataset and rely on large-scale pretraining to learn the weights of the model. The first component of our pretraining objective focuses on aligning matching image-text pairs close together and away from other images and texts in the batch. This objective has been incredibly successful in recent literature [12, 15]. However, the contrastive objective in its current form does not focus on capturing rich local image semantics necessary for dense prediction tasks like segmentation and detection. Therefore, we propose to pair the contrastive pretraining objective with a local-to-global consistency objective that uses self-distillation as shown in Figure 1. **SILC** gets its name from the two training objectives consisting of **S**elf-Distillation from Images and **I**mage-**L**anguage **C**ontrastive Alignment from Image-Text pairs.

### 2.1. Aligning Image and Text.

The contrastive pretraining objective relies on the Info-NCE framework [9]. It utilizes large amount of web-scale image-text dataset to learn an alignment between paired image and text. Given a minibatch $\mathcal{B} = \{(I_1, T_1), (I_2, T_2), \dots\}$, where $(I_i, T_i)$ denotes a matching pair of image and text, the contrastive objective encourages matching image and text pairs to lie close together in a shared embedding space. The image $I_i$ is processed by a learnable Vision Transformer $\mathcal{F}$ to get its feature embedding. Similarly, the tokenized text $T_i$ is processed by a learnable Text Transformer $\mathcal{G}$ to get its feature embedding. These feature embeddings are normalized by their $l_2$ norm to get $\mathbf{f}_i = \frac{\mathcal{F}(I_i)}{\|\mathcal{F}(I_i)\|_2} \in \mathbb{R}^J$ for the image $I_i$ and $\mathbf{g}_i = \frac{\mathcal{G}(T_i)}{\|\mathcal{G}(T_i)\|_2} \in \mathbb{R}^J$ for the paired text $T_i$ where $J$ is the feature dimension of the shared embedding space. The dot product of $\mathbf{f}_i$ and $\mathbf{g}_i$ computes their cosine similarity and is optimized with a pair of cross-entropy losses as proposed by CLIP [12] or a sigmoid loss as proposed by SigLIP [15]. The batch-wise contrastive losses of CLIP/ SigLIP, represented as $\mathcal{L}_{\text{image-text}}$, rely on a large batch size to align image-text pairs. This objective tuned over a large amount of data learns a shared embedding space between image and text and thus can be used for zero-shot transfer to multitude of computer vision tasks.

### 2.2. Distilling Local Image Features.

Enforcing local-to-global consistency has emerged as a powerful technique to accomplish this on large unlabelled image data [1, 10, 16] in self-supervision literature. However, these methods can not be directly used for open vocabulary models as they are trained without any language information. In the second component of our training framework, we take inspiration from this subset of literature and additionally add local-to-global consistency as a training objective for images in our image-text dataset.

We add this criterion for the image encoder $\mathcal{F}$. We add a projection as a learnable MLP on top of the image encoder to map from the original shared embedding space of dimension $J$ to $K$ where $K > J$. The student $\mathcal{F}_S$ is the main image encoder with a learnable projection head. Since we rely on noisy web scale image-text data, we do not have an oracle teacher for the student to match. We therefore construct our teacher $\mathcal{F}_T$ as a exponential moving average of the student $\mathcal{F}_S$ from the previous training iterations to realize our self-distillation framework:

$$\mathcal{F}_T \leftarrow \lambda \mathcal{F}_T + (1 - \lambda)\mathcal{F}_S, \tag{1}$$

where $\lambda$ controls the update step of the teacher. For a given image $I_i$, the teacher processes its global crop to produce $\mathbf{p}_i^t \in \mathbb{R}^K$ and the student processes its local crop to produce $\mathbf{p}_i^s \in \mathbb{R}^K$. To prevent the teacher from collapsing to a trivial solution, we apply sharpening on the outputs of teacher with $\tau_t$ and student with $\tau_s$. To encourage each feature dimension to contribute to the output feature, we additionally introduce a centering operation on the prediction of the teacher. The centering term $\mathbf{c} \in \mathbb{R}^K$ is initialized with 0 and is updated by a momentum update with a factor of $m$ with the first order batch statistics of the teacher's prediction at each step as follows: $\mathbf{c} \leftarrow m\mathbf{c} + (1 - m)\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|} \mathbf{p}_i^t$.

To learn local-to-global correspondences, the student is faced with an information asymmetry. The student is given a local view of an image which is realized as a random crop over a small region of the image. The teacher, however, has access to a global view of the image containing more information about the scene. The student is tasked

with matching the semantics of the teacher while only having partial information. This is realized as a knowledge-distillation loss where the student and the teacher's feature vectors are first converted to a probability distribution by applying a softmax on the teacher prediction $\mathcal{P}_t(I_i^{gl}) = \texttt{softmax}((\mathbf{p}_i^t - \mathbf{c})/\tau_t)$ and student prediction $\mathcal{P}_s(I_i^{lc}) = \texttt{softmax}(\mathbf{p}_i^s/\tau_s)$. The student is optimized to match the teacher with a cross-entropy loss,

$$\mathcal{L}_{\text{self}-\text{dist}} = -\mathcal{P}_t(I_i^{gl})^\intercal \log(\mathcal{P}_s(I_i^{lc})). \tag{2}$$

While this objective has been explored in self-supervised learning [1, 10], to the best of our knowledge, we are the first work to show its complimentary nature to image-text contrastive learning on web-scale dataset. We show that when combined with text, this objective allows the model to develop a local understanding of the semantics of an image grounded in language. We find two important modifications compared to previous works that allows it to be complimentary to image-text contrastive learning. 1. Each global view used in $\mathcal{L}_{\text{self}-\text{dist}}$ needs to be aligned with text, otherwise the two objectives diverge. This is realized by computing the image-text contrastive loss for each global view while maintaining the same batch size. 2. The momentum scheduler of the EMA should not converge to 1.0. Otherwise the teacher stops learning from image-text loss as the update step becomes too small in the later stage of the training. We therefore use a fixed momentum.

## 3. Experiments.

We compare our SILC pretraining framework with both CLIP [12] and SigLIP [15] on the same test bench and perform extensive experimentation. SILC models based on the CLIP objective are represented by **SILC-C** and the SigLIP versions are represented by **SILC-S**. We show that SILC sets a new state of the art on a variety of tasks: zero-shot classification, few-shot classification, retrieval and zero-shot segmentation.

**Classification and Retrieval.** We compare our pretraining framework with CLIP and SigLIP under the same training and evaluation protocol in Table 1. We compare at ViT/B16 and see that the introduction of self-distillation to both consistently improve their performance on zero-shot classification and retrieval. On zero-shot classification on ImageNet, SILC-C* improves on CLIP (WebLI) by 1.2 points, similarly we notice an improvement of 2.6 points on CIFAR-100 showing the benefit of local feature self-distillation. We make similar observation on retrieval where SILC-C* shows improvements on image to text as well as text to image retrieval. Moving to SigLIP versions of the model, we see a similar trend where the introduction of self-distillation objective allows SILC-S* to consistently improve almost all metrics over the evaluated tasks. We therefore conclude that capturing better local semantics results in learning stronger visual features which also helps tasks that require global understanding of the image. Comparing SILC* models with SILC, we notice that the finetuning on the cleaner subset unlocks additional performance for the model without significant extra training. SILC models set a new state-of-the-art for these tasks at ViT/B16 model size.

**Zero-Shot Semantic Segmentation.** Zero-shot semantic segmentation aims to measure the grounding performance of a VLM usually from its patch embeddings. Comparing against CLIP and SigLIP, we see that both SILC-C* and SILC-S* show significantly superior zero-shot semantic segmentation performance. In fact, both variants achieve multiple mIOU points improvements over all 5 datasets. This validates our hypothesis that the combination of image-text contrastive learning and local-to-global correspondence learning allows the model to develop better understanding of local semantics of the image grounded in language. From Table 1, we observe that the CLIP objective in general results in superior zero-shot segmentation than the SigLIP objective. This is also apparent as we compare SILC-C* with SILC-S*. Moreover, we observe that finetuning on a cleaner subset further improves the zero-shot segmentation performance of both SILC-C and SILC-S. We observe that the CLIP variant SILC-C also outperforms SILC-S here.

**Ablation on Model Components.** We ablate on the various design choices of our model and their impact on various tasks. We train all models for 5 Billion example-seen and report the performance in Table 2. Since our method processes additional image augmentations in the contrastive loss, we first test if our improvements are a consequence of processing more augmentations. We observe that the introduction of additional image augmentations (second row) improve the classification and retrieval metrics but their impact on zero-shot segmentation is not as significant. When we add an EMA over this model's weights similar to our model (third row), we notice a slight improvement as seen in previous SSL literature. Finally when we add the self-distillation from local crops, we see an improvement across the board on all tasks. In particular, we observe the strongest improvement on segmentation tasks highlighting our proposal's impact on them.

## 4. Conclusion.

We propose to integrate local-to-global correspondence learning by self-distillation as a complementary objective to the popular VLM contrastive objective of CLIP [12] and SigLIP [15]. We show that the introduction of this results in remarkable performance improvements on several computer vision tasks. We see a consistent performance improvement on zero-shot classification and retrieval. We fur-

| Model | Zero-Shot Classification | | Zero-Shot Segmentation | | | | | Retrieval COCO | |
|---|---|---|---|---|---|---|---|---|---|
| | ImageNet | CIFAR100 | A-150 | PC-59 | CityScapes | VOC-20 | COCO-stuff | I2T@1 | T2I@1 |
| CLIP (WebLI) [15] | 74.1 | 68.4 | 15.0 | 24.0 | 22.6 | 69.5 | 15.0 | 61.7 | 43.9 |
| **SILC-C* (Ours)** | 75.3 | 71.0 | 17.2 | 29.3 | 25.1 | 73.5 | 18.2 | 62.5 | 44.9 |
| **SILC-C (Ours)** | **76.2** | **72.3** | **19.3** | **31.6** | **26.9** | **77.5** | **20.8** | **66.1** | **49.1** |
| SigLIP [15] | 75.1 | 69.8 | 13.6 | 22.9 | 20.8 | 64.7 | 13.4 | 62.6 | 44.9 |
| **SILC-S*(Ours)** | 75.8 | 69.2 | 16.7 | 28.6 | 23.4 | 72.1 | 17.3 | 63.0 | 44.6 |
| **SILC-S(Ours)** | **76.6** | **70.6** | **18.6** | **30.9** | **25.2** | **76.3** | **19.7** | **66.2** | **48.7** |

Table 1. **Comparing SILC* with baselines**, we observe that our pretraining framework results in a significant improvement over both CLIP and SigLIP objectives. We further finetune SILC* on a cleaner subset to get our final model SILC and see that it unlocks additional performance without significant extra retraining. The best performance for each variant is **bolded**, the second best is underlined.

| Model | IM0shot | COCO Retrieval | | ZS Segmentation | | |
|---|---|---|---|---|---|---|
| | T1 | I2T@1 | T2I@1 | A-150 | Stuff | PC-59 |
| CLIP (WebLI) | 71.7 | 59.1 | 42.9 | 11.8 | 12.9 | 20.1 |
| + additional views | 73.6 | 60.6 | 43.2 | 11.7 | 13.0 | 20.0 |
| + EMA | 73.7 | 61.3 | 43.1 | 11.9 | 13.3 | 20.5 |
| + Self Dist (**SILC-C***) | **74.3** | **62.7** | **43.9** | **12.2** | **15.3** | **21.1** |

Table 2. **We ablate over each component** of our model to verify our design choices. The addition of image augmentation and EMA to CLIP (WebLI) improves classification and retrieval metrics while only slightly impact the segmentation. Adding local-to-global consistency by self-distillation, we observe an improvement across the board especially on segmentation metrics.

ther test our VLM on zero-shot segmentation and show that our training framework results in significant improvements without using any dense ground truth.

## References


[1] 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, 2, 3

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1

[3] Seokju Cho, Heeseong Shin, Sunghwan Hong, Seungjun An, Seungjun Lee, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2303.11797*, 2023. 1

[4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 1

[5] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, 2022. 1

[6] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 2020. 1

[7] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1

[8] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *ICLR*, 2023. 1

[9] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[10] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2, 3

[11] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. In *OpenAI*, 2018. 1

[12] 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 *ICLR*, 2021. 1, 2, 3

[13] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *ECCV*. Springer, 2022. 1

[14] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *NeurIPS*, 2023. 1

[15] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 1, 2, 3, 4

[16] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *ICLR*, 2022. 1, 2