Supplementary for 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

1 Detail about supplementary.

In this supplementary, we provide the following additional experiments and detail about our work.

- Table 1: Comparison with open source CLIP variants on Classification and Retrieval.
- Table 2: Comparison of SILC and baselines trained on open-source dataset.
- Section 2: Classification and Retrieval performance across different model size.
- Section 3: Performance of Teacher vs Student.
- Section 4: SILC applied to CNN backbone.
- Section 5: Additional Results on Zero-Shot Semantic Segmentation including PACL [14].
- Section 6.1: Importance of aligning all global views with text.
- Section 7: Additional Qualitative Results.
- Table 8: Comparison between SILC and MaskCLIP [7]
- Section 8: Additional details about evaluation, training and limitations.

2 Classification and Retrieval performance of additional SILC models.

In our main manuscript Table 1, we show that SILC models improve on CLIP and SigLIP at ViT/B16 size. We additionally train CLIP (WebLI), SILC-C* and SILC-C with ViT/L16 to show that our improvements are consistent at larger model size too. We also train SILC-C* and SILC-C at ViT/B8 to study the trade off between model size vs patch size. We report the results in Table 3. Comparing SILC-C* with CLIP (WebLI) at ViT/L16, we observe that our model consistently improves over the baseline to set a new state-of-the-art at this model size too. SILC-C* achieves a 1.3 points improvement over CLIP (WebLI) on ImageNet zero-shot classification. Similar improvements are noted over other

⋆ Research Consultant with Google, * Equal advising, ○ Intern at Google during the project.

classification and retrieval metrics. Finetuning SILC-C* to get SILC-C shows consistent improvements over all metrics showing that the cleaner subset benefits the larger model too. We also train SILC-C* with ViT/B8. ViT/B8 has the same number of learnable parameters for the Transformer as B/16 but uses half the patch size. We observe that the smaller patch size allows this model to consistently outperform the B/16 model. However, the smaller patch-size also means that the transformer has to process a longer sequence of tokens at each encoder block. As a result, ViT/B8 has approximately the same compute requirement as ViT/L16. We observe that The B/8 model performs slightly worse than the ViT/L16 model. Finally we see that the ViT/B8 model also benefits from finetuning on the cleaner subset of WebLI and SILC-C ViT/B8 consistently improves on SILC-C* ViT/B8.

3 Performance of Teacher vs Student for SILC*.

Our training setup consists of the student that is updated with gradient descent and a teacher that is updated with an EMA update. For comparisons in our main manuscript, we report the performance for the teacher. We additionally compare the teacher with the student in Table 4. During training we observe that the teacher converges faster than the student but both converge to about the same performance towards the end of training for zero-shot classification, few-shot classification and retrieval. However for zero-shot segmentation, the teacher achieves superior performance compared to the student. Similar observation has been made by earlier self-supervised works [3, 15] for self-supervised models. However in their case, the teacher always outperforms the students. In our setup, since the student is updated with image-text loss, it achieves similar performance to the teacher on classification and retrieval.

4 SILC applied to CNN backbone.

To show SILC's universality beyond ViT, we ablate SigLIP and and SILC-S* for 2B Example-Seen at ResNetv2-50. SILC-S* achieves IM0shot/COCO I2T Ret/ COCO T2I Ret **67.7/59.5/39.2** to SigLIP's 64.9/57.9/38.2. We then scale SILC-S* to RNv2-101 for 20B Example-Seen to get the best CNN based VLM at this model size achieving IM0shot/COCO I2T Ret/ COCO T2I Ret **79.2/70.0/52.8**.

5 Additional Results for Zero-shot Semantic Segmentation.

TCL [4], the previous state-of-the-art in zero-shot semantic segmentation, ensembles their learned model with MaskCLIP [24] by tuning a mixing factor on the predictions of the two models. However, this mixing factor violates the zero-shot protocol proposed by [19] as the model has access to segmentation labels

Model	ViT type	ImageNet				Retrieval COCO	
		Validation	v2	ReaL	ObjectNet	I2T@1	T2I@1
CLIP [16]	ViT/B16	68.3	61.9	-	55.3	52.4	33.1
OpenCLIP [11]	ViT/B16	70.2	62.3	-	56.0	59.4	42.3
MetaCLIP [20]	ViT/B16	72.1	65.1	-	61.4	59.4	41.3
EVA-CLIP [18]	ViT/B16	74.7	67.0	-	62.3	58.7	42.2
DFN-2B [9]	ViT/B16	76.2	68.2	-	63.2	60.4	43.4
CLIP (WebLI) [23]	ViT/B16	74.1	66.8	80.9	69.6	61.7	43.9
SILC-C* (Ours)	ViT/B16	75.3	68.4	82.5	74.1	62.5	44.9
SILC-C (Ours)	ViT/B16	76.2	69.1	82.8	74.6	66.1	**49.1**
SigLIP [23]	ViT/B16	75.1	68.0	81.9	70.1	62.6	44.9
SILC-S* (Ours)	ViT/B16	75.8	68.7	83.0	73.6	63.0	44.6
SILC-S (Ours)	ViT/B16	**76.6**	**69.4**	**83.5**	**74.6**	**66.2**	48.7
SILC-C* (Ours)	ViT/B8	77.5	70.5	84.0	78.1	64.5	46.0
SILC-C (Ours)	ViT/B8	**78.2**	**71.6**	**84.4**	**78.7**	**67.3**	**50.3**
CLIP [16]	ViT/L14	75.5	69.0	-	69.9	56.3	36.5
OpenCLIP [11]	ViT/L14	74.0	61.1	-	66.4	62.1	46.1
MetaCLIP [20]	ViT/L14	79.2	72.6	-	74.6	60.0	43.8
CLIPA-v2 [12]	ViT/L14	79.7	72.8	-	71.1	64.1	46.3
Datacomp1B [10]	ViT/L14	79.6	73.1	-	69.9	65.1	47.0
EVA-CLIP [18]	ViT/L14	79.8	72.9	-	75.3	63.7	47.5
DFN-2B [9]	ViT/L14	**81.4**	74.6	-	74.0	65.6	48.6
SigLIP [23]	ViT/L16	80.5	74.2	85.9	77.9	69.5	51.5
CLIP (WebLI) [23]	ViT/L16	79.7	73.3	85.3	77.3	67.7	48.9
SILC-C* (Ours)	ViT/L16	81.0	74.6	86.3	81.6	68.4	50.9
SILC-C (Ours)	ViT/L16	**81.4**	**75.5**	**86.7**	**82.2**	**70.1**	**52.8**
OpenCLIP [11]	ViT/G14	80.0	73.6	-	72.8	67.4	51.4
Datacomp1B [10]	ViT/G14	82.7	77.0	-	76.9	67.8	50.0
EVA-CLIP [18]	ViT/E14	82.1	75.6	-	79.4	68.7	51.1
SILC-S* (Ours)	ViT/G16	**83.7**	**77.8**	**87.9**	**85.4**	**73.2**	**54.7**

Table 1: We compare our SILC models with publically available contrastive image-text models and show that SILC models achieves the best performance. Best number for each model configuration is **bolded**.

during the mixing factor tuning. We additionally report the performance of TCL using author's checkpoint by removing the ensemble with MaskCLIP in Table 5. We show that this results in slight drop in performance. We advice future works to not touch segmentation labels to tune parts of their models to be consistent with the zero-shot protocol. The TCL [4] results reported by the authors in their main paper additionally use PAMR to refine the predicted segmentation of their model and remove some noise. The authors also report the performance of their model without PAMR in their supplementary which we have reported in our main manuscript. We also list TCL with PAMR numbers in Table 5 to show that post refinement can give boost in performance but it can mask the actual performance of the learned model. Refinement steps can improve all methods as shown in TCL's supplementary. Therefore, we do not use refinement in our work as we are interested in the raw zero-shot segmentation performance of the model. We additionally perform multi-scale evaluation for our best model SILC-C. Multiscale evaluation consists of the followings steps 1. Compute logits at multiple scales. 2. Resize them to original label size. 3. Average the logits and compute predictions. We observe that this further improves the zero-shot segmentation performance of SILC-C.

Model	Dataset	Example-Seen	ImageNet 0 shot	ImageNet Few shot			COCO Retrieval	
			T1	1shot	5shot	10shot	I2T@1	T2I@1
SigLIP [23]	CC3M	50M	17.5	8.9	17.7	21.2	14.6	10.6
SILC-S* (Ours)	CC3M	50M	**25.2**	**14.8**	**26.5**	**30.5**	**22.0**	**17.4**
MaskCLIP [7]	YFCC15M	500M	44.5	-	-	-	41.4	25.5
SigLIP [23]	CC12M	500M	38.4	19.4	35.1	39.9	37.4	25.2
SILC-S* (Ours)	CC12M	500M	**47.1**	**28.4**	**44.9**	**49.0**	**42.8**	**31.3**
SigLIP [23]	LAION400M	5B	67.4	33.5	53.5	58.0	58.7	41.2
SILC-S* (Ours)	LAION400M	5B	**70.8**	**37.8**	**57.5**	**61.1**	**61.5**	**43.2**
OpenCLIP [11]	LAION2B	13B	70.2	-	-	-	59.4	42.3
SigLIP [23]	LAION2B	5B	68.9	34.6	54.8	59.3	62.3	44.4
SILC-S* (Ours)	LAION2B	5B	72.4	38.9	58.6	62.4	64.4	46.2
SILC-S* (Ours)	LAION2B	20B	**74.6**	**42.0**	**61.1**	**65.2**	**66.0**	**47.9**

Table 2: Consistency on open source datasets. We show that SILC models consistently outperforms baselines when trained on open-source datasets. SILC-S* trained for 5B E-S on LAION2B already outperforms the much more tuned OpenCLIP which was designed on this dataset. SILC-S* trained for 20B E-S achieves the best reported performance on open source datasets at ViT/B16. This further shows the strong performance of SILC models.

Comparison with PACL (WebLI). We report zero-shot semantic segmentation results on an additional baseline PACL [14] in Table 5. Since PACL checkpoints and code are not available, we contacted the authors and closely followed their instructions in our implementation. We train a small MLP as a residual on top of our CLIP (WebLI) B/16 model similar to the authors. We use the cleaner small subset of WebLI with 100 Million image-text pairs for this experiment and report the performance in Table 5. We observe that PACL performs worse than TCL and SILC models. Since our reproduced numbers are different from the reported numbers in PACL manuscript, we contacted the authors and discussed their evaluation protocol in detail. PACL uses segmentation label supervision at test time and tunes a threshold on the model's prediction to only extract image regions where the model has a high confidence. The segmentation performance is then only evaluated over these regions and not the full label from the dataset. Hence the PACL [14] performance reported in their main manuscript is not directly comparable with our protocol. Since we are interested in the raw zero-shot semantic segmentation performance of the model over the full image, we do not perform this step and show that SILC models outperform PACL for our protocol.

6 Additional Ablation.

6.1 Impact of not aligning all global views with text.

We utilize multiple global views in our local-to-global correspondence learning branch similar to previous works in self-supervised learning [3, 15]. However, for this objective to be complimentary to image-text contrastive learning, we found that we need to align all global views with text. Otherwise the two objectives

Model		Zero-Shot Classification		Few-shot classification						Retrieval	
		ImageNet	CIFAR100	ImageNet			CIFAR100			COCO	
		T1	T1	1shot	5shot	10shot	1shot	5shot	10shot	I2T@1	T2I@1
CLIP (WebLI) [23]	ViT/B16	74.1	68.4	42.8	63.2	67.3	39.4	59.6	64.6	61.7	43.9
SILC-C* (Ours)	ViT/B16	75.3	71.0	44.6	64.3	67.8	42.8	64.6	69.6	62.5	44.9
SILC-C (Ours)	ViT/B16	**76.2**	**72.3**	**45.3**	**65.0**	**68.5**	**45.2**	**66.9**	**71.3**	**66.1**	**49.1**
SigLIP [23]	ViT/B16	75.1	69.8	44.0	64.2	68.4	39.0	61.7	66.3	62.6	44.9
SILC-S*(Ours)	ViT/B16	75.8	69.2	45.2	64.6	68.4	40.3	63.3	67.4	63.0	44.6
SILC-S(Ours)	ViT/B16	**76.6**	**70.6**	**45.9**	**65.2**	**68.9**	**41.8**	**64.9**	**68.9**	**66.2**	**48.7**
SILC-C* (Ours)	ViT/B8	77.5	72.6	48.9	67.3	70.7	47.9	68.6	73.1	64.5	46.0
SILC-C (Ours)	ViT/B8	**78.2**	**73.2**	**49.5**	**67.8**	**71.1**	**49.3**	**69.7**	**73.8**	**67.3**	**50.3**
CLIP (WebLI) [23]	ViT/L16	79.7	77.5	52.9	72.1	75.5	42.6	69.3	73.7	67.7	48.9
SILC-C* (Ours)	ViT/L16	81.0	80.5	54.8	73.9	76.8	53.2	75.8	79.5	68.4	50.9
SILC-C (Ours)	ViT/L16	**81.4**	**81.4**	**55.6**	**74.2**	**76.9**	**53.7**	**77.2**	**80.5**	**70.1**	**52.8**

Table 3: Performance of additional SILC models. We show that SILC-C* outperforms CLIP (WebLI) at ViT/L16 too. Moreover, we show that SILC-C achieves consistent improvement over SILC-C* at ViT/B8, ViT/B16 and ViT/L16. Best number for each model configuration is **bolded**. Second best is underlined.

Model	ImageNet zero shot	CIFAR100 zero shot	ImageNet Few shot			COCO Retrieval		ZS Segmentation		
	T1	T1	1shot	5shot	10shot	I2T@1	T2I@1	A-150	Stuff	PC-59
SILC-C*Teacher	75.3	71.0	44.6	64.3	67.8	62.5	44.9	17.2	18.2	29.3
SILC-C*Student	75.3	71.0	44.6	64.3	67.8	62.5	44.9	16.1	17.3	27.4

Table 4: Comparing SILC*Teacher and Student performance, we observe that both teacher and student behave similarly on classification and retrieval tasks. However, the teacher achieves superior performance on zero-shot segmentation.

diverge. The image-text loss for unaligned global views start to increase as the model over-fits them for local-to-global correspondence learning. This hurts the model performance as shown in Table 6 (last row).

7 Additional Qualitative Results.

7.1 Additional Qualitatives on Zero-Shot Semantic Segmentation.

We report additional qualitative results for Zero-Shot Semantic Segmentation in Figure 1 for A-150 and Figure 2 for PC-59. They demonstrate that SILC-C produces less noisy segmentations compared to CLIP and is less prone to class confusions such as booth/computer, field/grass, road/screen, swivel chair/chair, blind/curtain, counter/countertop, counter/kitchen, rock/mountain, rock/sand, animal/sea, armchair/sofa, and food/glass.

7.2 Additional Qualitatives on Open Vocabulary Semantic Segmentation.

We report additional qualitative results for Open Vocabulary Semantic Segmentation in Figure 3 for A-150 and Figure 4 for PC-459. They demonstrate that SILC-C better distinguishes semantically similar classes such as bookcase/shelf, countertop/counter, cabinet/shelf, swivel chair/ chair, stool/chair, pier/bridge,

Model	A-150	PC-59	Cityscapes	VOC-20	COCO-Stuff
TCL + PAMR [4]	17.1	33.9	24.0	83.2	22.1
PACL (WebLI) [14]	13.2	21.0	16.0	60.4	12.9
TCL no ensemble [4]	14.1	28.7	22.0	76.7	18.6
TCL [4]	14.9	30.3	23.1	77.5	19.6
SigLIP [23]	13.6	22.9	20.8	64.7	13.4
SILC-S* (Ours)	16.7	28.6	23.4	72.1	17.3
SILC-S (Ours)	18.6	30.9	25.2	76.3	19.7
SILC-C* (Ours)	17.2	29.3	25.1	73.5	18.2
SILC-C (Ours)	19.3	31.6	26.9	77.5	20.8
SILC-C+ Multiscale (Ours)	**22.5**	**36.2**	**33.5**	**83.8**	**24.1**

Table 5: Additional Zero-shot Semantic Segmentation comparisons. We report additional results for the previous state-of-the-art TCL. We additionally report result for another baseline PACL. SILC consistently outperforms the baselines on the same evaluation protocol i.e. raw predictions of the model.

Model	ImageNet 0 shot	ImageNet Few shot			COCO Retrieval	
	T1	1shot	5shot	10shot	I2T@1	T2I@1
CLIP (WebLI)	71.7	36.4	57.7	62.5	59.1	42.9
+ additional views	73.6	38.7	60.8	65.7	60.6	43.2
+ EMA	73.7	38.4	60.7	65.5	61.3	43.1
+ Self Dist (**SILC-C***)	**74.3**	**39.9**	**61.2**	**65.7**	**62.7**	**43.9**
CLIP (WebLI) + EMA + Self Dist	67.7	24.6	40.6	45.8	52.3	36.9

Table 6: 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. On the other hand, directly adding self-distillation without aligning all the global views with contrastive loss (additional views) hurts performance.

desk/shelf, train/metal, building/shed, wall/brick, sign/poster, cloth/plastic, ground/sand, and boat/water.

8 Additional Details.

8.1 Evaluation Protocol.

We follow the original CLIP [16] paper for the zero-shot classification and retrieval evaluations. We follow the original ViT [8] paper for few-shot classification evaluation. The evaluation code is used from the `big_vision` codebase [1, 2]. For our segmentation evaluations, we export our model weights to PyTorch. We follow previous works [4, 17, 21, 24] and implement our zero-shot segmentation evaluation in MMSeg [6] with Sliding-Window evaluation. We directly use the

Initialization	Training data	COCO	LVIS	
		AP	AP_{all}	AP_{rare}
CLIP (WebLI)	WebLI N-grams	40.4	31.9	29.2
SILC-C*(Ours)	WebLI N-grams	41.8	33.3	**30.4**
SILC-C(Ours)	WebLI N-grams	**42.3**	**33.4**	30.0
SigLIP	WebLI N-grams	40.9	32.8	30.4
SILC-S*(Ours)	WebLI N-grams	**42.7**	34.2	**32.4**
SILC-S(Ours)	WebLI N-grams	42.5	**34.3**	32.1

Table 7: Training OWLv2 for Object Detection with SILC models offers consistent improvement over CLIP and SigLIP for open vocabulary object detection. These models are trained with pseudo labels from WebLI N-grams [13] and evaluated zero-shot on COCO and LVIS.

Model	Zero-Shot Classification		Retrieval		Zeroshot segmentation				
	ImageNet	CIFAR100	COCO		A-150	PC-59	Cityscapes	VOC-20	COCO-Stuff
	T1	T1	I2T@1	T2I@1					
MaskCLIP (WebLI) [7]	74.4	69.0	61.4	43.6	16.3	27.2	23.0	72.8	15.9
CLIP (WebLI) [23]	74.1	68.4	61.7	43.9	15.0	24.0	22.6	69.5	15.0
SILC* (Ours)	75.3	71.0	62.5	44.9	17.2	29.3	25.1	73.5	18.2
SILC (Ours)	**76.2**	**72.3**	**66.1**	**49.1**	**19.3**	**31.6**	**26.9**	**77.5**	**20.8**

Table 8: Comparing SILC*with MaskCLIP [7], we observe that our pretraining framework consistently outperforms this baseline too. We reproduce MaskCLIP with WebLI data and observe that it improves on baseline CLIP (WebLI) for zero-shot classification and zero-shot segmentation. However, SILC-C* and SILC-C consistently outperform it on all metrics. The best performance is **bolded**, the second best is underlined.

model's prediction for segmentation and do not perform any refinement. For Open Vocabulary segmentation, we directly use the codebase from Cat-Seg [5] and do not perform any hyper-parameter tuning. All results for Cat-Seg are reported using the training protocol from the authors.

8.2 Additional Training Details.

We provide training detail for SILC* and SILC in the main manuscript. We provide additional training detail in this supplementary. SILC* at ViT B/16 can be trained on 256 TPUv4 chips meanwhile the B/8 and L/16 models require 512 chips. The training takes around 5 days. For the fine-tuning stage for SILC, we use a initial learning rate of $1e^{-4}$ and use a rsqrt scheduler [22] with 50000 cool down steps. We do not use warm up or weight decay at this stage. The MLP used for our self-distillation loss consists of two layers with gelu activation and dimension of 2048. This is followed by a bottleneck of dimension 256 followed by a projection to the output dimension K of size 65536. We do not perform

Model	Classification		Captioning	Question Ans.	
	ImageNet	SUN397	COCO	GQA	VQAv2
CLIP (WebLI)	82.3	82.4	118.1	52.5	63.5
SILC-C*(Ours)	83.8	**83.4**	120.8	53.1	64.6
SILC-C(Ours)	**83.9**	**83.4**	**122.0**	**53.6**	**64.7**
SigLIP	82.5	82.2	117.5	51.9	63.0
SILC-S*(Ours)	83.7	82.9	121.2	53.2	64.5
SILC-S(Ours)	**83.9**	**83.2**	**122.0**	**54.2**	**65.2**

Table 9: Evaluating SILC visual representation with LiT-Decoder in a multi-task setup, we observe consistent improvements on all tasks compared to CLIP and SigLIP. These improvements are especially apparent for tasks that require local understanding of the image i.e. Captioning and Question Answering.

tuning of each loss's contribution and directly optimise the sum of loss coming from our model's two components.

8.3 Limitations.

While SILC models improve on limitations of CLIP to better encode local semantics, they still come with limitations of contrastive image-text framework. While we significantly improve on zero-shot semantic segmentation performance, the performance is still far away from practical usage. We still require to train the model for open vocabulary segmentation to get better performance. Similarly, object detection, captioning etc. needs to be trained separately on top. Finally,SILC requires significant compute to train. Future works can focus on making the pretraining more compute efficient.

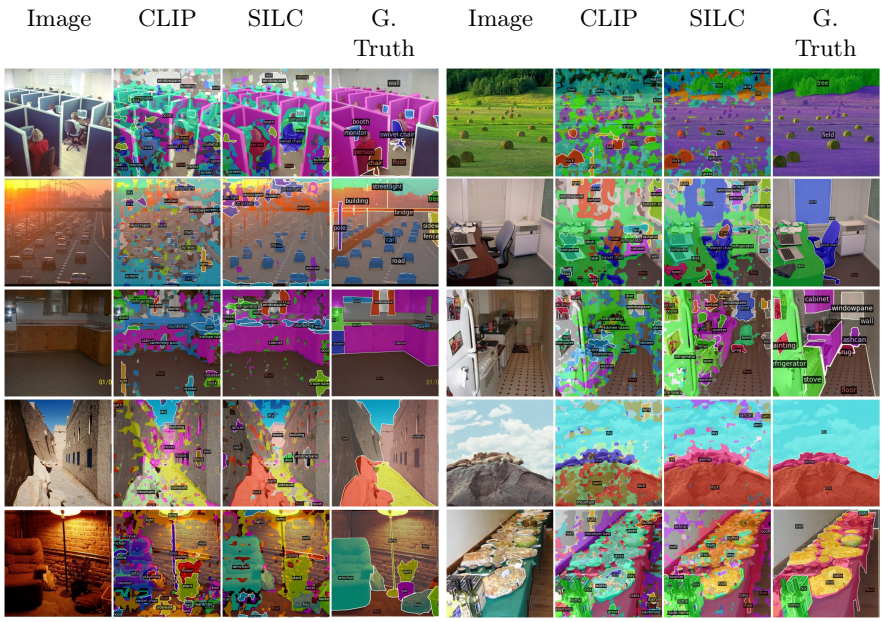

Fig. 1: Additional qualitative results for zero-shot segmentation on A-150.

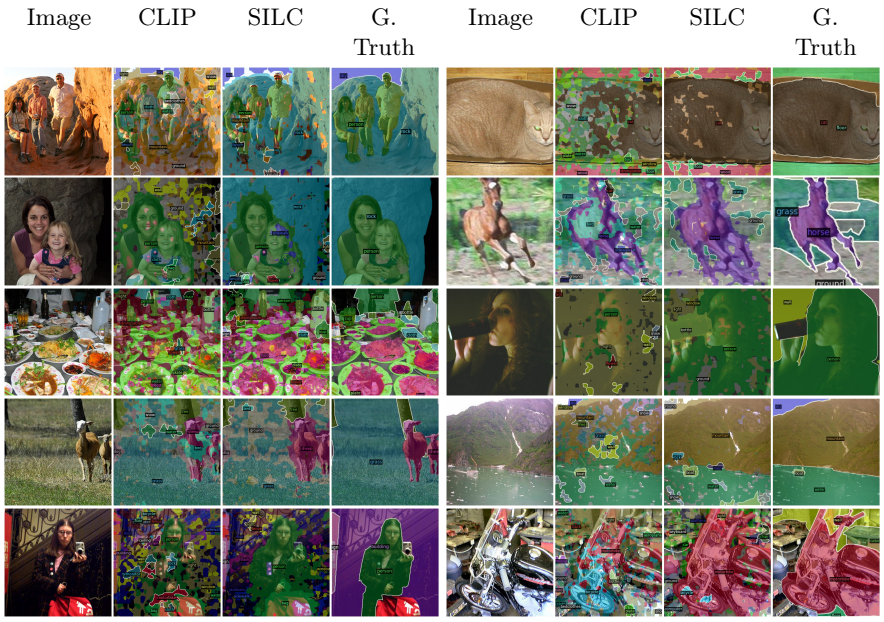

Fig. 2: Additional qualitative results for zero-shot segmentation on PC-59.

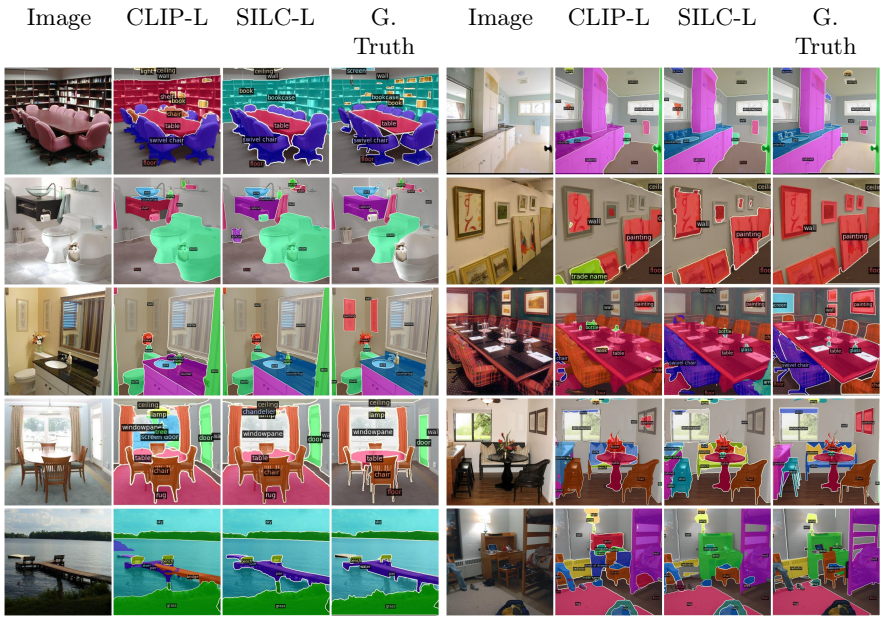

Fig. 3: Additional qualitative results for open-vocabulary segmentation on A-150.

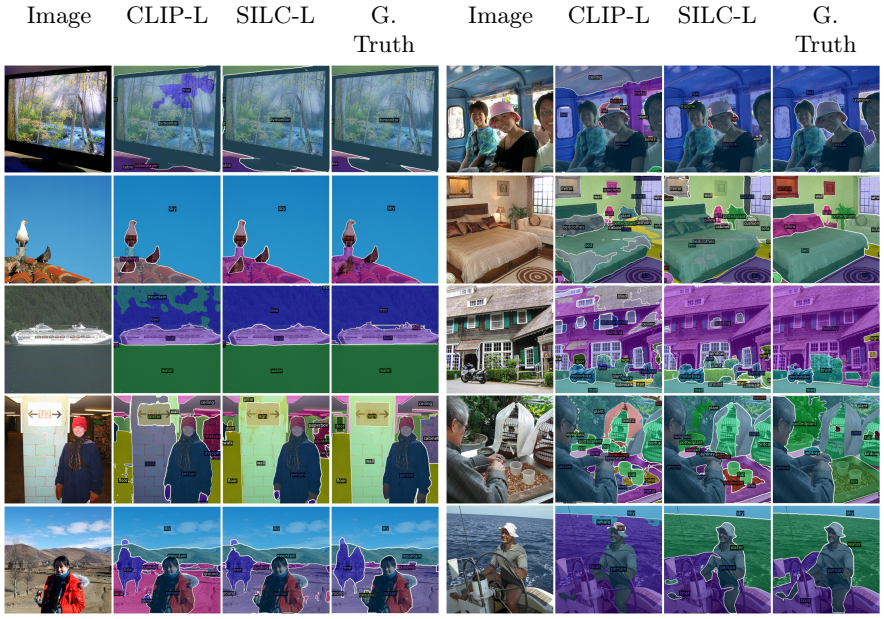

Fig. 4: Additional qualitative results for open-vocabulary segmentation on PC-459.

Bibliography

[1] Beyer, L., Zhai, X., Kolesnikov, A.: Better plain vit baselines for imagenet-1k (2022) 6

[2] Beyer, L., Zhai, X., Kolesnikov, A.: Big vision. https://github.com/google-research/big_vision (2022) 6

[3] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021) 2, 4

[4] Cha, J., Mun, J., Roh, B.: Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In: CVPR (2023) 2, 3, 6

[5] Cho, S., Shin, H., Hong, S., An, S., Lee, S., Arnab, A., Seo, P.H., Kim, S.: Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. arXiv preprint arXiv:2303.11797 (2023) 7

[6] Contributors, M.: MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation (2020) 6

[7] Dong, X., Bao, J., Zheng, Y., Zhang, T., Chen, D., Yang, H., Zeng, M., Zhang, W., Yuan, L., Chen, D., et al.: Maskclip: Masked self-distillation advances contrastive language-image pretraining. In: CVPR (2023) 1, 4, 7

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 6

[9] Fang, A., Jose, A.M., Jain, A., Schmidt, L., Toshev, A., Shankar, V.: Data filtering networks. arXiv preprint arXiv:2309.17425 (2023) 3

[10] Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next generation of multimodal datasets. Advances in Neural Information Processing Systems **36** (2024) 3

[11] Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). https://doi.org/10.5281/zenodo.5143773, https://doi.org/10.5281/zenodo.5143773, if you use this software, please cite it as below. 3, 4

[12] Li, X., Wang, Z., Xie, C.: Clipa-v2: Scaling clip training with 81.13

[13] Minderer, M., Gritsenko, A., Houlsby, N.: Scaling open-vocabulary object detection. arXiv preprint arXiv:2306.09683 (2023) 7

[14] Mukhoti, J., Lin, T.Y., Poursaeed, O., Wang, R., Shah, A., Torr, P.H., Lim, S.N.: Open vocabulary semantic segmentation with patch aligned contrastive learning. In: CVPR (2023) 1, 4, 6

[15] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Di-

nov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 2, 4
[16] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICLR (2021) 3, 6
[17] Shin, G., Xie, W., Albanie, S.: Reco: Retrieve and co-segment for zero-shot transfer. In: NeurIPS (2022) 6
[18] Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023) 3
[19] Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE transactions on pattern analysis and machine intelligence **41**(9) (2018) 2
[20] Xu, H., Xie, S., Tan, X.E., Huang, P.Y., Howes, R., Sharma, V., Li, S.W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying clip data. arXiv preprint arXiv:2309.16671 (2023) 3
[21] Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: CVPR (2022) 6
[22] Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. In: CVPR (2022) 7
[23] Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV (2023) 3, 4, 5, 6, 7
[24] Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: ECCV (2022) 2, 6