A Simple Framework for Open-Vocabulary Zero-Shot Segmentation

Thomas Stegmüller*, Tim Lebailly*, Nikola Đukić, Behzad Bozorgtabar,
Tinne Tuytelaars, Jean-Philippe Thiran

EPFL

KU Leuven

March 2025

Cross-Modality Object-level Supervision

Objective:

- Zero-shot segmentation
- Scalable training
- Efficient inference

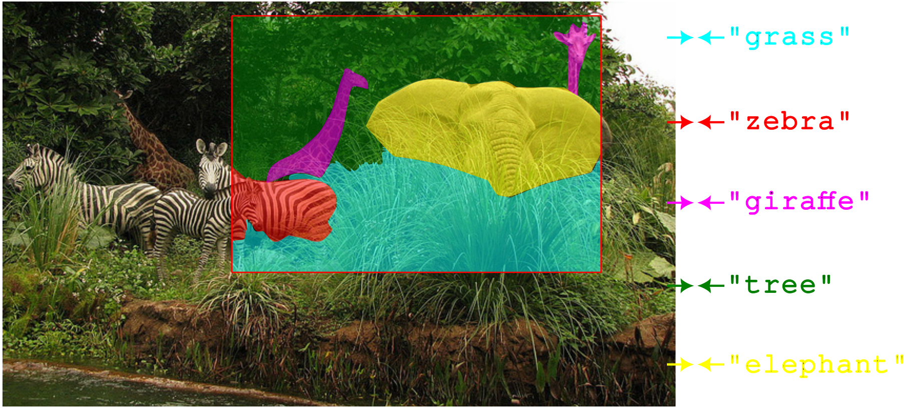

Figure: Cross-modality object-level supervision

CLIP → LiT → SimZSS

Key ingredients:

1. Freezing the vision tower *à la* LiT [13] enables spatial alignment.
2. Alignment should be at the concept level, not the caption level.

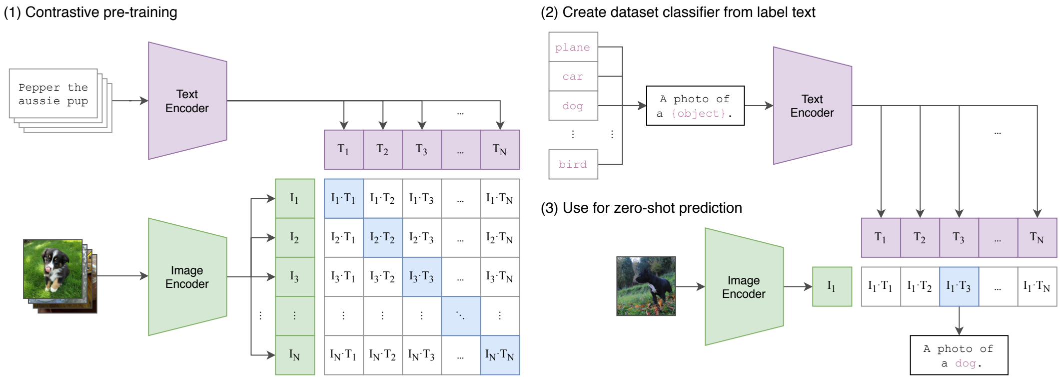

Figure: Overview of CLIP. Source: [7].

Exploiting SSL's Spatial Awareness

Semantic clustering is already present in visual representations of SSL transformers:

- Natural language supervision is suboptimal for learning visual representations [13].
- Self-supervised transformers demonstrate strong spatial awareness [1, 9].

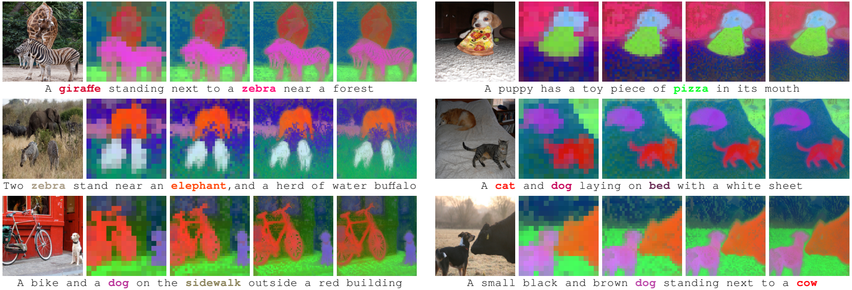

Figure: Patch-level representations and text concepts visualized in RGB space.

Obtaining Pairs of Vision-language Concept-level Representations

Concept-level text representations from part-of-speech tagger:

$$f_t(\mathbf{x}_t) = \mathbf{z}_t \qquad \mathbf{c}_t^l = \frac{1}{|\mathcal{S}_l|} \sum_{i \in \mathcal{S}_l} \mathbf{z}_t^i \qquad \tilde{\mathbf{c}}_t^l = g\left(\mathbf{c}_t^l\right)$$

Query the vision modality to obtain pairs:

$$f_v(\mathbf{x}_v) = \mathbf{z}_v \qquad \mathbf{s} = \mathtt{softmax}\left(\frac{\mathbf{z}_v \tilde{\mathbf{c}}_t^l}{\tau}\right) \qquad \mathbf{c}_v = \mathbf{z}_v^\top \mathbf{s}$$

Figure: Vision-language alignment of text concepts and dense visual representations.

Object-level Cross-modal Alignment

Store unique concepts in the batch and keep track of the underlying represented concepts:

$$\mathbf{C}_t \in \mathbb{R}^{\tilde{b}\times d_t} \qquad \mathbf{C}_v \in \mathbb{R}^{\tilde{b}\times d_v} \qquad \mathbf{q} \in \{0, 1, ..., k-1\}^{\tilde{b}}$$

Compute the weights of a linear classifier $\mathbf{h} \in \mathbb{R}^{k\times d_v}$:

$$\mathbf{h}_i = \sum_j \mathbb{1}_{\{\mathbf{q}_j = i\}} g\left(\mathbf{C}_t\right)_j \qquad \mathbf{p} = \underset{k}{\texttt{softmax}}\left(\mathbf{C}_v \mathbf{h}^\top\right)$$

Cross-modality cross-entropy loss:

$$\mathcal{L}_l = \frac{1}{\tilde{b}} \sum_i \sum_j -\mathbb{1}_{\{\mathbf{q}_i = j\}} \log\left(\mathbf{p}_{ij}\right)$$

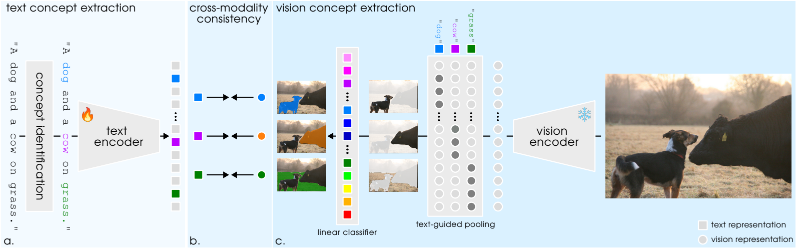

Figure: Overview of SimZSS.

Zero-shot Segmentation Results

Table: **Zero-shot foreground segmentation.** Pixel-wise predictions are obtained by projecting patch representations onto pre-computed text embeddings of the class names, followed by up-sampling. The mIoU scores are reported across five standard segmentation datasets. † refers to our reproduction using DINOv2 pre-trained vision backbones. The remaining results are as reported in [11].

Method	❄️ Params	🔥 Params	Pascal VOC	Pascal Context	COCO-Stuff	Cityscapes	ADE20K
Miscellaneous							
ReCo [8]	313M	0	57.7	22.3	14.8	21.1	11.2
GroupViT [12]	0	55M	79.7	23.4	15.3	11.1	9.2
TCL [2]	156M	21M	77.5	30.3	19.6	23.1	14.9
MaskCLIP [4]	291M	0	74.9	26.4	16.4	12.6	9.8
OVDiff [5]	1,226M	0	81.7	33.7	-	-	14.9
CLIP-DINOiser [11]	-	-	80.9	35.9	24.6	31.7	20.0
LAION-400M							
CLIP [7] (ViT-B)	94M	63M	35.1	7.7	4.2	1.8	2.0
LiT† [13] (ViT-B)	94M	63M	80.5	31.8	23.3	24.7	18.7
SimZSS (ViT-B)	94M	63M	85.1	34.2	24.9	27.8	19.6
COCO Captions							
LiT† [13] (ViT-B)	94M	63M	86.1	35.5	25.6	25.8	18.1
SimZSS (ViT-S)	21M	40M	87.2	37.3	23.8	29.2	17.9
SimZSS (ViT-B)	94M	63M	**90.3**	**43.1**	**29.0**	**33.0**	**21.8**

Zero-shot Classification Results

Observations:

- No trade-off between zero-shot classification and segmentation.
- Segmentation benefits more from curation than scale.

Table: **Zero-shot classification.** Image-level predictions are obtained by projecting the image `[CLS]` token onto pre-computed text embeddings of class names. Accuracy is reported for various visual pre-training and vision-language alignment methods. † refers to our reproduction using DINOv2 pre-trained vision backbones. The remaining results are as reported in [13].

Method	Visual pre-training	Backbone	Pre-training dataset	Alignment dataset	Alignment samples	Labels	ImageNet-1K	Average
LiT [13]	MoCo-v3 [3]	ViT-B/16	ImageNet-1K	CC12M+YFCC100M	-	✗	55.4	-
LiT [13]	DINOv1 [1]	ViT-B/16	ImageNet-1K	CC12M+YFCC100M	-	✗	55.5	-
LiT [13]	AugReg [10]	ViT-B/16	ImageNet-21k	CC12M+YFCC100M	-	✓	55.9	-
LiT† [13]	DINOv2 [6]	ViT-B/14	LVD-142M	COCO Captions	4M	✗	22.6	24.4
LiT† [13]	DINOv2 [6]	ViT-B/14	LVD-142M	LAION-400M	400M	✗	63.6	37.5
CLIP [7]	-	ViT-B/16	-	LAION-400M	12.8 B	✗	67.0	**47.2**
Ours								
SimZSS	DINOv2 [6]	ViT-B/14	LVD-142M	COCO Captions	4M	✗	24.3	26.1
SimZSS	DINOv2 [6]	ViT-B/14	LVD-142M	LAION-400M	400M	✗	64.1	38.9
SimZSS	DINOv2 [6]	ViT-B/14	LVD-142M	LAION-400M	1.6B	✗	**69.3**	41.3

Conclusion

Strengths:

- Efficient in both data and computation.
- Maintains classification performance while enabling segmentation.
- Achieves strong accuracy without compromising inference speed.

Limitations:

- Referring segmentation remains unexplored.
- Consistency is enforced only for concepts explicitly mentioned in captions.
- Performance still lags behind the upper bound (linear segmentation).

Table: **Computational and memory efficiency.** The efficiency of SimZSS is compared to that of related methods, *i.e.*, LiT and CLIP. When feasible, we report results using the local training batch size; otherwise, the largest power of 2 that fits into memory is utilized. The reported values are obtained on a single node equipped with 4x AMD MI250x (2 compute die per GPU, *i.e.*, $\texttt{worldsize} = 8$).

Method	Batch size per compute die	Memory per compute die [GB]	Time per step [ms]	Throughput [image/s]
CLIP	256	~ 40	1196.0	1712
LiT	1024	~ 27	2049.2	3997
SimZSS	1024	~ 38	2069.8	3957

Thank you!

References I

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00951. URL `https://doi.org/10.1109/ICCV48922.2021.00951`.

[2] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 11165–11174. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01074. URL `https://doi.org/10.1109/CVPR52729.2023.01074`.

[3] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9620–9629. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00950. URL `https://doi.org/10.1109/ICCV48922.2021.00950`.

[4] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 10995–11005. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01058. URL `https://doi.org/10.1109/CVPR52729.2023.01058`.

References II

[5] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation. *ArXiv preprint*, abs/2306.09316, 2023. URL https://arxiv.org/abs/2306.09316.

[6] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

[8] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/daabe43c3e1d06980aa23880bfbe1f45-Abstract-Conference.html.

References III

[9] Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*, 2021.

[10] Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022.

[11] Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation, 2024.

[12] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022.

[13] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18123–18133, 2022.