Understanding Fine-tuning CLIP for Open-vocabulary Semantic Segmentation in Hyperbolic Space

Zelin Peng[1], Zhengqin Xu[1], Zhilin Zeng[1], Changsong Wen[1], Yu Huang[1],
Menglin Yang[2], Feilong Tang[3,4], and Wei Shen[1(✉)]
[1]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2]Hong Kong University of Science and Technology (Guangzhou) [3]MBZUAI, [4]Monash University.
https://github.com/SJTU-DeepVisionLab/HyperCLIP

Abstract

CLIP, a foundational vision-language model, has emerged as a powerful tool for open-vocabulary semantic segmentation. While freezing the text encoder preserves its powerful embeddings, recent studies show that fine-tuning both the text and image encoders jointly significantly enhances segmentation performance, especially for classes from open sets. In this work, we explain this phenomenon from the perspective of hierarchical alignment, since during fine-tuning, the hierarchy level of image embeddings shifts from image-level to pixel-level. We achieve this by leveraging hyperbolic space, which naturally encoders hierarchical structures. Our key observation is that, during fine-tuning, the hyperbolic radius of CLIP's text embeddings decreases, facilitating better alignment with the pixel-level hierarchical structure of visual data. Building on this insight, we propose HyperCLIP, a novel fine-tuning strategy that adjusts the hyperbolic radius of the text embeddings through scaling transformations. By doing so, HyperCLIP equips CLIP with segmentation capability while introducing only a small number of learnable parameters. Our experiments demonstrate that HyperCLIP achieves state-of-the-art performance on open-vocabulary semantic segmentation tasks across three benchmarks, while fine-tuning only approximately 4% of the total parameters of CLIP. More importantly, we observe that after adjustment, CLIP's text embeddings exhibit a relatively fixed hyperbolic radius across datasets, suggesting that the granularity required for this segmentation task might be quantified using the hyperbolic radius.

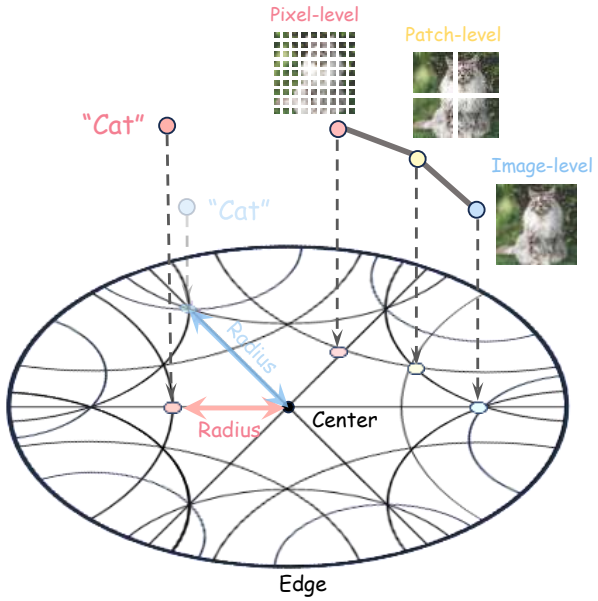

Figure 1. **Visualization of hyperbolic space by a Poincaré ball model**: The height of points reflects the hierarchical level of input data. The black segments refer to the shortest path between the points in the Poincaré ball model. The red arrow line denotes the hyperbolic radius of text embeddings. Embeddings closer to the center have a smaller radius, representing low-level visual information (e.g., pixels), while points farther from the center have a larger radius, representing high-level visual information (e.g., images). During fine-tuning, the hyperbolic radius of text embeddings is adjusted to match the radius associated with pixels, thereby endowing CLIP with segmentation ability (Best viewed in color).

1. Introduction

The goal of open-vocabulary semantic segmentation is to develop a segmentation model capable of labeling each pixel in an image with categories that extend beyond a predefined closed set, based on textual descriptions. Vision-language foundation models [5, 8, 9, 11, 13–16, 18–21, 23, 26, 27, 29, 33–35, 40], particularly CLIP [29], are frequently employed to provide open-vocabulary recognition capabilities. Consequently, open-vocabulary semantic segmentation fundamen-

✉ Corresponding Author: wei.shen@sjtu.edu.cn

tally involves transferring these vision-language foundation models, originally trained with image-level supervision, to achieve pixel-level predictions.

Current methods [6, 37, 39, 42] typically fine-tune CLIP on a closed set with segmentation annotations, i.e., COCO [3], to equip it with the segmentation capability. A prevailing opinion is that simply freezing CLIP's text encoder maximally preserves the powerful text embeddings of a vast array of classes, which is believed to be beneficial for good generalization on open sets. Many studies follow this opinion [17, 22, 39]. However, recent studies [6, 25, 37] offer an alternative perspective on fine-tuning: simultaneously fine-tuning both the image and text encoders results in superior segmentation performance on open sets compared to former approaches. To explain and resolve this seemingly paradoxical phenomenon, we first leverage a non-Euclidean manifold—specifically, hyperbolic space [4]—to understand why fine-tuning the text encoder appears to improve segmentation performance on open sets. Building on this understanding, our objective is to design an explainable fine-tuning strategy with a minimal number of introduced tunable parameters, which is able to equip CLIP with strong segmentation performance while minimizing the loss of its generalization ability.

Existing research [1] has found that images inherently exhibit hierarchical structures, consisting of multiple levels: pixels, patches, objects, entire scenes, etc. Given that, in open-vocabulary semantic segmentation, CLIP's image embeddings transition from image level to pixel level, we infer that the performance improvement brought by the recent studies [6, 37] could be attributed to adjusting CLIP's text embeddings from their original hierarchical level to better align with the level represented by pixels, thereby enhancing cross-modal alignment. However, most existing fine-tuning methods operate in Euclidean space, which makes it challenging to quantify such a transition through different hierarchical levels. Hyperbolic spaces have gained significant interest in recent years, owing to their ability to naturally and compactly encode hierarchical structures [30, 31, 36], as shown in Fig. 1. The hyperbolic radius, defined as the distance from a point to the center in hyperbolic space, represents its hierarchical level [32]. Points closer to the center (smaller radius) correspond to low-level visual embeddings, e.g., those from pixels, while points near the edge (larger radius) correspond to high-level visual embeddings, e.g., those from images.

Motivated by this, we project the text embeddings into hyperbolic space and visualize the changes in their radii before and after fine-tuning, as shown in Fig. 2. This fine-tuning is realized using the state-of-the-art method provided in [6]. Specifically, one can observe that the radius after fine-tuning is smaller than that of pre-trained CLIP. In particular, we identify a pattern during fine-tuning: the hierarchical level of the text embeddings adjusts in conjunction with the hierarchical level of the embeddings provided by the image encoder. Our finding brings a novel perspective for fine-tuning CLIP: Adjusting the hyperbolic radius of text embeddings from edge to center endowed CLIP with segmentation ability.

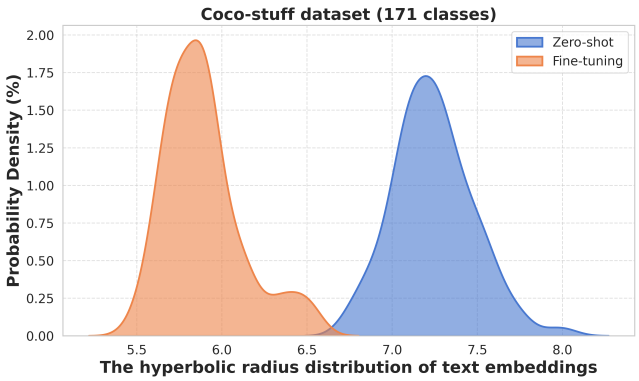

Figure 2. **Understanding fine-tuning CLIP's text encoder from the perspective of hyperbolic radius**. We use hyperbolic radius to illustrate the change in the hierarchical level of 171 classes after fine-tuning. We observe that the hyperbolic radius of each class becomes smaller, suggesting that adjusting the hyperbolic radius equips CLIP with segmentation ability (Best viewed in color).

Building on the analysis above, we propose a novel fine-tuning strategy for CLIP in hyperbolic space, termed HyperCLIP. HyperCLIP aims to directly adjust the hyperbolic radius of CLIP's embeddings to facilitate alignment with the appropriate hierarchical level for segmentation tasks. Specifically, we introduce several block-diagonal scaling matrices. Multiplying these matrices with CLIP's feature representations using a Möbius matrix multiplication operation is equivalent to performing scaling transformations that adjust the radius of CLIP's embeddings. This enables the adjustment of their hierarchical levels, ultimately enhancing the model's performance. As a result, HyperCLIP can equip CLIP with segmentation ability with only a small number of introduced learnable parameters.

Extensive experiments demonstrate that our method sets new state-of-the-art results in open-vocabulary semantic segmentation across three benchmarks, while fine-tuning only approximately 4% of the total parameters of CLIP. More importantly, we observe that after adjustment, CLIP's text embeddings maintain a relatively fixed hyperbolic radius across different datasets. This suggests that the granularity of the segmentation task might be quantified using the hyperbolic radius, warranting further exploration. The full version of this work is [28].

2. Methodology

In this section, we introduce HyperCLIP for open-vocabulary semantic segmentation. HyperCLIP is based on the Poincaré

Model	VLM	Additional Backbone	Fine-tuning Space	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
Partial Fine-Tuning									
ZS3Net [2]	-	ResNet-101	E	-	-	-	19.4	38.3	-
LSeg [17]	CLIP ViT-B/32	ResNet-101	E	-	-	-	-	47.4	-
OpenSeg [12]	ALIGN	ResNet-101	E	4.4	7.9	17.5	40.1	-	63.8
ZegFormer [7]	CLIP ViT-B/16	ResNet-101	E	4.9	9.1	16.9	42.8	86.2	62.7
ZSseg [38]	CLIP ViT-B/16	ResNet-101	E	7.0	-	20.5	47.7	88.4	-
OVSeg [22]	CLIP ViT-B/16	ResNet-101c	E	7.1	11.0	24.8	53.3	92.6	-
ZegCLIP [44]	CLIP ViT-B/16	-	E	-	-	-	41.2	93.6	-
Selective Fine-Tuning									
SED [37]	CLIP ConvNeXt-B	-	E	11.4	18.6	31.6	57.3	94.4	-
CAT-Seg [6]	CLIP ViT-B/16	-	E	<u>12.0</u>	<u>19.0</u>	<u>31.8</u>	<u>57.5</u>	<u>94.6</u>	<u>77.3</u>
Parameter-efficient Fine-Tuning									
SAN [39]	CLIP ViT-B/16	Side Adapter	E	10.1	12.6	27.5	53.8	94.0	-
Ours	CLIP ViT-B/16	-	**H**	**12.3**	**19.2**	**32.1**	**58.5**	**95.6**	**78.9**

Table 1. **Comparison with state-of-the-art methods on standard benchmarks.** The best-performing results are presented in bold, while the second-best results are underlined. "E": Euclidean Space. "H": Hyperbolic Space.

ball model, which aims to adjust the hyperbolic radius of CLIP's embeddings, thereby equipping CLIP with segmentation ability.

2.1. Fine-tuning in Hyperbolic Space

The core technique in HyperCLIP involves linear transformations, i.e., scaling, in the hyperbolic space. One of the primary methods [41] to learn the weights of linear transformations in hyperbolic space is hyperLoRA [41]. Given a feature embedding $\mathbf{z}$ in Euclidean space, this method first maps $\mathbf{z}$ to the hyperbolic space at a local reference point, typically the center, using a exponential map. Linear transformations with learnable weight matrix $\mathbf{W}$ is then applied to $\mathbf{z}$ within the hyperbolic space to obtain the hyperbolic representation. Then, a logarithmic map is applied to map the hyperbolic representation back to the Euclidean space. Formally, in the Poincaré ball model, this process is realized as:

$$\mathbf{W}\mathbf{z} = \log_{\mathbf{0}}^{\mathbb{D},c}(\mathbf{W} \otimes_c \exp_{\mathbf{0}}^{\mathbb{D},c}(\mathbf{z})), \tag{1}$$

where $\otimes_c$ is the möbius matrix multiplication operation. The primary aim of this method [41] is to capture more intricate hierarchical relationships within hyperbolic space. In contrast, our goal is to directly adjust the hyperbolic radius of CLIP's embeddings, leading to a scaling transformation.

2.2. Hyperbolic Radius Adjustment

The objective of HyperCLIP is to introduce a scaling transformation to directly adjust the hyperbolic radii of CLIP's embeddings in hyperbolic space, thereby equipping CLIP with segmentation ability. Concretely, it necessitates the utilization of a diagonal matrix, denoted as $\mathbf{S}$, to scale a CLIP's feature embedding $\mathbf{z} \in \mathbb{R}^{b\times d}$, where d is the feature dimension of $\mathbf{z}$, b is the spatial dimension of input images or number of classes. When extending this process to hyperbolic space, $\mathbf{S}$ can directly adjust the hyperbolic radius of $\mathbf{z}$. Specifically, we realize the diagonal matrix $\mathbf{S}$ in a block-wise manner,

$$\mathbf{S} = \mathrm{diag}(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_K), \tag{2}$$

where $\mathbf{S}_k \in \mathbb{R}^{n\times n}$ denotes a square matrix of the k-th block, and the block number K is calculated by $K = d/n$. By adjusting the size of n, we can achieve a transition between naive scaling transformation ($n = 1$, i.e., modifying only the diagonal elements) and arbitrary transformation ($n = d$, i.e., fully fine-tuning in hyperbolic space).

Overall Fine-tuning. During the fine-tuning phase, HyperCLIP introduces a similar process to Eq. 1 is incorporated into CLIP's encoder. Specifically, HyperCLIP uses the following forward pass:

$$\mathbf{S}\mathbf{z} = \log_{\mathbf{0}}^{\mathbb{D},c}(\mathbf{S} \otimes_c \exp_{\mathbf{0}}^{\mathbb{D},c}(\mathbf{z})). \tag{3}$$

Concurrently, the original components of CLIP load their weights from the pre-trained checkpoint, with their parameters remaining frozen.

3. Experiments

3.1. Experimental Setup

Datasets. Following prior works [6, 37], we use the COCO-Stuff dataset [3] as our training set. This dataset consists of around 118,000 densely annotated images, covering 171 distinct semantic categories. For inference, we compare our method against state-of-the-art approaches across several semantic segmentation benchmarks, including ADE20K [43], PASCAL VOC [10], and PASCAL-Context [24].

3.2. Comparisons with State-of-the-art Methods

Here, we compare our proposed method with several state-of-the-art methods, as shown in Table 1, using six test sets

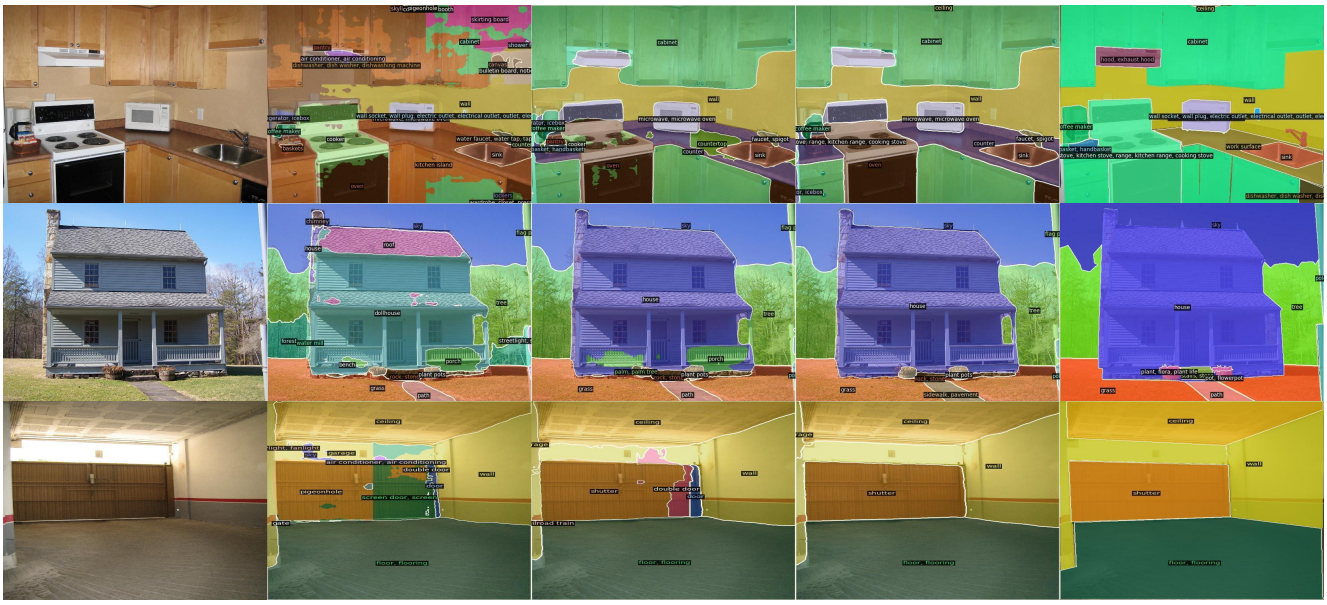

Image SAN [39] CAT-Seg [6] Ours Ground Truth

Figure 3. Comparison of qualitative reults on ADE20K [43] with 847 categories. We compare our method with other two state-of-the-art methods, i.e., SAN [39] and CAT-Seg [6]. The results show our method performs more precise segmentation in different scenarios.

across three benchmarks. Overall, we achieve the best results. Most existing open-vocabulary semantic segmentation methods follow partial fine-tuning strategies, i.e., fine-tuning CLIP's image encoder. Although these methods provide sufficient flexibility for aligning with text descriptions generated by the text encoder, they fail to adjust the text embeddings to the appropriate hierarchical level for effective text-to-image alignment, often leading to suboptimal segmentation performance. Differently, CAT-Seg [6] simultaneously fine-tunes both the text encoder and image encoder of CLIP, achieving performance comparable to ours on some of the datasets. However, its fine-tuning scheme is manually controlled through different layer combinations, necessitating a careful design to balance generalization and segmentation ability, while ours does not suffer from such an issue. Then, compared to SAN [39], another parameter-efficient fine-tuning method that introduces only a limited number of tunable parameters, our approach significantly outperforms it, achieving improvements of 6.6% on the `PC-459` dataset and 4.7% on the `PC-59` dataset with ViT-B/16 as the base model. These results demonstrate the effectiveness of our method in preserving generalization while mastering segmentation capability.

Qualitative results. Here, we visualize our method's representative example segmentation results against prevailing methods, i.e., CAT-Seg [6] and SAN [39], in the `A-847` dataset. As shown in Fig. 3, we can observe that our method can make more accurate predictions on object location and category. Even if objects are small and the background is complicated, our method still performs well.

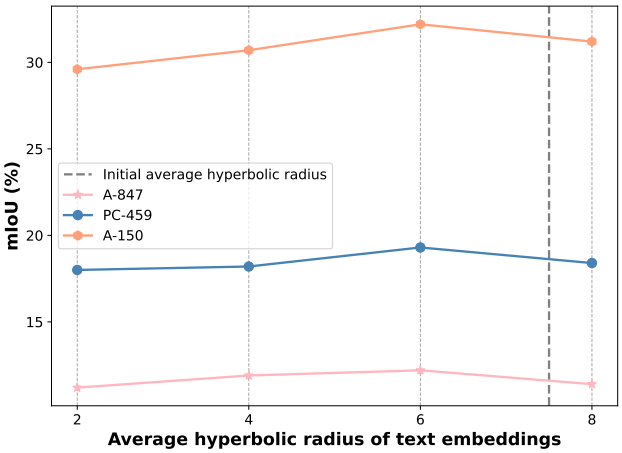

Figure 4. **Visualization on the effect of different average hyperbolic radii of text embeddings.** The controlling of different hyperbolic radii is achieved via a mean squared error (MSE) loss that enforces a constraint between the expected radius and the mean radius of the text embeddings during fine-tuning.

Acknowledgments

This work was supported by the NSFC under Grant 62322604, 62176159 and 62401361, and in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102.

References

[1] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2): 115, 1987. 2

[2] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 2, 3

[4] James W Cannon, William J Floyd, Richard Kenyon, Walter R Parry, et al. Hyperbolic geometry. *Flavors of geometry*, 31 (59-115):2, 1997. 2

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 1

[6] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation, 2024. 2, 3, 4

[7] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 3

[8] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Advances in neural information processing systems*, 35:32942–32956, 2022. 1

[9] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022. 1

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 3

[11] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, 2020. 1

[12] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 3

[13] Ming Hu, Kun Yuan, Yaling Shen, Feilong Tang, Xiaohao Xu, Lin Zhou, Wei Li, Ying Chen, Zhongxing Xu, Zelin Peng, et al. Ophclip: Hierarchical retrieval-augmented learning for ophthalmic surgical video-language pretraining. *arXiv preprint arXiv:2411.15421*, 2024. 1

[14] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985, 2021.

[15] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

[16] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021. 1

[17] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 2, 3

[18] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[20] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[21] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. 1

[22] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 2, 3

[23] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 1

[24] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. 3

[25] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Yaoming Wang, Lingxi Xie, Qi Tian, and Wei Shen. Parameter-efficient fine-tuning in hyperspherical space for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2405.18840*, 2024. 2

[26] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Lingxi Xie, Qi Tian, and Wei Shen. Parameter efficient fine-tuning via cross block orchestration for segment anything model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3743–3752, 2024. 1

[27] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Xiaokang Yang, and Wei Shen. Sam-parser: Fine-tuning sam efficiently by parameter space reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4515–4523, 2024. 1

[28] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Changsong Wen, Yu Huang, Menglin Yang, Feilong Tang, and Wei Shen. Understanding fine-tuning clip for open-vocabulary semantic segmentation in hyperbolic space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4562–4572, 2025. 2

[29] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[30] Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27263–27272, 2024. 2

[31] Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *International conference on machine learning*, pages 4460–4469. PMLR, 2018. 2

[32] Dídac Surís, Ruoshi Liu, and Carl Vondrick. Learning the predictability of the future. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12607–12617, 2021. 2

[33] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 1

[34] Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. Intervening anchor token: Decoding strategy in alleviating hallucinations for mllms. In *The Thirteenth International Conference on Learning Representations*.

[35] Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. Ufo: A unified transformer for vision-language representation learning. *arXiv preprint arXiv:2111.10023*, 2021. 1

[36] Melanie Weber, Manzil Zaheer, Ankit Singh Rawat, Aditya K Menon, and Sanjiv Kumar. Robust large-margin learning in hyperbolic space. *Advances in Neural Information Processing Systems*, 33:17863–17873, 2020. 2

[37] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation, 2023. 2, 3

[38] 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 *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 3

[39] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 2, 3, 4

[40] Haochen Xue, Feilong Tang, Ming Hu, Yexin Liu, Qidong Huang, Yulong Li, Chengzhi Liu, Zhongxing Xu, Chong Zhang, Chun-Mei Feng, et al. Mmrc: A large-scale benchmark for understanding multimodal large language model in real-world conversation. *arXiv preprint arXiv:2502.11903*, 2025. 1

[41] Menglin Yang, Aosong Feng, Bo Xiong, Jihong Liu, Irwin King, and Rex Ying. Hyperbolic fine-tuning for large language models. *arXiv preprint arXiv:2410.04010*, 2024. 3

[42] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In *NeurIPS*, 2023. 2

[43] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 3, 4

[44] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3