SAM-CLIP: Merging Vision Foundation Models towards Semantic and Spatial Understanding

Anonymous Author(s)
Affiliation
Address
email

Abstract

The landscape of publicly available vision foundation models (VFMs), such as CLIP and SAM, is expanding rapidly. VFMs are endowed with distinct capabilities stemming from their pretraining objectives. For instance, CLIP excels in semantic understanding, while SAM specializes in spatial understanding for segmentation. In this work, we introduce a simple recipe based on multi-task distillation to efficiently *merge* VFMs into a unified model that assimilates their expertise. By applying our method to SAM and CLIP, we derive SAM-CLIP : a unified model that amalgamates the strengths of SAM and CLIP into a *single backbone*, making it apt for edge device applications. We show that SAM-CLIP learns *richer visual representations*, equipped with both localization and semantic features, suitable for a broad range of vision tasks. We further show that SAM-CLIP not only retains the foundational strengths of its precursor models but also introduces *synergistic functionalities*, most notably in zero-shot semantic segmentation, where SAM-CLIP establishes new state-of-the-art results. It outperforms previous models that are specifically designed for this task by a large margin, including +6.8% and +5.9% mean IoU improvement on Pascal-VOC and COCO-Stuff datasets, respectively.

1 Introduction

Vision Foundation Models (VFM) such as CLIP [37], SAM [20], MAE [15], and DINOv2 [34] provide strong backbones that can be utilized for a wide range of vision tasks after finetuning. Additionally, some of these models exhibit notable zero-shot capabilities, such as classification from text prompts [37] and segmentation from geometric prompts (points and bounding boxes) [20]. Depending on their pretraining objectives, VFMs can act as feature extractors suitable for diverse downstream tasks. For instance, models that employ contrastive losses during training [6, 37, 34], utilize low-frequency signals, and generate features that can linearly separate samples based on their semantics [36]. Conversely, the pretraining objectives for MAE and SAM involve denoising masked images and instance mask segmentation, respectively, leading to the acquisition of features utilizing high-frequency signals with localization knowledge but limited semantic understanding (Figure 3).

Deploying separate models for different downstream tasks is inefficient (high memory footprint and runtime, especially on edge devices) and lacks opportunity for cross-model learning [42]. *Multitask learning* [52] is a paradigm capable of addressing this issue. However, it often requires costly training and simultaneous access to all tasks[11]. Training foundation models often relies on an unsupervised or semi-supervised approach, requiring substantial computational resources. For example, state-of-the-art CLIP models are trained on extensive datasets, such as LAION [43] and DataComp [12], consuming massive amount of computational power. Similarly, SAM's pretraining on 1.1 billion masks is computationally demanding. A multi-objective pretraining method requires comparable or more data and compute as single objective VFM training. This is in addition to other multi-task

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

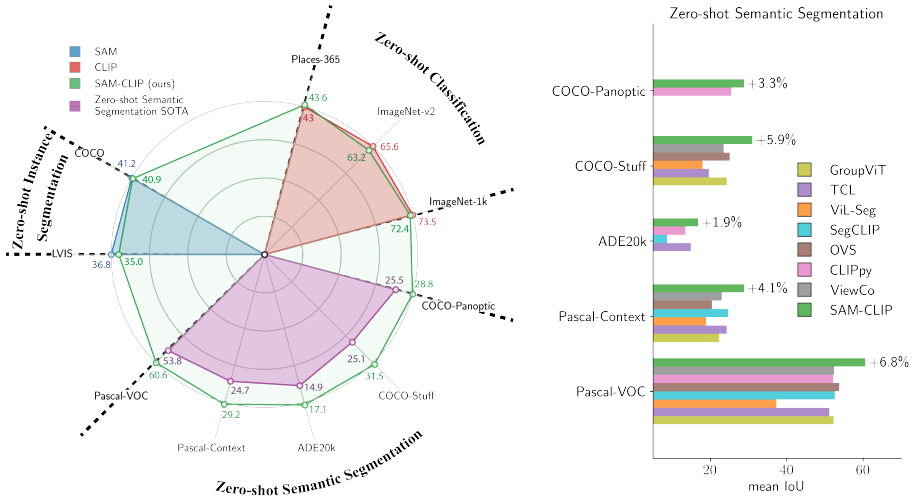


Figure 1: SAM-CLIP inherits zero-shot capabilities of SAM (instance segmentation) and CLIP (classification) using a single shared backbone (**left**). Further, SAM-CLIP is capable of a new task, zero-shot semantic segmentation, and obtains state-of-the-art results on several benchmarks (**right**).

learning challenges such as interfering gradients, training instabilities [9], and access to pretraining datasets that are often proprietary [37], which limit the scalability and feasibility of this approach.

To overcome these challenges, model merging has emerged as a rapidly growing area of research [46, 51]. The majority of merging techniques focus on combining multiple task-specific models into a single model without requiring additional training. For instance, this can be achieved through techniques such as model weights interpolation [17], parameter importance analysis [29], or leveraging invariances in the models [1]. These techniques, however, put too much stress on not using data or not performing additional training/finetuning resulting in decreased performance or lack of generalization to diverse set of tasks [46]. Our goal is to merge VFMs that are trained with fundamentally different objectives, have distinct capabilities, and possibly interact with other modalities. In this setup, naive merging approaches results in significant forgetting [30] (Appendix B).

We aim to fill the gap between training-free model merging and multitask training by drawing techniques from continual learning [24, 35] and knowledge distillation [16]. We treat model merging as a continual learning problem, where, given a pretrained base VFM, the knowledge of a second auxilary VFM is merged without forgetting of the initial knowledge. On one side, in contrast to weight averaging techniques, we allow access to *small part of* pretraining data or its surrogates during the merging process. We leverage multi-task distillation on the replay data to avoid forgetting the original knowledge of base VFM during the merging process. On the other side, our merging process is significantly more efficient than traditional multitask training by requiring less than 10% of the data and compute compared to their original pretraining (Section 2).

We instantiate our proposed merging approach by combining SAM and CLIP into a *single multi-task model*, called SAM-CLIP , suitable for edge device deployment. This merged model inherits prompt-based zero-shot capabilities from both CLIP and SAM with minimal forgetting: specifically, zero-shot classification and image-text retrieval from CLIP, and zero-shot instance segmentation from SAM (see Figure 1 left). Further, we illustrate that SAM-CLIP learns richer visual representations compared to SAM and CLIP, endowed with both spatial and semantic features, resulting in improved head-probing performance on new tasks (see Figure 3). Finally, SAM-CLIP shows an emerging capability of zero-shot transfer to a new task: *zero-shot semantic segmentation* thanks to combined skills inherited from SAM and CLIP. This task involves generating a segmentation mask based on a free-form text prompt. It requires both semantic understanding from text and segmentation capabilities, skills SAM-CLIP learns from CLIP and SAM, respectively. We demonstrate that SAM-CLIP achieves state-of-the-art performance on zero-shot semantic segmentation (Figure 1 right).

2 Proposed Approach

We constrain our discussion to the specific case where SAM serves as the base VFM, while a CLIP model serves as the auxiliary VFM. This pair presents an intriguing combination, as both models have

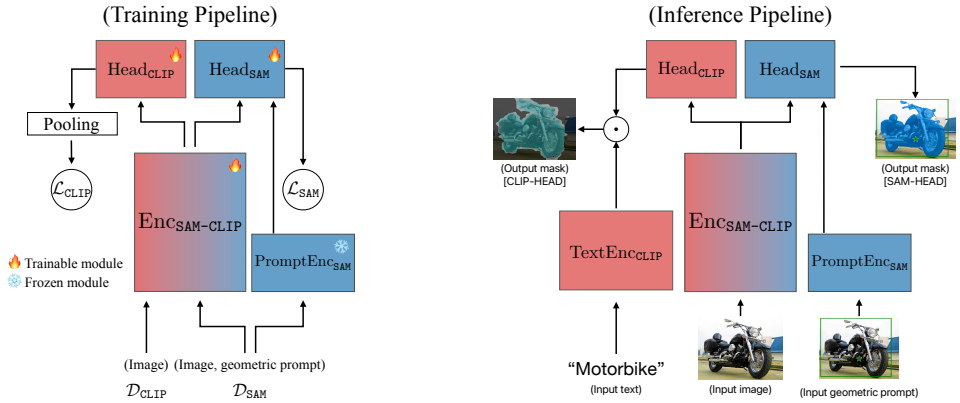


Figure 2: Multi-head architecture of SAM-CLIP for training (**left**) and inference (**right**).

been successfully deployed in diverse tasks and exhibit complementary capabilities. SAM excels in localization and high-resolution image segmentation but has limitations in semantic understanding. Conversely, CLIP offers a powerful image backbone for semantic understanding. We demonstrate it by several probing experiments (see Figure 3). We assume access to limited subsets of datasets (or their proxies) used to train the base and auxiliary VFMs, which function as memory replay in our CL setup. These are denoted as $\mathcal{D}_{\texttt{SAM}}$ and $\mathcal{D}_{\texttt{CLIP}}$.

We employ a multi-head architecture, illustrated in Figure 2. Our base VFM, SAM, has an image encoder ($\text{Enc}_{\texttt{SAM}}$), a prompt encoder ($\text{PromptEnc}_{\texttt{SAM}}$), and a light mask decoder ($\text{MaskDec}_{\texttt{SAM}}$). The auxiliary VFM, CLIP, has an image encoder ($\text{Enc}_{\texttt{CLIP}}$) and a text encoder ($\text{TextEnc}_{\texttt{CLIP}}$). Our goal is to merge both image encoders to a single backbone called $\text{Enc}_{\texttt{SAM-CLIP}}$ which is initialized by $\text{Enc}_{\texttt{SAM}}$. Further, we consider lightweight heads corresponding to each VFM, namely, $\text{Head}_{\texttt{SAM}}$ and $\text{Head}_{\texttt{CLIP}}$. $\text{Head}_{\texttt{SAM}}$ is initialized with $\text{MaskDec}_{\texttt{SAM}}$ and $\text{Head}_{\texttt{CLIP}}$ is initialized with random weights (since CLIP does not come with a head that we can deploy). We deploy other modality encoders (i.e., $\text{PromptEnc}_{\texttt{SAM}}$ and $\text{TextEnc}_{\texttt{CLIP}}$) with no change (frozen).

As a baseline merging approach, we perform KD on $\mathcal{D}_{\texttt{CLIP}}$ utilizing a cosine distillation loss [13]:

$$\mathcal{L}_{\texttt{CLIP}} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\texttt{CLIP}}}\left[1-\phi^{\text{Pooling}}(\text{Head}_{\texttt{CLIP}}(\text{Enc}_{\texttt{SAM-CLIP}}(\boldsymbol{x})))^T\text{Enc}_{\texttt{CLIP}}(\boldsymbol{x})\right], \tag{1}$$

where ϕ^{Pooling} is a pooling operator converting patch-level features from $\text{Head}_{\texttt{CLIP}}$ to a normalized image-level embedding. In this setup, parameters of both $\text{Head}_{\texttt{CLIP}}$ and $\text{Enc}_{\texttt{SAM-CLIP}}$ are learnable, while the CLIP encoder, $\text{Enc}_{\texttt{CLIP}}$, is frozen and used as a teacher. While this infuses SAM with CLIP's semantic abilities, it incurs at the cost of catastrophic forgetting of SAM's original capabilities even after deploying mitigative methods such as Wise-FT [48] (see supplementary materials).

To address these challenges, we propose a rehearsal-based multi-task distillation. This serves two primary goals: 1) facilitate the efficient transfer of knowledge from the auxiliary VFM to the base model, and 2) preserve the original capabilities of the base model. Inspired by [21], we consider a two-stage training: head-probing and multi-task distillation.

I. Head probing: In this stage, we first freeze the image backbone, $\text{Enc}_{\texttt{SAM-CLIP}}$, and only train $\text{Head}_{\texttt{CLIP}}$ with the loss in Equation (1). Intuitively, with this approach we first learn some reasonable values for parameters of $\text{Head}_{\texttt{CLIP}}$ (which is initialized randomly) before allowing any change in $\text{Enc}_{\texttt{SAM-CLIP}}$ that is prone to forgetting.

II. Multi-task distillation: In this stage, we allow all heads as well as our image encoder to be learnable. We perform a multi-task training on $\mathcal{L}_{\texttt{CLIP}} + \lambda\mathcal{L}_{\texttt{SAM}}$, with:

$$\mathcal{L}_{\texttt{SAM}} = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{g})\sim\mathcal{D}_{\texttt{SAM}}}\mathcal{L}_{\text{FD}}(\text{Head}_{\texttt{SAM}}(\text{Enc}_{\texttt{SAM-CLIP}}(\boldsymbol{x}),\text{PromptEnc}_{\texttt{SAM}}(\boldsymbol{g})),\boldsymbol{z}), \tag{2}$$

where, $\boldsymbol{x}$ is raw image, $\boldsymbol{g}$ is a geometric prompt, $\boldsymbol{z} = \text{MaskDec}_{\texttt{SAM}}(\text{Enc}_{\texttt{SAM}}(\boldsymbol{x}))$ is segmentation mask score produced by frozen SAM teacher, and $\mathcal{L}_{\text{FD}}$ refers to a linear combination of Focal [25] and Dice [32] used in the original SAM training adapted for distillation. We train on $\mathcal{D}_{\texttt{SAM}} \cup \mathcal{D}_{\texttt{CLIP}}$ with total loss of $\mathcal{L}_{\texttt{CLIP}} + \lambda\mathcal{L}_{\texttt{SAM}}$. During training, each batch has some samples from $\mathcal{D}_{\texttt{CLIP}}$ and some form $\mathcal{D}_{\texttt{SAM}}$, which contribute to $\mathcal{L}_{\texttt{CLIP}}$ and $\mathcal{L}_{\texttt{SAM}}$, respectively. To encourage less forgetting we use an order of magnitude smaller learning rate for parameters of $\text{Enc}_{\texttt{SAM-CLIP}}$ and $\text{Head}_{\texttt{SAM}}$ compared to $\text{Head}_{\texttt{CLIP}}$ at this stage.

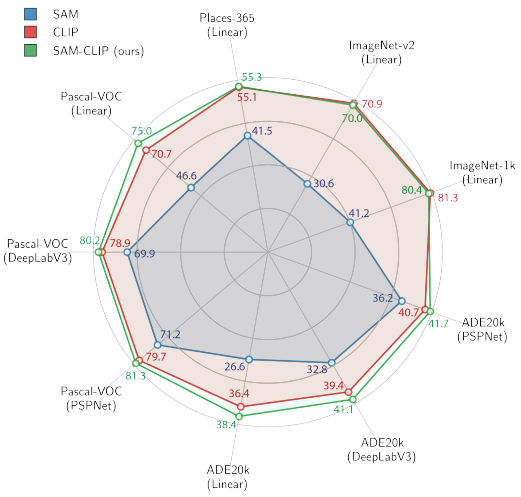
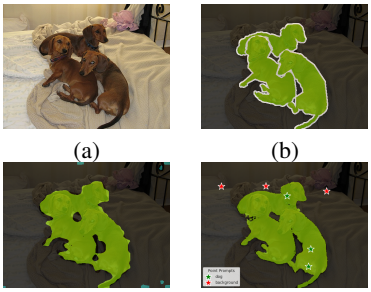


Figure 3: Head-probing evaluation of each vision backbone for classification and semantic segmentation tasks demonstrating enriched visual features of SAM-CLIP .



(a) (b) (c) (d)

Figure 4: Passing an input image through the image encoder (a), $\text{Head}_{\texttt{CLIP}}$ can predict a semantic segmentation mask (c), and $\text{Head}_{\texttt{SAM}}$ can refine it to a more fine-grained mask with auto-generated geometric prompts (d) matching ground-truth (b).

3 Experiments

Experimentation details are presented in the supplementary materials.

Zero-Shot Image Classification. To examine the CLIP-related capabilities of SAM-CLIP, we perform zero-shot image classification on ImageNet [8], ImageNet-v2 [39] and Places365 [54]. Results shown in Figure 1 validate the efficacy of our approach in inheriting CLIP's capabilities.

Zero-Shot Instance Segmentation. For the SAM component of SAM-CLIP , we evaluate its performance in instance segmentation, a task at which the original SAM model excels [20], with COCO [26] and LVIS [14] datasets. Results (Figure 1) show that SAM-CLIP is close to the original SAM ViT-B on the two benchmarks, not suffering from catastrophic forgetting.

Zero-Shot Transfer to Semantic Segmentation. We extend our evaluation to (text-prompted) zero-shot semantic segmentation over 5 datasets, Pascal VOC [10], Pascacl Context [33], ADE20k [55], COCO-Stuff [2] and COCO-Panoptic [19, 26]. SAM-CLIP establishes new state-of-the-art performance on all 5 datasets as shown in Figure 1 (right).

Composing Both CLIP and SAM Heads for Better Segmentation. Given that SAM-CLIP is a multi-task model with SAM and CLIP heads, one would naturally ask if the two heads can work together towards better performance on some tasks. Here, we showcase that a simple composition of SAM-CLIP 's CLIP and SAM heads (low-resolution mask from CLIP head followed by high-resolution refinement by SAM head) can lead to even better zero-shot semantic segmentation. Example of this pipeline is shown at Figure 5. For fair comparison, when we compare with previous works in Figure 1 we report SAM-CLIP zero-shot segmentation performance with 448px resolution using $\text{Head}_{\texttt{CLIP}}$ only. Using our high-resolution pipeline we obtain further gain: for example **mIoU on Pascal-VOC increases from 60.6% to 66.0%**.

Head-Probing Evaluations on Learned Representations. By merging the SAM and CLIP models, we anticipate that the resultant model will inherit advantages at the representation level from both parent models. Specifically, SAM excels at capturing low-level spatial visual details pertinent to segmentation tasks, while CLIP specializes in high-level semantic visual information encompassing the entire image. We hypothesize that the merged model combines these strengths, thereby enhancing its utility in broad range of downstream vision tasks. To investigate this hypothesis, we conduct head-probing (i.e., learn a task specific head with a frozen image backbone) evaluations on SAM, CLIP, and SAM-CLIP, utilizing different segmentation head structures (linear head, DeepLab-v3 [5] and PSPNet [53]) across two semantic segmentation datasets, Pascal-VOC and ADE20k, and linear probing for image classification task on ImageNet and Places365 datasets. Results are presented in Figure 3 demonstrating SAM-CLIP superior visual feature representation capabilities.

References

[1] Samuel K Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022.

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

[3] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11165–11174, 2023.

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[9] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

[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.

[11] Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34:27503–27516, 2021.

[12] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

[13] 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. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[14] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019.

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[17] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. *Advances in Neural Information Processing Systems*, 35:29262–29277, 2022.

[18] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021.

[19] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019.

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[21] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

[22] Xianhang Li, Zeyu Wang, and Cihang Xie. An inverse scaling law for clip training. *NeurIPS*, 2023.

[23] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022.

[24] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[27] Quande Liu, Youpeng Wen, Jianhua Han, Chunjing Xu, Hang Xu, and Xiaodan Liang. Open-world semantic segmentation via contrasting and clustering vision-language embedding. In *European Conference on Computer Vision*, pages 275–292. Springer, 2022.

[28] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pages 23033–23044. PMLR, 2023.

[29] Michael S Matena and Colin A Raffel. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.

[30] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[31] Sachin Mehta, Farzad Abdolhosseini, and Mohammad Rastegari. Cvnets: High performance library for computer vision. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 2022.

[32] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[33] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[34] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[35] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.

[36] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? In *The Eleventh International Conference on Learning Representations*, 2022.

[37] 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.

[38] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models. ICCV, 2023.

[39] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

[40] Pengzhen Ren, Changlin Li, Hang Xu, Yi Zhu, Guangrun Wang, Jianzhuang Liu, Xiaojun Chang, and Xiaodan Liang. Viewco: Discovering text-supervised segmentation masks via multi-view semantic consistency. *arXiv preprint arXiv:2302.10307*, 2023.

[41] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[42] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

[43] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[44] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

[45] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[46] Yi-Lin Sung, Linjie Li, Kevin Lin, Zhe Gan, Mohit Bansal, and Lijuan Wang. An empirical study of multimodal model merging. *arXiv preprint arXiv:2304.14933*, 2023.

[47] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[48] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022.

[49] 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.

[50] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, and Weidi Xie. Learning open-vocabulary semantic segmentation models from natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2935–2944, 2023.

[51] Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Resolving interference when merging models. *arXiv preprint arXiv:2306.01708*, 2023.

[52] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.

[53] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[54] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[55] 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.