Perceptual Grouping in Contrastive Vision-Language Models

Kanchana Ranasinghe*, Brandon McKinzie, Sachin Ravi,
Yinfei Yang, Alexander Toshev, Jonathon Shlens†
Apple
kranasinghe@cs.stonybrook.edu

Abstract

Recent advances in zero-shot image recognition suggest that vision-language models learn generic visual representations with a high degree of semantic information that may be arbitrarily probed with natural language phrases. Understanding an image, however, is not just about understanding what content resides within an image, but importantly, where that content resides. In this work we examine how well vision-language models are able to understand where objects reside within an image and group together visually related parts of the imagery. We demonstrate how contemporary vision and language representation learning models based on contrastive losses and large web-based data capture limited object localization information. We propose a minimal set of modifications that results in models that uniquely learn both semantic and spatial information. We measure this performance in terms of zero-shot image recognition, unsupervised bottom-up and top-down semantic segmentations, as well as robustness analyses. We find that the resulting model achieves state-of-the-art results in terms of unsupervised segmentation, and demonstrate that the learned representations are uniquely robust to spurious correlations in datasets designed to probe the causal behavior of vision models.

1. Introduction

Recent vision-language models trained under weak supervision demonstrate a remarkable ability to learn generic and transferable visual representations [21, 43, 68, 95], but showcase a profound inability to associate visual content with individual objects (Fig. 1, bottom row). In other words, models trained on large weakly-supervised data have a limited ability to group together visually related content [31]. Because the representations have a poor understanding of *where* an object resides, they easily conflate background with foreground content. Hence, the learned representations are unable to learn the spatial layout of a scene [77, 80], and are susceptible to learning spurious correlations between a semantic label and extraneous content [55, 73].

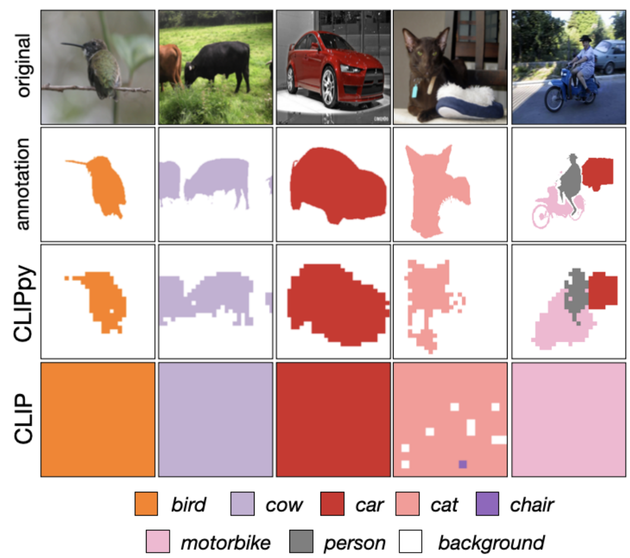


Figure 1. **Semantic localization in vision-language models.** We measure the ability of vision-language models to predict a label at each spatial position in a zero shot manner based on the similarity of location tokens to the corresponding language tokens on selected examples. CLIP / ALIGN [43, 68] have minimal understanding of the spatial location of individual objects (row 4). Our proposed CLIPpy (row 3) predicts the label at locations that correspond closely to human annotation for semantic segmentation (row 2). All predictions were performed with no access to any segmentation data during training or inference. More visualizations in App. B.

Recent work [91, 92] attempts to bridge this gap through grouping mechanisms under the same weakly supervised training paradigm, but focus more on foreground objects (neglecting background classes). Another direction is task specific unsupervised fine-tuning [23, 103] which loses the generic and transferable nature of these representations.

In this work, we explore vision-language models that learn from similar weakly labeled data, but a) retain the generic and transferable nature of features, and b) learns where all (background & foreground) visual content resides within an image. Unlike previous works using grouping specific architectures [91, 92] or dense human annota-

*Work performed as part of Apple internship.
†Work performed at Apple.

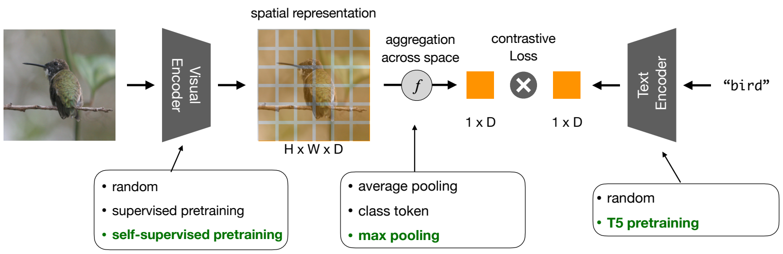


Figure 2. **Architecture diagram.** We demonstrate that two minimal design decisions (indicated in green) are of paramount importance for CLIP [68] models to perform perceptual grouping under image-level weak supervision.

Component	CLIP [68]	CLIP†	CLIPpy
Image Backbone	ViT-B/16	ViT-B/16	ViT-B/16
Text Backbone	T-B	T-5	T-5
Image Init	Random	Random	DINO
Text Init	Random	Random	Sent T-5
Image Pooling	CLS	CLS	Max
Text Pooling	Avg	Avg	Avg
Dataset	300M*	CC-12M	CC-12M
VOC mIoU (%)	16.4	17.5	**50.8** (+33.3)
VOC JS (%)	28.6	37.3	**47.5** (+10.2)

Figure 3. We highlight CLIPy differences from CLIP. CLIP† is our implementation following train settings identical to CLIPpy. *indicates OpenAI private data.

tions [31, 33, 48], we explore a minimal set of modifications to existing CLIP models [68] that leads to visual grouping while retaining their weakly supervised and scalable training procedure. We find that two small adjustments – employing specific pretraining strategies and adjusting spatial feature aggregation – results in models that are equally effective in zero-shot image recognition, but also retain spatial information regarding object locations (see Fig. 1, 3rd row).

The resulting model termed CLIPpy exhibits *perceptual grouping*: ability to select and combine related visual signals into semantically meaningful regions [60, 72, 89]. Endowing models with perceptual grouping – whether in a bottom up (based solely on visual content) or top down (guided by external information, language in this case) manner – in learned representations has been a long standing goal in computer vision [58, 59]. Further, this emergence of localization ability uniquely leads to robustness to counterfactual manipulations.

2. Methodology

Our work builds on CLIP [68], introducing two key modifications that emerge grouping behavior: alternate aggregation options and pre-training strategies (Fig. 2). Training uses the same contrastive objective under weak supervision.

2.1. Aggregation

The goal of aggregation is to collapse the image embedding from $[H, W, D]$ to D dimension. **Average pooling** across space is an established technique to obtain a final embedding independent of image resolution [57, 78]. Another approach for ViT is **class token (CLS)**, which learns an aggregated embedding from patch tokens. In this work we systematically explore these aggregation strategies to find that the application of **maximum pooling** across the spatial dimensions – while extremely simple – is also by far most effective (Tab. 4). We hypothesize that this may be due to gradient updates being focused solely on a single spatial location, and not spread across all spatial dimensions.

2.2. Pretraining

Language Model. For better sentence level representation, we utilize pre-training from Sentence-T5 [64], selected over auto-regressive models [6, 22] because their contrastive loss is aligned to our setup.

Image Model. We investigate using supervised and self-supervised ([10]) pre-training strategies. We focus on the latter direction due to impressive localization performance of those features [10, 36]. The visual encoder representation space can be viewed as containing per-image features (post-aggregation) vs per-spatial location features (pre-aggregation). We hypothesize that semantics tied boundaries of this representation space should operate at the latter granularity to induce perceptual grouping. Furthermore, we suggest that initializations facilitating the former will detriment grouping behaviour. In particular, visual pre-training strategies separating image-level representations by semantics (e.g. supervised ImageNet pre-training) will diminish perceptual grouping. Self-supervised pre-training strategies focused on more granular within image representations (e.g. [10]) will tend to enhance perceptual grouping. This hypothesis is empirically validated in ablations (see Table 4).

2.3. Visual Token Sub-Sampling

Motivated by vision transformers' ability to process sequences of length different to train time, we generate higher resolution segmentations during inference by sampling more image patches. In order to increase robustness to such varying resolution, we utilize up to $2\times$ higher resolution images during training but randomly drop 80% of visual tokens to minimize additional compute overhead (similar to [38, 52]). While improving inference quality, this also provides training stability possibly due to its regularization effect (see App. E for more details).

2.4. Inference

Classification and top-down grouping follows zero-shot analyses from [68] (see App. I) with latter doing at each patch similar to [30] . Bottom-up grouping follows analysis in DINO demos [10]. PCA on image features (from visual encoder pre-aggregation) gives top n (=8) principal components, used as cluster centers. Each same image feature is assigned to one of these n clusters based on proximity (cosine similarity) to the centers resulting in n clusters (groups).

	Dataset	IN	IN-v2
ALIGN [43]	ALIGN-1800M	76.4	70.1
CLIP [68]	CLIP-400M	65.5	60.8
CLIP †	CC-12M	46.0	40.3
GroupViT [91]	CC-12M+YFCC	42.9	-
GroupViT †	CC-12M	25.6	23.8
CLIPpy	CC-12M	45.3	40.0
ALIGN †	HQITP-134M	51.1	45.6
CLIP †	HQITP-134M	61.4	56.4
CLIPpy	HQITP-134M	60.3	54.8

Table 1. **CLIPpy achieves competitive zero-shot image recognition.** IN and IN-v2 denote ImageNet and ImageNet-v2 accuracy, respectively. † indicates our implementation. [43] evaluated at 640×640; others evaluated at 224×224. CLIPpy shows ±0.5 and ±0.9 IN acc. (5 runs) on CC-12M and HQITP-134M, respectively.

	Dataset	Train	VOC
MoCo [91]	ImageNet	self	28.2
DINO [10]		self	45.9
DSM [61]		self	37.2
COMUS [97]		self	47.3
DINO [91]	CC-12M & YFCC-100M	self	41.8
CLIP [91]		text	28.6
GroupViT [91]		text	51.8
CLIP †	CC-12M	text	37.3
GroupViT †		text	42.8 (+5.5)
CLIPpy		text	**47.5** (+10.2)
CLIP †	HQITP-134M	text	38.9
CLIPpy		text	**54.6** (+15.7)

Table 2. **CLIPpy effectively performs bottom-up grouping.** We report Jaccard Similarity, an instance average of the IoU between proposed and annotated segmentations, independent of the object label. † denotes our implementation.

	arch	dataset	ADE20K	COCO	VOC	COCO (obj)
GroupViT † [91]	ViT	CC-12M	6.2	12.7	40.1	17.5
MaskCLIP † [103]	ViT		6.8	8.1	22.1	13.8
OVS [92]	ViT		7.1	-	44.6	25.1
CLIP †	ViT		5.0	7.8	17.5	13.2
CLIPpy	ViT		**13.1** (+8.1)	**23.8** (+16.0)	**50.8** (+33.3)	**28.5** (+15.3)
ALIGN [30]	CNN	ALIGN-1800M	9.7	15.6	-	-
CLIP [68]	ViT	CLIP-400M	5.8	8.7	16.4	-
ALIGN †	CNN	HQITP-134M	7.5	14.4	29.7	-
CLIP †	ViT		5.1	8.0	18.1	-
CLIPpy	ViT		**13.5** (+8.4)	**25.5** (+17.5)	**52.2** (+34.1)	32.0

Table 3. **CLIPpy provides competitive top-down grouping (semantic segmentation) with no segmentation or location annotations.** All models trained without any segmentation annotations. Results grouped by training dataset (bold highlights best per dataset). Numbers are mean IoU. † indicates our implementation.

3. Experiments

Experimental Setup. We train our models on two datasets: Conceptual Captions 12M (CC-12M) [11] and High Quality Image Text Pairs (HQITP-134M) consisting of 12 million and 134 million image-text pairs, respectively (App. C for details). More training details in App. E.

We first measure performance of CLIP [68] and ALIGN [43] on zero-shot image classification on ImageNet and ImageNet-v2. Table 1 highlights these results. In the following experiments we attempt to address the following questions:

- Limitations of current vision-language models?
- Do we observe perceptual grouping in these models? (Tabs. 2 and 3).
- How resilient are they to counterfactual manipulations? (Fig. 5).

Finally, we report ablations on each component in Tab. 4.

3.1. Limitations of vision-language models

Visual representations learned in vision-language models exhibit an impressive ability to generalize across tasks [43,68]. However they also exhibit a profound shortcoming – learned visual representations maintain minimal information about *where* an object resides, failing to properly recognize what parts of an image constitute an object. Fig. 1 (bottom row) showcases failure of a CLIP model; namely, the model improperly conflates visual content not associated with an object with the actual object. One consistently observes the central object of interest being incorrectly predicted to reside at every spatial location. This failure of vision-language models to properly understand spatial organization of information is consistent with earlier observations [63,77,99].

In contrast, if we perform the same analysis on CLIPpy, we see that the model retains significant information about spatial information (Fig. 1, 3rd row). We take these visualizations as an impetus for further investigation.

3.2. Emergence of Bottom-Up Grouping

Unsupervised segmentation performance is a direct measure of bottom up perceptual grouping. We apply CLIPpy at test time to perform segmentation without prompting for any labels. Fig. 4 shows how image embeddings naturally group into spatially distinct clusters mirroring the image structure. We emphasize that this analysis does *not* rely on text prompts *nor* segmentation labels, but merely emerges from the image features alone. Hence the model has learned to *group* perceptually related pixels merely based on the pixel content and image-level captions during training. We quantify the accuracy of this bottom-up segmentation using Jaccard Index [10,91], and show results in Tab. 2. Our intu-

Figure 4. **Visualizations of bottom-up grouping by CLIPpy**. Each color represents one grouping learned on a given image.

ition is that CLS and average pooling breaks spatial structure of features and mixes image-level features across features at all spatial locations unlike in CLIPpy, leading to observed better bottom-up grouping.

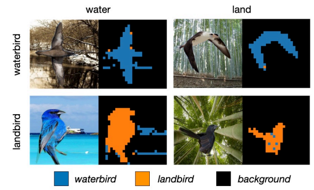
3.3. Emergence of Top-down Grouping

We next measure top-down grouping ability with zero-shot semantic segmentation. Fig. 1 visualizes predicted zero-shot segmentations and Tab. 3 quantifies these results using mean intersection over union (mIoU). CLIPpy outperforms all other approaches on semantic segmentation when trained on the same datasets, both for CC-12M and HQITP-134M. GroupViT [91] and OVS [92] containing grouping specific architectures and pre-training strategies provide important points of comparison. CLIPpy obtains clear performance improvements over these methods across all datasets.

3.4. Perceptual grouping may improve robustness

We next explore how observed perceptual grouping could improve robustness of image understanding. The synthetic benchmark is *Waterbirds* [73] places segmented birds over land or water background posing a two-way classification task of whether a bird belongs to the `waterbird` or `landbird` category. What makes this problem particularly challenging is when the background is not commensurate with the type of bird (e.i. landbird in water background).

We evaluate CLIPpy and a baseline CLIP model and report results in Fig. 5. Given the structured outputs from grouping, we perform a specialized inference procedure (App. H for details). For the CLIP baseline, model performance depends heavily on the background (δ column), indicating reliance on background features for prediction. Meanwhile, CLIPpy, while still susceptible to some spurious correlations, is far more robust than baseline CLIP. As points of comparison, all prior work train a supervised model on the training split. In contrast, our predictions are zero-shot, and we do not use the training set. That said, the best supervised training methods achieve a domain gap Δ of 4% to 8% [55], comparable to our results. We take these results to indicate that our zero-shot approach leveraging perceptual grouping provides another approach for addressing spurious correlations and learning robust image features.



CLIP	water	land	Δ
waterbird	80.2	48.1	**-32.1**
landbird	38.8	71.7	**-32.9**

CLIPpy	water	land	Δ
waterbird	76.9	74.9	**-2.0**
landbird	80.0	84.1	**-4.1**

Figure 5. **Perceptual grouping mitigates sensitivity to spurious correlations.** (left) Selected segmentation examples by CLIPpy of waterbirds and landbirds on each background. (right) Accuracy on the *test* split (5794 examples) of Waterbirds on CLIP and CLIPpy evaluated at 448×448 resolution. The domain gap Δ reports the drop in accuracy between on and off diagonal entries within a row.

I-P	I-F	T-F	IN	VOC
Cls	✓	✗	39.9	3.4
Max	✓	✗	24.2	10.4
Max	✗	✓	35.9	29.5
Max	✗	✗	**42.3**	**50.8**

Aggreg.	ImageNet	Pascal VOC	
	Accuracy	mIoU	Jaccard
Max	42.3	**50.8**	**47.5**
Avg	44.0	11.6	38.1
Cls	46.0	4.0	40.4

Table 4. **Ablation studies:** *(left)* Freezing pre-trained backbones. I-P for image backbone pooling, I-F for image backbone frozen, and T-F for text backbone frozen. Top-1 accuracy (%) for ImageNet (IN) and mean IoU for VOC reported. *(right)* Aggregation methods: max pooling (Max), average pooling (Avg), and class token (Cls). All models initialized with the same pre-trained features.

4. Discussion

In this work we demonstrate how contrastive vision-language models have a profound lack of understanding object location. We described a minimal set of changes to existing models - modifying the aggregation method and introducing optimal pre-training strategies - to endow the model with both bottom-up and top-down perceptual grouping. We emphasize that our changes are minimal but sufficient to match if not exceed the performance of custom-built architectures [91, 92] in achieving perceptual grouping.

We establish how our resulting model provides state-of-the-art results in terms of both bottom-up and top-down grouping – even though the model has been afforded *no* segmentation annotations whatsoever. Finally, we demonstrate the utility of these representations by demonstrating how perceptual grouping may be leveraged to learn visual features that are robust to spurious correlations.

We take these results to indicate that contrastive vision-language models may provide the emergence of perceptual grouping without supervision. We do see limitations in this approach as segmentation suffers with increasing visual clutter and label cardinality (e.g. ADE-20K). We suspect that recent advent of larger-scale open datasets [8, 74] and advanced in self-supervised learning [36, 61] may offer opportunities to demonstrate further benefits for endowing models with perceptual grouping. We also note the possibility of biases in our training data that may be reflected in our models. We highlight how all reported numbers contain an equivalent version on CC-12M only to enable reproducibility.

References

[1] Pablo Arbeláez, Bharath Hariharan, Chunhui Gu, Saurabh Gupta, Lubomir Bourdev, and Jitendra Malik. Semantic segmentation using regions and parts. In *CVPR*, pages 3378–3385. IEEE, 2012. 14

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 14

[3] Donghyeon Baek, Youngmin Oh, and Bumsub Ham. Exploiting a joint embedding space for generalized zero-shot semantic segmentation. In *ICCV*, 2021. 14

[4] Zhipeng Bao, Pavel Tokmakov, A. Jabri, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. Discovering objects that can move. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11779–11788, 2022. 14

[5] Irwan Bello, William Fedus, Xianzhi Du, Ekin Dogus Cubuk, Aravind Srinivas, Tsung-Yi Lin, Jonathon Shlens, and Barret Zoph. Revisiting resnets: Improved training and scaling strategies. *NeurIPS*, 34:22614–22627, 2021. 13

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. 2

[7] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *NeurIPS*, 2019. 14

[8] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022. 4

[9] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision*, 2018. 9

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *CVPR*, pages 9650–9660, 2021. 2, 3, 9, 12, 13, 14

[11] 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*, pages 3558–3568, 2021. 3, 12

[12] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *CVPR*, pages 15789–15798, 2021. 12

[13] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 2017. 13

[14] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *arXiv*, 2021. 13

[15] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16794–16804, 2021. 9, 14

[16] Dorin Comaniciu and Peter Meer. Robust analysis of feature spaces: Color image segmentation. In *CVPR*, pages 750–755. IEEE, 1997. 14

[17] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 13

[18] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *ICML*, pages 2189–2200. PMLR, 2021. 14

[19] Yufeng Cui, Lichen Zhao, Feng Liang, Yangguang Li, and Jing Shao. Democratizing contrastive language-image pre-training: A clip benchmark of data, model, and supervision. *arXiv preprint arXiv:2203.05796*, 2022. 13

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 13, 14

[21] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *CVPR*, pages 11162–11173, 2021. 1

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 12

[23] Jian Ding, Nan Xue, Guisong Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11573–11582, 2021. 1, 13, 14

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 11, 12, 13

[25] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, Jianfeng Gao, and Lijuan Wang. Coarse-to-fine vision-language pre-training with fusion in the backbone. *ArXiv*, abs/2206.07643, 2022. 13

[26] Gamaleldin F. Elsayed, Aravindh Mahendran, Sjoerd van Steenkiste, Klaus Greff, Michael Curtis Mozer, and Thomas Kipf. Savi++: Towards end-to-end object-centric learning from real-world videos. *ArXiv*, abs/2206.07764, 2022. 14

[27] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 13, 15

[28] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic

segmentation by contrasting object mask proposals. *ICCV*, pages 10032–10042, 2021. 14

[29] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *NeurIPS*, 34:23885–23899, 2021. 14

[30] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*, 2021. 2, 3, 12, 14

[31] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. In *ECCV*, 2022. 1, 2, 13

[32] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, June 2014. 14

[33] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Zero-shot detection via vision and language knowledge distillation. *arXiv e-prints*, pages arXiv–2104, 2021. 2, 13

[34] Zhangxuan Gu, Siyuan Zhou, Li Niu, Zihan Zhao, and Liqing Zhang. Context-aware feature generation for zero-shot semantic segmentation. In *ACM MM*, pages 1921–1929, 2020. 14

[35] Zhangxuan Gu, Siyuan Zhou, Li Niu, Zihan Zhao, and Liqing Zhang. From pixel to patch: Synthesize context-aware features for zero-shot semantic segmentation. *arXiv:2009.12232*, 2020. 14

[36] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *ICLR*, 2022. 2, 4, 9, 14

[37] David F. Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James R. Glass. Jointly discovering visual objects and spoken words from raw sensory input. *IJCV*, 128:620–641, 2019. 9

[38] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2021. 2

[39] Olivier J. H'enaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelovi'c. Object discovery and representation networks. In *European Conference on Computer Vision*, 2022. 14

[40] Ping Hu, Stan Sclaroff, and Kate Saenko. Uncertainty-aware learning for zero-shot semantic segmentation. In *NeurIPS*, 2020. 14

[41] Jingwei Ji, Shyamal Buch, Alvaro Soto, and Juan Carlos Niebles. End-to-end joint semantic segmentation of actors and actions in video. In *ECCV*, pages 702–717, 2018. 14

[42] Xu Ji, Andrea Vedaldi, and João F. Henriques. Invariant information clustering for unsupervised image classification and segmentation. *ICCV*, pages 9864–9873, 2019. 9, 14

[43] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *ICML*, 2021. 1, 3, 12, 13

[44] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, pages 1780–1790, 2021. 13

[45] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv:1703.04977*, 2017. 14

[46] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, pages 5637–5664. PMLR, 2021. 14

[47] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *CVPR*, pages 2661–2671, 2019. 14

[48] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *ICLR*, 2022. 2, 13, 14

[49] Jiahao Li, Greg Shakhnarovich, and Raymond A. Yeh. Adapting clip for phrase localization without further training. *ArXiv*, abs/2204.03647, 2022. 13, 14

[50] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10955–10965, 2021. 13

[51] Peike Li, Yunchao Wei, and Yi Yang. Consistent structural relation learning for zero-shot segmentation. *NeurIPS*, 2020. 14

[52] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *CVPR*, 2023. 2

[53] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 13

[54] 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 *ECCV*, 2014. 13, 15

[55] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *ICML*, pages 6781–6792. PMLR, 2021. 1, 4, 14

[56] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *ArXiv*, abs/2006.15055, 2020. 14

[57] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[58] Jitendra Malik. Visual grouping and object recognition. In *Proceedings 11th International Conference on Image Analysis and Processing*, pages 612–621. IEEE, 2001. 2, 14

[59] Jitendra Malik, Pablo Arbeláez, João Carreira, Katerina Fragkiadaki, Ross Girshick, Georgia Gkioxari, Saurabh Gupta, Bharath Hariharan, Abhishek Kar, and Shubham Tulsiani. The three R's of computer vision: Recognition, reconstruction and reorganization. *Pattern Recognition Letters*, 72:4–14, 2016. 2, 14

[60] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 1982. 2

[61] Luke Melas-Kyriazi, C. Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8354–8365, 2022. 3, 4, 14

[62] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *NeurIPS*, 33:20673–20684, 2020. 14

[63] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *NeurIPS*, 2021. 3

[64] Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*, 2021. 2

[65] Giuseppe Pastore, Fabio Cermelli, Yongqin Xian, Massimiliano Mancini, Zeynep Akata, and Barbara Caputo. A closer look at self-training for zero-label semantic segmentation. In *CVPRW*, pages 2693–2702, 2021. 14

[66] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, volume 32, 2019. 13

[67] Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Zhe Lin, Philip Torr, and Jiaya Jia. Open-world entity segmentation. *arXiv preprint arXiv:2107.14228*, 2021. 14

[68] 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. *ICML*, 2021. 1, 2, 3, 9, 12, 13, 14, 15

[69] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020. 11

[70] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400. PMLR, 2019. 13, 14

[71] Xiaofeng Ren and Jitendra Malik. Learning a classification model for segmentation. In *CVPR*, volume 2, pages 10–10. IEEE Computer Society, 2003. 14

[72] Pieter R Roelfsema et al. Cortical algorithms for perceptual grouping. *Annual review of neuroscience*, 29(1):203–227, 2006. 2

[73] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 1, 4, 13, 14, 15

[74] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 4

[75] Feihong Shen, Jun Liu, and Ping Hu. Conterfactual generative zero-shot semantic segmentation. *arXiv:2106.06360*, 2021. 14

[76] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 22(8):888–905, 2000. 14

[77] Sanjay Subramanian, Will Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022. 1, 3

[78] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015. 2

[79] Mingxing Tan and Quoc V Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019. 11, 13

[80] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, pages 5238–5248, 2022. 1

[81] Guiyu Tian, Shuai Wang, Jie Feng, Li Zhou, and Yadong Mu. Cap2seg: Inferring semantic and spatial context from captions for zero-shot image segmentation. In *ACM MM*, pages 4125–4134, 2020. 14

[82] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *IJCV*, 104(2):154–171, 2013. 14

[83] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 11

[84] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In *NeurIPS*, volume 32, 2019. 11

[85] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-

task benchmark and analysis platform for natural language understanding. In *ICLR*, 2019. 11

[86] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 13

[87] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 15

[88] Xin Wen, Bingchen Zhao, Anlin Zheng, X. Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. *ArXiv*, abs/2205.15288, 2022. 14

[89] Max Wertheimer. Laws of organization in perceptual forms. In W. Ellis, editor, *A Source Book of Gestalt Psychology*, pages 71–88. Routledge and Kegan Paul, London, 1938. 2, 14

[90] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *CVPR*, 2019. 14

[91] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. *CVPR*, 2022. 1, 3, 4, 9, 13, 14

[92] 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. *ArXiv*, abs/2301.09121, 2023. 1, 3, 4, 13, 14

[93] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *ICLR*, 2022. 12, 13

[94] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019. 13

[95] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1, 12

[96] Sukmin Yun, Hankook Lee, Jaehyung Kim, and Jinwoo Shin. Patch-level representation learning for self-supervised vision transformers. In *CVPR*, pages 8354–8363, June 2022. 12

[97] Andrii Zadaianchuk, Matthaeus Kleindessner, Yi Zhu, Francesco Locatello, and Thomas Brox. Unsupervised semantic segmentation with self-supervised object-centric representations. *ArXiv*, abs/2207.05027, 2022. 3

[98] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *ArXiv*, abs/2111.08276, 2021. 13

[99] Shuangfei Zhai, Navdeep Jaitly, Jason Ramapuram, Dan Busbridge, Tatiana Likhomanenko, Joseph Y Cheng, Walter Talbott, Chen Huang, Hanlin Goh, and Joshua M Susskind. Position prediction as an effective pretraining strategy. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26010–26027. PMLR, 17–23 Jul 2022. 3

[100] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *ArXiv*, abs/2206.05836, 2022. 13

[101] Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open vocabulary scene parsing. In *ICCV*, 2017. 14

[102] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through ade20k dataset. *IJCV*, 2018. 13, 16

[103] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, 2021. 1, 3, 13, 14