PERCEPTUAL GROUPING IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

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

Learning a representation for visual imagery requires resolving not only what resides within an image, but also where that information resides (Marr, 1982). In many applications, knowledge of *where* information resides is sometimes more important than a precise description of the content (Geiger et al., 2012; Sun et al., 2020). Hence, our ability to learn more generic and robust visual representations requires learning the geometry of visual semantics, and how visual information may be grounded by specific regions of the visual field.

Vision-language models have demonstrated a remarkable ability to learn generic visual representations that may be readily reused across a large array of visual tasks and domains (Jia et al., 2021; Radford et al., 2021; Yu et al., 2022; Desai & Johnson, 2021). Such models are trained on extremely large corpora of weakly labeled image and text (caption) pairs using contrastive and/or caption based losses, yet the resulting learned representations are quite robust, and may be arbitrarily probed for open-vocabulary (i.e. zero-shot) image recognition problems.

Although vision-language modeling provides a considerable advance towards a generic visual representation (Geirhos et al., 2021), the learned representations demonstrate a profound inability to associate visual content with individual objects (Fig. 1, bottom left). In other words, models trained on large weakly-supervised data have a limited ability to group together visually related content (Ghiasi et al., 2022). 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 (Subramanian et al., 2022; Thrush et al., 2022), and are susceptible to learning spurious correlations between a semantic label and extraneous content (Sagawa et al., 2019; Liu et al., 2021).

In this work, we wish to build vision-language models which learn from weakly labeled data, but have the added benefit of properly learning where visual content resides within an image. Previous attempts have considered elaborations of vision-language architectures that leverage additional training data, specalized architectures or heuristics (Ghiasi et al., 2022; Xu et al., 2022; Yao et al., 2022). We instead focus on building a system that is able to perceptually group regions of visual imagery by identifying a minimal number of changes to existing vision-language models to encourage spatial

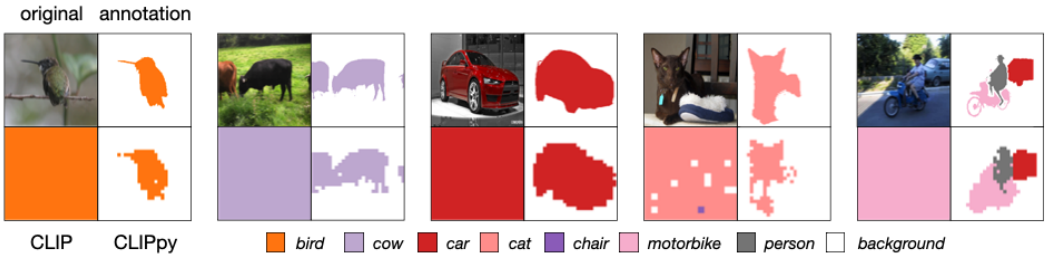


Figure 1: **Semantic localization in vision-language models.**. We measure the ability of vision-language models to predict a label at each spatial position in a zero shot manner based on the similarity to the corresponding language tokens on selected examples. CLIP / ALIGN (Jia et al., 2021; Radford et al., 2021) have minimal understanding of the spatial location of individual objects. Our proposed CLIPpy predicts the label at locations that correspond closely to annotated semantic segmentations (Everingham et al., 2010). All predictions were performed with no access to any segmentation data during training or inference. More examples and complete color legend for labels in App A.

localization. We find that two small adjustments – employing pretrained weights and adjusting the manner of aggregation across space – results in models that are equally effective in zero-shot image recognition, but also retain spatial information about the location of each object (Fig. 1, bottom right).

The resulting model termed CLIPpy exhibits *perceptual grouping* – that is, the ability to select and combine related visual signals into semantically meaningful regions (Wertheimer, 1938; Marr, 1982). Endowing models with perceptual grouping – whether in a bottom up or top down manner – in learned representations has been a long standing goal in computer vision (Malik, 2001; Malik et al., 2016). In this work, our contributions are as follows:

- Identify and characterize the systematic failure of vision-language models to properly identify where objects reside within an image, and group together semantically related content.
- Design a minimal set of changes to endow a model with perceptual grouping. The resulting model achieves state-of-the-art zero-shot segmentation *without* training on *any* segmentation data.
- Emergence of localization ability in our models uniquely leads to robustness to counterfactual manipulations. The degree of robustness matches if not surpasses previous state-of-the-art supervised learning methods employing specialized training methodologies.

2 Related Work

Vision-language models. Vision-language models have advanced considerably in the last decade. Early work learned correspondences between pretrained image and language embeddings in order to perform zero-shot image recognition (Frome et al., 2013; Socher et al., 2013). Subsequent work attempted to learn a model that directly predicted a sentence (i.e. caption) from an associated image (Karpathy & Fei-Fei, 2015; Vinyals et al., 2015; Kiros et al., 2014; Mao et al., 2014) (see also Desai & Johnson (2021)). Recent work has revisited a scaled-up version of learning correspondences between image and language features by using a contrastive loss across the batch (Radford et al., 2021; Jia et al., 2021). See (Pham et al., 2021; Yu et al., 2022) for the latest efforts for scaling up these models.

Vision-language models for grounding. Because of the strength of vision-language models for zero-shot image recognition, several groups have extended this work to better identify the language with parts of an image. Such grounding efforts have focused on employing various heuristics to learn alignments between regions of images and words in a caption (Yao et al., 2022; Cui et al., 2022) Additional work has employed language as a free form way of arbitrarily probing images in an ongoing dialogue (Yuan et al., 2021) and open vocabulary detection (Kamath et al., 2021).

Semantic segmentation. Image segmentation is a core problem in computer vision (Szeliski, 2010), and several prominent benchmarks exist for measuring the success of models on a prescribed set of labels (Everingham et al., 2010; Lin et al., 2014; Zhou et al., 2018). These benchmarks have led to a large corpus of papers focused on developing new architectures treating image segmentation as a dense supervised learning problem (e.g. Chen et al. (2017; 2018)).

Annotating image segmentations for supervised learning is expensive. This expense has motivated learning segmentations from weakly labeled data (e.g. Pinheiro & Collobert (2015)) with the goal of

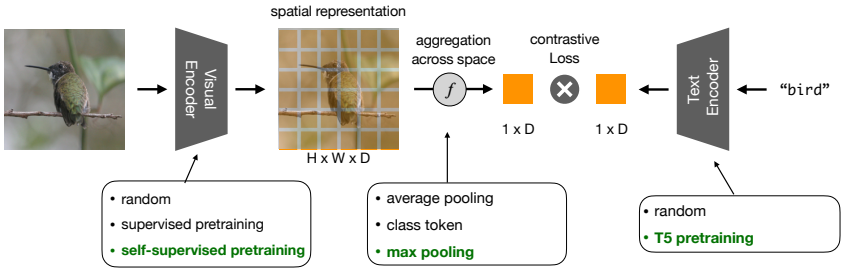


Figure 2: **Diagram of architecture.** Image and caption are separately embedded into a Euclidean, where the image features are spatially aggregated. A contrastive loss trains the global image embedding to be close to the caption embedding. There are several design decisions that we demonstrate are of paramount importance to obtain image models that understand perceptual grouping.

predicting segmentations for objects never previously observed. Several works have extended this to pursue zero-shot segmentation with the goal of open vocabulary segmentation. Ghiasi et al. (2022) leveraged language embeddings to segment images of unobserved visual concepts. Li et al. (2022) employed language to ground image segmentations. Finally, Zabari & Hoshen (2021) employed interpretability on CLIP models to generate pseudo-labels to supervise a segmentation model.

Perceptual grouping for bottom-up recognition. The topic of perceptual grouping has a long, rich history in human visual perception (Wertheimer, 1938) and computer vision (Malik, 2001). The central premise behind perceptual grouping is to learn representations of visual imagery that identify the affinity between visual similar objects. Perceptual grouping has been offered and explored as a method for building systems that are able to generalize to new visual domains (Qi et al., 2021; Malik et al., 2016). It is this latter goal that most closely inspires this work.

Early efforts to achieve these goals led to a series of methods in computer vision to generate groups of pixels based on known spatially-local affinities (Comaniciu & Meer, 1997; Shi & Malik, 2000; Ren & Malik, 2003). Subsequently, modern incarnations have led to region proposal networks for object detection (Uijlings et al., 2013) to advances in semantic segmentation (Arbeláez et al., 2012). Recent methods employ self-supervision to learn such grouping (Cho et al., 2021; Hamilton et al., 2022). Most close to our work, GroupViT provides a custom architecture trained on vision and text pairs learning perceptual grouping by optimizing a discretized attention mask (Xu et al., 2022).

Learning robust visual representations. Assessing learned visual representations has been the subject of intense investigation. Most of the field has started this endeavor with the supposition that ImageNet accuracy provides a reasonable proxy (Girshick et al., 2014; Kornblith et al., 2019). However, recent work has highlighted notable deficiencies in such learned representations (Geirhos et al., 2021; Recht et al., 2019; Koh et al., 2021) including a sensitivity to low level textures, failures in the presence of domain shifts, and a tendency for models to rely on spurious correlations.

These failures inspired a large literature to mitigate learning spurious correlations (Sagawa et al., 2019; Liu et al., 2021; Arjovsky et al., 2019) by focusing on new optimization techniques. Progress on this issue may address parallel issues in fairness (Creager et al., 2021). Resulting methods have largely focused on synthetic data and arrived at algorithms for rebalancing data and shaping learned embeddings (Nam et al., 2020; Liu et al., 2021). Nonetheless, theoretical results suggest pessimistic bounds unless additional structure informs the problem (see refs. in Sagawa et al. (2019)).

3 Methods

We first set the stage by discussing established core architectures and contrastive learning formulation. Next, we discuss modifications that are the focus of the analysis in this work. In particular, we discuss aggregation options and pre-training alternatives.

3.1 Architecture and Training

We provide a quick overview of the vision-language architecture (Fig. 2). Consider a batch size N, spatial height H, spatial width W, and depth D. X is a tensor that has a shape of $[N, H, W, D]$ and is the output of an image encoder. Y is a tensor that is shape $[N, D]$ and is the output of a text encoder.

Language Model. We employ a strong language model baseline derived from the transformer architecture (Vaswani et al., 2017) and implemented in T5 (Raffel et al., 2020). T5 models use an encoder-decoder architecture that is trained using a generative span corruption task, and have achieved state-of-the-art on a broad range of NLP tasks including GLUE (Wang et al., 2019b) and Super-Glue (Wang et al., 2019a). We use the encoder only and discard the decoder part. We employ the T5-base which consists of 12 transformer layers, 12 attention heads, and 768 dimensions.

Image Model. We explore two architectures for image featurization, CNN-based and Vision-Transformers, although we focus the majority of work on the latter. First, we employ the EfficientNet architecture (Tan & Le, 2019) as a high performant CNN architecture, which has been used previously in vision-language models. The specifics of the meta-architecture were derived from considerations based on neural architecture search. Second, we employ the Vision Transformer (ViT) architecture (Dosovitskiy et al., 2020). We refer the reader to (Dosovitskiy et al., 2020; Vaswani et al., 2017) for details. Briefly, ViT is largely inherited from the NLP literature and consists of a hierarchical associative memory. Each layer, termed a Transformer, is composed of a Multi-headed Self-Attention (MSA) layer followed by a 2-layer feed-forward multi-layer perceptron (MLP). The primary parameter of ViT is the patch size P specifying the $P \times P$ patch of pixels constituting a token in the architecture.

Contrastive Representation Learning. Let x_i and y_i denote the image and text embeddings of the i'th example in the batch. A contrastive loss may be specified as the cross entropy across a batch (Radford et al., 2021; Jia et al., 2021). The cross entropy is calculated between a one-hot encoding specifying the correspondence between the image and text examples, and a normalized distribution specifying the similarity between image and text embeddings.

$$L = \underbrace{-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(x_i^\top y_i/\tau)}{\sum_{j=1}^{N}\exp(x_i^\top y_j/\tau)}}_{\text{image-to-text}} + \underbrace{-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(y_i^\top x_i/\tau)}{\sum_{j=1}^{N}\exp(y_i^\top x_j/\tau)}}_{\text{text-to-image}}$$

The normalization for the image-to-text and text-to-image similarity is computed by summing over the potential matches (indexed by j) to the text and image examples within a batch, respectively. Note that τ is the temperature of the softmax for the normalization.

3.2 Aggregation

The goal of the aggregation methods is to collapse the image embedding from a $[H, W, D]$ tensor to a D dimensional vector. **Average pooling** across space is an established technique for ensuring that the final embedding is independent of the image resolution (Szegedy et al., 2015; Long et al., 2015), and has been adopted for CNN-based architectures in vision-language models (Jia et al., 2021). Alternatively, **maximum pooling** has been explored, in particular with success for point clouds (Qi et al., 2017) and image-audio (Harwath et al., 2019). Another approach typical for ViT borrowed from language modeling (Devlin et al., 2018) is **class token (CLS)** that is prepended to the image patch tokens (Dosovitskiy et al., 2020). A class token learns an embedding that aggregates information across all patch tokens in order to predict the image label. The class token may be used to summarize the content for an entire image for ViT-based models (Radford et al., 2021; Caron et al., 2021). Subsequent work in vision-language models has explored learning pooling strategies (Chen et al., 2021; Yao et al., 2022), heuristically selecting a set of similar neighbors (Yun et al., 2022) or learning attention-based mechanisms (Yu et al., 2022).

In this work we systematically explore these aggregation strategies. In early experiments we found that many complex strategies for aggregation yielded poor results (App. G). We found that the simple application of maximum pooling across the spatial dimensions – while extremely simple – was also by far most effective (Sec. 4.5). We hypothesize that the success of maximum pooling may be due to the gradient updates being focused solely on a single spatial location, and not spread across all spatial dimensions.

3.3 Pretraining

Language Model. For better sentence representation, we initialize the **T5 encoder** from the pre-trained Sentence-T5 checkpoints (Ni et al., 2021) which adopts the original T5 models to sentence embedding models using a contrastive loss. The model is first trained on 2 billion question-answers pairs from community question-and-answer websites (Cer et al., 2018), and is subsequently trained again using a contrastive loss on the Stanford Natural Language Inference (SNLI) dataset containing 275K examples focused on entailment questions (Bowman et al., 2015; Gao et al., 2021).

Image Model. We investigate initializing the image model with several methods. First, we investigate initializing the image model using **supervised pre-training** and removing the final layer for logistic regression (Girshick et al., 2014; Kornblith et al., 2019). We next investigated **self-supervised methods** derived from self-distillation (e.g. (Caron et al., 2021)). We focused on this latter direction because such models demonstrated impressive performance in terms of localization.

4 EXPERIMENTS

Experimental Setup. We train vision-language models on two datasets: Conceptual Captions 12M (CC-12M) (Changpinyo et al., 2021) and High Quality Image Text Pairs (HQITP-134M) consisting of 12 million and 134 million image-text pairs, respectively (App. B for details). For both datasets, the text is tokenized, and the image is resized and center cropped to 224×224 pixels. We report results on EfficientNet-B5 employed by ALIGN (Jia et al., 2021), and ViT-B/16 employed by CLIP (Radford et al., 2021) although we focus more on the latter. We train models on 32 GPUs across 4 machines with PyTorch (Paszke et al., 2019). See App. C for details.

We evaluate the model across image classification, localization and robustness tasks. All reported results are based on zero-shot analyses in which the model is prompted at inference time for a selection of potential labels (App. J for prompts). For image classification, we employ the validation splits of ImageNet (Deng et al., 2009), ImageNet-v2 (Recht et al., 2019), and the test split of Waterbirds (Sagawa et al., 2019).

	dataset	IN	IN-v2
ALIGN [a]	ALIGN-1800M	76.4	70.1
CLIP [b]	CLIP-400M	65.5	60.8
ALIGN †	HQITP-134M	51.1	45.6
CLIP †	HQITP-134M	61.4	56.4
CLIPpy	HQITP-134M	60.3	54.8
CLIPpy	CC-12M	45.3	40.0

Table 1: **CLIPpy achieves competitive zero-shot image recognition.** IN and IN-v2 denote ImageNet and ImageNet-v2 accuracy (top-1), respectively. † indicates our implementation. Superscript letter denote the result: [a] Jia et al. (2021), [b] Radford et al. (2021).

For segmentation tasks, we employ zero shot analysis at each spatial location. This is performed by employing the CAM method (Ghiasi et al., 2021; Zhou et al., 2016) to generate a prediction across space by exploiting the transitive property of average pooling and argmax. To measure success, we employ the validation splits of PASCAL VOC (Everingham et al., 2010), ADE20K (Zhou et al., 2018; Cheng et al., 2021) and COCO (Lin et al., 2014; Chen et al., 2015). Each dataset contains 20, 150 and 133 labels, respectively.

Given that most competitive baselines are trained on private datasets, we first attempt to reproduce results by designing and training vision-language models on a corpus of image-text pairs. In more detail, we train on HQITP-134M, and observe competitive performance given our data limitations, as well as on on the public CC-12M dataset to provide comparable numbers. We measure the performance of CLIP and ALIGN on zero-shot image classification on ImageNet and ImageNet-v2. We evaluate all of the proposed architectural changes in Sec. 3, which in aggregate are dubbed CLIPpy, and thoroughly analyse them.

Tab. 1 highlights our results. We take this as a starting point for subsequent work. In the following experiments we attempt to address the following questions:

- What are the limitations of current vision-language models? (Fig. 1)
- Do we observe perceptual grouping in vision language models? (Fig. 3, Tab. 2).
- How resilient are vision-language models to counterfactual manipulations? (Fig. 4).
- How important are each of the proposed model modifications? (Fig. 3).

4.1 LIMITATIONS OF VISION-LANGUAGE MODELS

The learned visual representations in vision-language models exhibit an impressive ability to generalize across tasks (Radford et al., 2021; Jia et al., 2021). However these models also exhibit a profound shortcoming – the 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 left) showcases the failure of a CLIP model; namely, the model improperly conflates visual content not associated with an object with the actual object. This can be observed by measuring the similarity of each embedding at each spatial location with a label set using the CAM method (Sec.

4). One consistently observes that the central object of interest is incorrectly predicted to reside at every spatial location. For instance, in the left example, the CLIP model predicts that a `bird` resides at every spatial location. In a CNN architecture, where spatial information is inherently preserved, we observe some improvement, but the larger issue of poor localization remains (see App. D for details).

The failure of vision-language models to properly understand the spatial organization of information is consistent with earlier observations. Ablation experiments in ViT models demonstrated that removing positional embeddings minimally detriments predictive performance (Dosovitskiy et al., 2020; Naseer et al., 2021; Zhai et al., 2022; Subramanian et al., 2022). Without positional information, ViT models effectively learn representations as a "bag of image patches", ignoring the spatial organization.

In contrast, if we perform the same analysis on CLIPpy, we see that the model retains significant information about spatial information (Fig. 1, bottom right). We take these visualizations as an impetus for further investigation. In particular, we start by quantifying the ability of the model to arbitrarily group together semantically related pixels, and compare this to previous works.

4.2 Emergence of Bottom-Up Perceptual Grouping

Performance on segmentation unsupervised by true object masks at training time is a direct measure of bottom up perceptual grouping. We apply CLIPpy at test time to perform semantic segmentation without prompting it for any labels [1]. Fig. 3b shows that the model visually groups semantically related regions of an image (see also Fig. 5) as the 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 associated text (see App. F for more).

We quantify the accuracy of this bottom-up segmentation to capture known segmentations within annotated images. We follow Caron et al. (2021), Xu et al. (2022), and perform a matching between all candidate annotations with a given segmentation proposed by the model. We compute the Jaccard Similarity (JS) between the inferred segmentations and the associated annotation. The JS measures the average intersection over the union across all segmentation instances regardless of object identity.

On Pascal VOC, CLIPpy achieves a JS of 54.6% outperforming all previous models (Fig. 3a); in comparison, CLIP achieves 38.9%. Additionally, we tested the model on two more challenging datasets and note that the model drops in performance perhaps indicative of more visually cluttered scenes (Fig. 3c). We take the results to indicate that CLIPpy perceptually groups semantically related content better than previous work, and provides state-of-the-art results in unsupervised segmentation.

4.3 Top-down Object Grouping

We demonstrated that CLIPpy is able to perceptually group visual content within an image. Next, we ask how well this perceptual grouping corresponds to semantically meaningful labels. To measure the emergence of top-down processing, we ask how well the perceptual grouping of the model may be steered by embeddings from the language model. We test this hypothesis by comparing the model's ability to perform zero-shot semantic segmentation across three datasets. Note that all of our results and comparisons are solely restricted to models trained on *no* segmentation annotations [2].

Fig. 1 provides a visualization of the predicted zero-shot segmentations (see also App. A), and Tab. 2 quantifies the 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. For example, on HQITP-134M, CLIPpy achieves 51.0% while our baseline CLIP achieves 18.1% IoU. Correspondingly, the official, trained CLIP implementation achieves 16.4%

[1] We follow the procedure outlined by Caron et al. (2021). Namely, we compute PCA across the spatial map of image features. Each principal component corresponds to a candidate feature and we cluster the proximity of each feature vector to these components. GroupViT (Xu et al., 2022) and DINO (Caron et al., 2021) employ 8 and 6 feature vectors based on their model architectures. For our visualizations, we employ 8 feature vectors.

[2] In App. E, we provide a summary of other zero-shot semantic segmentation results. Some of these prior results achieve superior performance, but we note that all of these results were trained explicitly on various forms of segmentation masks, if not the segmentation labels. These models were tested in terms of zero-shot semantic segmentation through a careful split of training and testing labels.

A.

	dataset	train?	VOC
DeiT [a]	ImageNet	class	24.6
MoCo [b]		self	28.2
DINO [a]		self	45.9
DINO [b]	CC-12M YFCC-100M	self	41.8
CLIP [b]		text	28.6
GroupViT [b]		text	51.8
CLIP *	CC-12M	text	37.3
CLIPpy		text	47.5
CLIP *	HQITP-134M	text	38.9
CLIPpy		text	**54.6**

B.

C.

	dataset	ADE20K	COCO
CLIP *	CC-12M	22.9	20.4
CLIPpy		28.9	26.0
CLIP *	HQITP-134M	24.2	21.6
CLIPpy		29.5	27.2

Figure 3: **CLIPpy effectively groups semantically related concepts.** (A and C) All numbers report the Jaccard Similarity, which is an instance average of the IoU between proposed and annotated segmentations independent of the object label. Superscript letter denote the result: [a] Caron et al. (2021), [b] Xu et al. (2022), * denotes our implementation. (B) Visualizations of perceptual grouping. Each color represent one grouping learned by the model on a given image. All models employ a ViT architecture except for GroupViT and operate with a spatial resolution between 448 and 480.

	arch	dataset	ADE20K	COCO	PASCAL VOC
GroupViT [a]	ViT	CC-12M			41.1
CLIPpy	ViT		13.1	23.8	**50.8**
ALIGN †	CNN	HQITP-134M	7.5	14.4	29.7
CLIP †	ViT		5.1	8.0	18.1
CLIPpy	ViT		**13.5**	**25.5**	**52.2**
ALIGN [b]	CNN	ALIGN-1800M	9.7	15.6	
CLIP [c]	ViT	CLIP-400M	5.8	8.7	16.4
GroupViT [a]	ViT	CC-12M, YFCC-100M		24.3	52.3

Table 2: **CLIPpy provides competitive localization 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. Superscript letter denote the result: [a] Xu et al. (2022), [b] Ghiasi et al. (2021), [c] Radford et al. (2021).

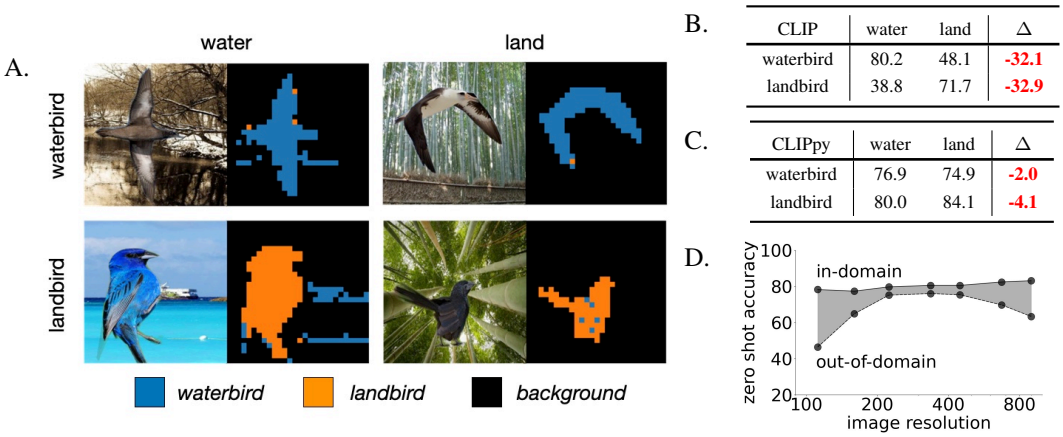
mIoU. We also tested an internal implementation of an ALIGN model based on a CNN backbone and achieved improved results, as may be expected given the strong spatial prior imposed by the model, however notably below previous state-of-the-art (App. D for details).

GroupViT provides an important point of comparison (Xu et al., 2022). This model is a custom ViT architecture trained on vision and text pairs designed to perform perceptual grouping by optimizing a discretized attention mask. GroupViT was trained on a comparable dataset of CC-12M, yet our simple changes outperform this custom architecture by over 10 percentage points (41.1 vs. 50.8). [3] We take these results to mean that our simple changes to existing vision-language models uncover powerful localization information.

4.4 Perceptual grouping may improve robustness

We have observed how parsimonious changes to vision-language models result in state-of-the-art unsupervised and zero-shot semantic segmentation. In this section, we ask how the resulting perceptual grouping may be exploited to improve the robustness of image understanding. A large literature has consistently observed that models systematically underperform under domain shifts (Recht et al., 2019). For instance, CLIP, ALIGN and CLIPpy underperform on ImageNet-v2 versus ImageNet validation (Tab. 1). Another means of assessing robustness is to measure how well a model *causally* predicts the label from the appropriate input variates (Pearl, 2009; Pearl & Mackenzie, 2018). To probe for causal dependencies, one can measure model performance to counterfactual examples where an input is selectively manipulated in order to test for sensitivity to spurious correlations.

[3] We note that even removing all pretraining data and solely training on CC-12M still retains notable performance on unsupervised segmentation (see Tab. 3).



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 4: **Perceptual grouping mitigates sensitivity to spurious correlations.** (A) Selected examples of waterbirds and landbirds on each background, respectively. The right panel shows the argmax of similarity between the three prompts and the image embedding at each spatial location for CLIPpy. (B) Accuracy on the *test* split (5794 examples) of Waterbirds on CLIP and (C) CLIPpy evaluated at 448×448 resolution. The domain gap Δ reports the drop in accuracy between on and off diagonal entries within a row. (D) Zero shot accuracy of CLIPpy across image resolution for landbirds on land (top) and water (bottom). Note log axis. Shading highlights Δ.

A common formulation for this problem is to artificially synthesize a malicious dataset where a trained model may correlate inappropriate image features to predict a label (Xiao et al., 2020; Moayeri et al., 2022; Jacobsen et al., 2018; Arjovsky et al., 2019). A large class of supervised learning algorithms have been developed to train on these datasets [4] with the aim of mitigating such spurious correlations (Sagawa et al., 2019; Liu et al., 2021; Nam et al., 2020). One common synthetic benchmark is *Waterbirds* (Sagawa et al., 2019) which places segmentations of birds in front of a background of land or water. The goal of any prediction system is a two-way classification of whether or not a bird is from the `waterbird` or `landbird` category. What makes this problem particularly challenging is when the background is not commensurate with the type of bird. For instance, a trained model may be prone to predict the type of bird due to the presence of water in the background in lieu of the visual appearance of the actual bird.

We first asked how our baseline CLIP model performs on this task when presented with a zero-shot three-way classification task (App. I for inference procedure). Model performance depends heavily on the background (Fig. 4b). For instance, the prediction accuracy of `waterbirds` drops by Δ=32.1% (80.2 → 48.1) in the presence of an incommensurate background. Clearly, the baseline CLIP model performs zero-shot prediction by relying on features from the background.

We next asked how CLIPpy performs given that it exhibits a unique ability to discriminate the spatial locations of objects. Fig. 4a shows selected examples from each class colored by the prediction at each spatial location. Clearly, the model is able to discriminate which locations correspond to each category. We quantify model accuracy across each task, and find the model far less sensitive to the background. For instance, in the case of waterbirds, CLIPpy accuracy, while slightly less than the baseline CLIP model, only drops by Δ = 2.0% (76.9 → 74.9) in spite of the background change (Fig. 4c). Interestingly, the domain gap Δ is minimal (∼4%) around a broad range of image input resolutions centered about the training resolution of the model (Fig. 4d). Hence, CLIPpy, while still susceptible to some spurious correlations, is far more robust than a standard vision-language model.

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. This difference makes a direct comparison of the raw accuracy difficult. That said, the best supervised training methods achieve a domain gap Δ of 4% to 8% (Tab. 1 and priv. correspondence, Liu et al. (2021)), comparable to our

[4] Synthetic datasets are deliberately constructed to contain a class imbalance such that a minority class may be particularly prone to systematic worse performance. Consequently, experimenters have focused on the worse-case performance on the minority class (Sagawa et al., 2019; Liu et al., 2021). Our work is instead focused on the domain gap to target the degree to which spurious correlations inappropriately influence predictions.

dataset	aggreg.	ImageNet accuracy	Pascal VOC mIoU	Pascal VOC Jaccard
CC-12M	Max	42.3	50.8	47.5
	Avg	44.0	11.6	38.1
	Cls	46.0	4.0	40.4
HQITP -134M	Max	59.0	50.1	54.6
	Avg	60.0	17.9	40.5
	Cls	60.2	4.1	41.3

dataset	image init	T5 init?	ImageNet accuracy	Pascal VOC mIoU	Pascal VOC Jaccard
CC-12M	DINO	✓	42.3	50.8	47.5
	IN-1K	✓	53.3	22.5	43.3
	random	✓	28.9	32.9	43.6
	DINO		34.1	44.3	47.2
	IN-1K		44.5	20.0	42.2
	random		25.6	23.5	43.1

Table 3: **Ablation studies** across zero-shot image classification and segmentation. *Left*: Ablation across aggregation methods including global max pooling (Max), global average pooling (Avg) and using the class token embedding (Cls). All models initialized with the same pretraining features. *Right*: Ablation across pretraining where we initialize the image encoder with DINO, supervised training on ImageNet-1K or random weights. For the text encoder, we initialize with T5 or random weights. Models employ maximum pooling (Max). Parallel ablations using HQITP-134M in App. H.

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.

4.5 Ablations Studies

We next performed a set of experiments to demonstrate how individual factors in CLIPpy led to improved localization performance. We first explored the selection of the aggregation method. Our model employs a maximum operation over all spatial locations. We likewise trained models which performed spatial averaging or employed the class token in ViT. Tab. 3 (left) shows results across two training sets. We see that standard procedures of class token and average pooling result in similar performance on zero-shot classification on ImageNet, but notable reductions in mIoU on Pascal VOC semantic segmentation. For instance, on the model trained with CC-12M, the mIoU on PASCAL VOC dropped from 50.8% to 4.0% representing a relative drop of 91.3%. Similarly, in the case of bottom-up segmentation on the same dataset, we demonstrate a 10 point drop in JS.

We also explored how the selection of the pretraining method effected the overall performance. Tab. 3 (right) explores the selective removal of pretraining on the image model, language model or both. All models employ the maximum pooling aggregation across spatial locations. Again, we see that CLIPpy exhibits significant drops in both zero-shot image recognition and localization by selectively dropping out each pre-training step. For instance, in the model trained on CC-12M, model performance drops from 42.3% to 25.6% top-1 accuracy. Likewise, the semantic segmentation mIoU drops from 50.8% to 23.5% accuracy. For bottom-up segmentation, initializing from pretrained models matters to a lesser degree. We suspect that these results indicate that each initialization provides valuable prior information not readily available in the joint training set for eliciting strong localization properties.

5 Discussion

In this work we demonstrated how vision-language models have a profound lack of understanding of object location. We described a minimal set of changes to existing vision-language models by modifying the aggregation method and the initialized model weights 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 to achieve perceptual grouping (Xu et al., 2022). We demonstrate that our resulting model provides state-of-the-art results in terms of unsupervised segmentation, and achieves competitive results in term of zero-shot semantic segmentation – 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 vision-language models may provide the emergence of perceptual grouping without supervision. We do see limitations in this approach as semantic segmentation suffers with increasing visual clutter and label cardinality (e.g. ADE-20K). We suspect that the recent advent of larger-scale open datasets (Schuhmann et al., 2021; Byeon et al., 2022) and new methods in self-supervised learning (Hamilton et al., 2022) may offer opportunities to demonstrate further benefits for endowing models with perceptual grouping.

REFERENCES

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

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

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.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

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

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.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *CVPR*, pp. 9650–9660, 2021.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

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*, pp. 3558–3568, 2021.

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

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.

Liang-Chieh Chen, Maxwell Collins, Yukun Zhu, George Papandreou, Barret Zoph, Florian Schroff, Hartwig Adam, and Jonathon Shlens. Searching for efficient multi-scale architectures for dense image prediction. *NeurIPS*, 31, 2018.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

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

Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. *CVPR*, pp. 16789–16799, 2021.

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

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

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.

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

Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *CVPR*, pp. 11162–11173, 2021.

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.

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.

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

Andrea Frome, Greg S Corrado, Jonathon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *NeurIPS*, 26, 2013.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. *ICCV*, pp. 10032–10042, 2021.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pp. 3354–3361. IEEE, 2012.

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.

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.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. In *ECCV*, 2022.

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

Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *ICLR*, 2022.

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.

Jörn-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. Excessive invariance causes adversarial vulnerability. *arXiv preprint arXiv:1811.00401*, 2018.

Xu Ji, Andrea Vedaldi, and João F. Henriques. Invariant information clustering for unsupervised image classification and segmentation. *ICCV*, pp. 9864–9873, 2019.

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.

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*, pp. 1780–1790, 2021.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pp. 3128–3137, 2015.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

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*, pp. 5637–5664. PMLR, 2021.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *CVPR*, pp. 2661–2671, 2019.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, 2011.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *ICLR*, 2022.

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.

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

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*, pp. 6781–6792. PMLR, 2021.

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

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

Jitendra Malik. Visual grouping and object recognition. In *Proceedings 11th International Conference on Image Analysis and Processing*, pp. 612–621. IEEE, 2001.

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.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.

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

Mazda Moayeri, Phillip Pope, Yogesh Balaji, and Soheil Feizi. A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes. In *CVPR*, pp. 19087–19097, 2022.

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.

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

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.

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.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V Le. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*, 2021.

Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

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.

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.

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.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pp. 5389–5400. PMLR, 2019.

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

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.

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.

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

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. *NeurIPS*, 26, 2013.

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.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, pp. 2446–2454, 2020.

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*, pp. 1–9, 2015.

Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

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

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*, pp. 5238–5248, 2022.

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.

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.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pp. 3156–3164, 2015.

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, 2019a.

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*, 2019b. URL https://openreview.net/forum?id=rJ4km2R5t7.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pp. 7794–7803, 2018.

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.

Max Wertheimer. Laws of organization in perceptual forms. In W. Ellis (ed.), *A Source Book of Gestalt Psychology*, pp. 71–88. Routledge and Kegan Paul, London, 1938.

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

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

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.

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.

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.

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.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

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

Nir Zabari and Yedid Hoshen. Semantic segmentation in-the-wild without seeing any segmentation examples. *arXiv preprint arXiv:2112.03185*, 2021.

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*, pp. 26010–26027. PMLR, 17–23 Jul 2022.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pp. 2921–2929, 2016.

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