A Semantic Space is Worth 256 Language Descriptions: Make Stronger Segmentation Models with Descriptive Properties

Junfei Xiao[1] Ziqi Zhou [2] Wenxuan Li[1] Shiyi Lan[3] Jieru Mei[1]
Zhiding Yu[3] Bingchen Zhao[4] Alan Yuille[1] Yuyin Zhou[2] Cihang Xie[2]

[1]Johns Hopkins University [2]UCSC [3]NVIDIA [4]University of Edinburgh

Abstract. We introduce **ProLab**, a novel approach using **pro**perty-level **lab**el space for creating strong interpretable segmentation models. Instead of relying solely on category-specific annotations, ProLab uses descriptive properties grounded in common sense knowledge for supervising segmentation models. It is based on two core designs. First, we employ Large Language Models (LLMs) and carefully crafted prompts to generate descriptions of all involved categories that carry meaningful common sense knowledge and follow a structured format. Second, we introduce a description embedding model preserving semantic correlation across descriptions and then cluster them into a set of descriptive properties (*e.g.*, 256) using K-Means. These properties are based on interpretable common sense knowledge consistent with theories of human recognition. We empirically show that our approach makes segmentation models perform stronger on five classic benchmarks (*e.g.*, ADE20K, COCO-Stuff, Pascal Context, Cityscapes and BDD). Our method also shows better scalability with extended training steps than category-level supervision. Our interpretable segmentation framework also emerges with the generalization ability to segment out-of-domain or unknown categories using in-domain descriptive properties. Code is available at https://github.com/lambert-x/ProLab.

1 Introduction

Semantic segmentation is widely used in many real-world applications such as autonomous driving [13,23,82], scene understanding [20,43,51,87], and medical image analysis [27, 42, 60]. Seminal works in this domain include [10, 48, 59, 74, 75], all of which have significantly advanced the field with their innovative architectures and strategies.

Despite their advanced design, models like DeepLab [10], UperNet [74], SegFormer [75], and Vision Perceiver [12] use a one-hot label space for categories, lacking inter-category semantic correlations. Attempts to address this, such as manual category merging [37] or modeling hierarchical label relationships [40], often result in performance drops and scalability challenges, exacerbated by expanding data and semantic spaces. On the other hand, recent works [38,89] have

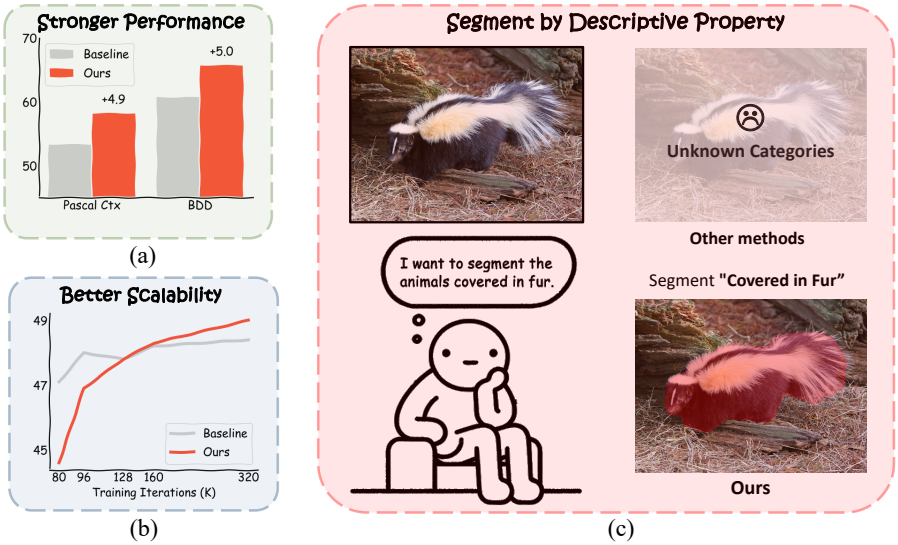


Fig. 1: Advantages of ProLab. Compared to classic category-level label space, ProLab improves segmentation models in three aspects: (a) **stronger performance** on classic segmentation benchmarks (b) **better scalability** with extended training steps (c) ability to **segment by descriptive properties**, which could generalize to out-of-domain categories or even unknown categories.

addressed label space issues by leveraging language embeddings from CLIP [56] for constructing label spaces. However, methods that use CLIP to model inter-class embeddings often struggle with human interpretability. This is primarily because CLIP, despite its capabilities, lacks an extensive common sense knowledge base. Additionally, CLIP encounters challenges due to its reliance on image-text paired data, which inherently suffers from the long-tail distribution issue, limiting the model's ability to less common or more nuanced scenarios.

To address these challenges, we developed **ProLab**, an innovative approach that creates a **pro**perty-level **lab**el space for semantic segmentation. This label space, derived from the rich common sense knowledge base of Large Language Models (LLMs), is filled with descriptive properties. It aligns segmentation models to a nuanced and human-interpretable common sense semantic space, leading to stronger performance and better generalization ability, shown in Fig. 1(a) and Fig. 1(b). Moreover, ProLab models interpretable semantic correlations between categories and is scalable and adaptable to expanding data volumes.

ProLab enables models to recognize objects based on a set of interpretable properties. For instance, our model can generate pixel-wise logits for properties like "paws have claws and pads for walking" or "round eyes." For general segmentation benchmarks evaluation, our method compares these property-level logits to the original label space which can be somehow in line with the human-reasoning process [14,35]. It might, for example, assign the category "dog" based

on activated properties like “paws have claws and pads for walking" and “seen in parks", showcasing its ability to align property-based recognition with conventional category-level identification.
Our contribution can be summarized as follows:

- We propose a novel method, **ProLab**, building an interpretable label space for semantic segmentation by retrieving common sense knowledge from pure language models instead of by using CLIP text encoders.
- **ProLab** consistently shows stronger performance than classic category-level supervision on five benchmarks: ADE20K [87], COCO-Stuff [43], Pascal Context [43], Cityscapes [13], and BDD [82].
- **ProLab** shows better scalability with extended training steps without having performance saturation.
- **ProLab** qualitatively exhibits strong generalization capabilities to segment out-of-domain categories with in-domain descriptive properties.

2 Related Work

2.1 Open-vocabulary recognition

Open-vocabulary recognition aims to address visual recognition problems in an open world by extending the semantic space to unlimited vocabularies. To address this problem, a universal trend is to leverage pre-trained vision-language models such as CLIP [56], where the language modules are trained to be visually aligned such that they can be mapped to open vocabularies.

Recent works such as [26] address open-vocabulary object detection and subsequent works extend the problem to various segmentation tasks with more or less similar approaches [8, 11, 25, 38, 39, 77, 83, 85, 91]. A critical difference between prior works and this paper is that our method focuses on the construction of semantic space using LLM knowledge instead of vision-language pre-training.

2.2 Language-supervised image segmentation

Besides open-vocabulary recognition, recent works also consider language-supervised dense prediction without using mask annotations [52, 76, 76, 88]. While enjoying similar open-vocabulary recognition capabilities using vision-language models, these models further explore the emerging dense prediction properties from language supervision. It should be mentioned that these works are inherently related to earlier works on the emerging localization from network activation [86] and weakly-supervised learning [5, 19].

2.3 Referring expression grounding

Another important language model application for vision tasks is referring expression grounding. Referring expression grounding aims to locate a target region in an image according to the referring language expression. There exists

a rich literature that focuses on the design of vision-language interaction modules [28,29,32,79,80] with common grounding benchmarks such as RefCOCO [33] and PhraseCut [69]. More recent works feature the use of well pre-trained vision-language models [41,44] and LLMs [3,9,36,55,66,71].

Our work is partly related to referring expression grounding in the sense that segmentation with interpretable properties can be broadly viewed as an inverse process going from regions to language expression. With moderate changes, it is possible to also retrieve these regions with free-form language similar to referring expression grounding.

2.4 Semantic space construction

The use of language knowledge for semantic space construction is possibly the most relevant area to this work. This area is rooted from many seminal early works in visual recognition, including but are not limited to the following:

Hierarchy and graph. Semantic hierarchy [2,15,22,40,49,92] and relational/-knowledge graph [16,67,90] has a long history [63] for large-scale visual recognition [17]. In these cases, one could formulate the recognition task as a structured prediction task, or leverage the hierarchy and graph as structured priors besides likelihood. Hierarchy and relational/knowledge graph are essentially special cases of the LLM encoded knowledge.

Attribute learning. Learning with attributes [1,21,61,62,81] is a task in which one focuses on learning the properties or "visual adjectives" beyond just taxonomy. A direct application of attribute learning is zero-shot recognition [21,53,81] and our method can be broadly viewed as a form of attribute learning.

Multi-dataset training. Constructing aligned label space is an important step for multi-dataset training since different dataset may use different vocabularies to represent the same visual concept [34, 37, 72, 89]. Lambert et al. [37] show that aligning the label space is key to the domain generalization capability of trained models. Kim et al. [34] proposes a label unification approach to systematically resolve the label definition conflicts, while recent works also resort to language models to automate this process [89]. Our method can be used for similar purposes using LLM.

3 Method

Conventionally, a semantic segmentation model f process an RGB image $x \in \mathcal{R}^{3\times H\times W}$ as input, generating pixel-wise predictions $p = f(x) \in \mathcal{R}^{N\times H\times W}$, where N signifies the number of categories in line with the label space $\{C_1, ...C_N\}$ of the designated training dataset(s). This is represented as:

$$y \in \mathbb{1}_{C_i}(x) := \begin{cases} 1 & \text{if } x \in C_i, \\ 0 & \text{if } x \notin C_i. \end{cases}, \quad i \in \{1, 2, 3, \ldots, N\}$$

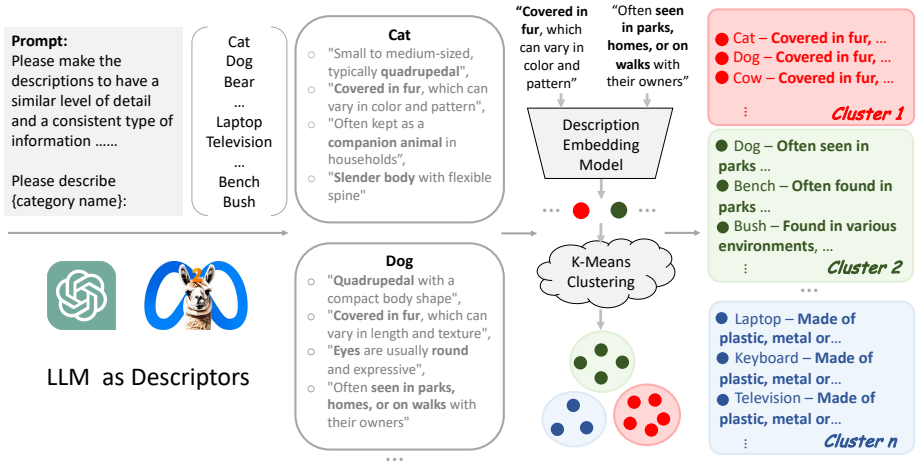


Fig. 2: Build a semantic space of descriptive properties. ProLab firstly employs a Large Language Model to extract common sense knowledge pertinent to all involved categories, utilizing crafted prompts to ensure a structured format. Subsequently, a description embedding model is used to encode these descriptions, preserving semantic correlations. Finally, the description embeddings are grouped into a series of unique descriptive properties through K-Means clustering.

However, this traditional one-hot label space fails to capture inter-class correlations, resulting in models lacking out-of-domain generalization ability. Our approach, in contrast, employs LLMs (e.g., GPT-3.5) to transform this one-hot category-level label space into a multi-hot property-level label space for supervision. Initially, LLMs function as descriptors to provide a set of descriptions regarding the properties of each distinct category (as detailed in §3.1). These descriptions are encoded into embeddings by a sentence embedding model and subsequently clustered into a series of interpretable properties $\{P_1, P_2, P_3 ... P_M\}$ (as detailed in §3.2). This is represented as:

$$y \in \mathbb{1}_{P_j}(x) := \begin{cases} 1 & \text{if } x \in C_i, C_i \in P_j \\ 0 & \text{if } x \notin C_i \end{cases}, \quad \begin{matrix} i \in \{1, 2, 3, \ldots, N\}, \\ j \in \{1, 2, 3, \ldots, M\} \end{matrix}$$

3.1 Property Knowledge Retrieval from LLM

Large language models. Our approach leverages LLMs as descriptive tools, illusrated in Figure 2. They offer a spectrum of detailed descriptions related to the characteristics of various categories within a label space. These models, particularly state-of-the-art ones like GPT-3.5, are adept at generating rich, meaningful descriptions at the property level. These descriptions not only resonate with human understanding but also serve as benchmarks for distinguishing between different categories. The capacity of Large Language Models (LLMs) to express



Please make the descriptions to **have a similar level of detail and a consistent type of information**, which should be **beneficial for clustering and machine learning** applications. Each description will follow a structured format:
- *Start with a **general description** of the object or scene.*
- *Describe the **shape, orientation, and primary physical characteristics**.*
- *Mention the **material, texture, or typical colors**.*
- *Note **common features or elements** associated with the object or scene.*
- *End with possible **additional details, variations, or environmental context**.*

Please describe {**category name**}:

Fig. 3: Prompt for descriptions. This crafted-prompt ensures precise guidance for Large Language Models in retrieving consistent and reliable property descriptions.

common sense knowledge about shape, texture, and other general properties, akin to human recognition processes, renders them a vital element in our approach. This ability significantly augments the interpretability and practicality of label spaces in our method.

Prompt. Crafting an effective prompt is essential for getting high-quality, detailed descriptions from Large Language Models (LLMs), which are instrumental in reconstructing the semantic space with rich property-level information. As illustrated in Figure 3, we deploy precise instructions to guide the LLM in characterizing categories with uniform detail and consistent information types.

This approach clusters properties into groups with meaningful semantics. The LLM provides descriptions across a broad spectrum of attributes, including shape, orientation, and primary physical features like material, texture, and characteristic colors. It also covers common elements associated with the object or scene. This structured prompting generates comprehensive descriptions that enhance the quality and usability of the semantic space.

3.2 Build Semantic Space with Descriptions

Description embedding model. Our approach employs sentence embedding models, such as Sentence Transformers [58] or BGE-Sentence [73], to transform descriptions into a representational space suitable for supervision. These models, purely based on language, are pre-trained exclusively on textual data. Their training involves the use of contrastive loss, which effectively constructs an embedding space capable of modeling semantic similarities between sentences. These similarities are quantified using cosine similarity scores, allowing the embedding space to accurately preserve semantic correlations among various property-level descriptions. This methodology ensures that the essence and nuanced differences of each description are effectively captured and represented in the model, facilitating a more robust and semantically rich supervision process.

Cluster description embeddings. Given that many properties are shared across multiple categories, it becomes necessary to group descriptions into clus-

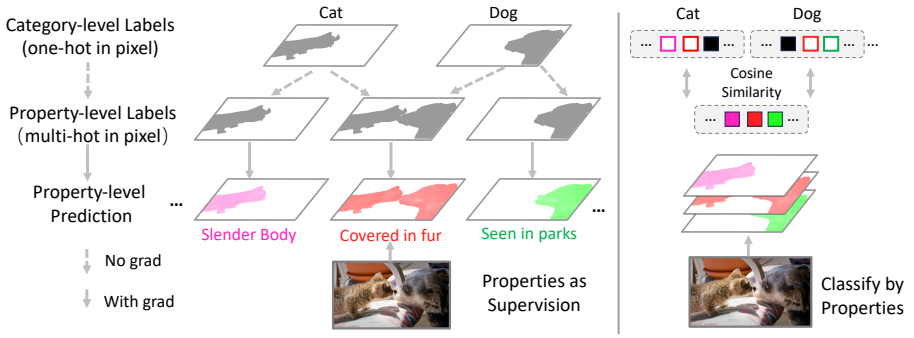


Fig. 4: Supervise and classify with properties. **Left**: the training procedure where descriptive properties are used for model supervision. **Right**: the inference procedure of categorizing items within the original category-level label space.

ters when they refer to identical or similar properties. To achieve this, we employ K-Means clustering [47] to cluster the embeddings generated by the description embedding model into a set of generalized properties. This clustering is performed while maintaining a controlled semantic distance between the properties, ensuring that they remain interpretable and relevant to humans. By aggregating similar descriptions, we streamline the semantic space, making it more practical and user-friendly for both computational processes and human understanding.

3.3 Supervise and Classify with Properties

Properties as supervision. In contrast to the category-level label space that a pixel is only labeled with one category, this property-level label space is a multi-label classification problem. As shown in Figure 4, the pixels which are originally labeled as “cat", now correspond to all the clustered properties of the “cat" descriptions, including “slender body" and “covered in fur". And for the pixels labeled with “dog", they are now labeled with all the dog’s properties including “covered with fur" and "seen or found in various environments". For each pixel i, our model predicts an embedding $\mathbf{e}_i$ at the pixel. Let $\mathbf{E} \in \mathbb{R}^{d\times k}$ be the property embedding bank. Then, the property-level logits for pixel i can be calculated: $\mathbf{z}_i = \sigma(\mathbf{e}_i\mathbf{E})$, where $\sigma(\cdot)$ denotes the sigmoid function. Let $\mathbf{y}_j$ be the multi-hot label vector for the class j of the label space. The cosine similarity is calculated as: $\text{sim}(\mathbf{y}_j, \mathbf{z}_i) = \frac{\mathbf{y}_j \cdot \mathbf{z}_i}{\|\mathbf{y}_j\|\|\mathbf{z}_i\|}$. With this multi-hot property-level label space, our model has stronger performance, better scalability, and emerges generalization ability to out-of-domain and even unknown categories.

Classify by properties. Our model is still adept at categorizing pixels into their original categories by leveraging property-level logits, as depicted in Figure 4. Specifically, this is achieved by assigning each pixel to the category whose property-level label exhibits the highest cosine similarity with that pixel:

$c_i = \arg\max_j (\text{sim}(\mathbf{y}_j, \mathbf{z}_i))$. This method effectively harnesses the nuanced information contained within the property-level descriptions, ensuring that pixel categorization aligns accurately with the most relevant semantic characteristics.

4 Experiments

4.1 Implementation Details

Segmenatation model. We use ViT-Adapter [12] with UperNet [74] as the segmentation framework which is a state-of-the-art method for the semantic segmentation task. Unless otherwise specified, ViT-Base serves as the standard vision backbone across all experiments. The output feature dimension is determined by the dimension of language embeddings, which is set to 384 or 768, which matches the dimension of description embeddings.

Large Language Models. In our experiments, we employ GPT-3.5 [6] and LLAMA2-7B [65] as our primary large language models. By default, we utilize GPT-3.5 for descriptive properties retrieval. To ensure the extraction of descriptive properties at a consistent level for different categories, we craft a prompt to accurately guide the LLMs in retrieving relevant and uniform property description, as detailed in Fig. 3.

Embedding model. For the language embedding model, we choose Sentence Transformers [58] and BGE-Sentence [73]. The output feature dimension is set to 384 or 768. All pretrained weights are imported from HuggineFace [68].

Training details. The ViT backbone in our model is initialized with the DeiT-Base weights [64], which have been pre-trained on ImageNet-1K [17]. For ADE, COCO-Stuff, and Pascal Context, the input resolution is set to 512x512, while for Cityscapes and BDD, a resolution of 768x768 is used. The optimizer of choice is AdamW, with LR configured at 6e-5. Our models are trained on 8-GPU machines, with a total batch size of 16. For supervising the property-level labels, cosine similarity loss is utilized. A one-hot training stage is adopted for warm-up when training large models. Further details are provided in the appendix.

Evaluation details. The models are evaluated using the classic single-scale test setting. The primary metric is the Mean Intersection over Union (mIoU). For classic evaluation, a dictionary of multi-hot property-level labels is employed to calculate cosine similarity scores with the property-level logits. The category with the highest cosine similarity score is determined as the predicted label. Further details are provided in the paper's appendix.

4.2 Segmentation with Interpretable Properties

Our method utilizes property-level supervision to generate activation maps reflecting distinct descriptive properties, which are then correlated with traditional category-level labels based on property-level logits similarity. Figure 5 displays the model's predictions related to different properties in two example images.

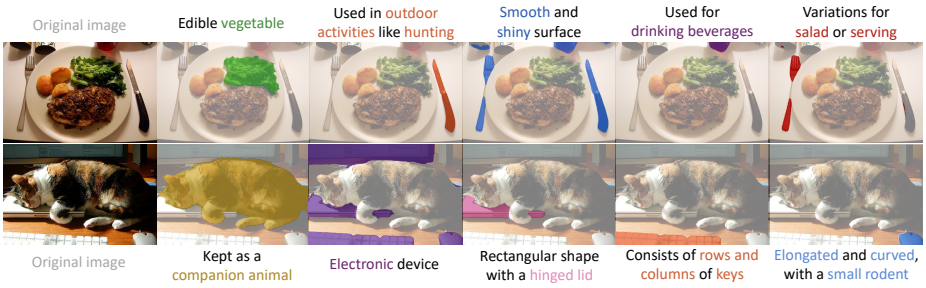


Fig. 5: Segmentation with interpretable properties. Our ProLab model enables property-level segmentation using descriptive prompts, enhancing interpretability and mirroring human-like understanding.

Each row shows activated areas corresponding to three distinct interpretable properties. In 1st row, "Broccoli" is identified through the "edible vegetable" property. However, "knife"and "fork" require a nuanced approach for categorization. They are grouped under the "smooth and shiny surface" property. The "knife" can be distinguished by its association with "activities like hunting." Consequently, our model segments areas that align with both properties as "knife", while employing additional properties to isolate the region representing "fork".

The second row demonstrates the model's ability to segment by diverse interpretable properties like "rectangular shape with a hinged lid," "companion animal," and "electronic device." Here, items such as keyboards, mice, laptops, and monitors are segmented under the "electronic device" property, which aligns with human cognition. This approach underlines the effectiveness of property-level segmentation for more accurate and understandable categorization, mirroring human-like decision-making in identifying objects.

4.3 Stronger Performance & Better Scalability

Stronger performance. To comprehensively evaluate our method, We conduct extensive experiments on five classic semantic segmentation datasets: three natural scene datasets (ADE20K [87], COCO-Stuff [43], Pascal Context [43]), and two self-driving datasets (Cityscapes [13], BDD [82]). We utilized ViT-Adapter, a state-of-the-art segmentation framework, as our baseline to evaluate the efficacy of our property-level label space. As detailed in Tab. 1, our method consistently outperformed the baseline across all five datasets, often by a substantial margin. Notably, it achieved significant enhancements in the Pascal Context and BDD datasets, with an increase to 58.2 (+4.9) mIoU and 65.7 (+5.0) mIoU.

We also make improvements on remaining datasets (ADE20K, COCO-Stuff, and Cityscapes). These results validate the effectiveness of our approach, demonstrating that our label space, constructed with descriptive properties, builds a better representation space. This space adeptly encodes semantic correlations across different categories, proving its superiority over traditional methods.

			Natural Scene			Self Driving	
Label Space	Backbone	Framework	ADE20K	COCO-S.	Pascal Ctx.	Cityscapes	BDD
Categories	DeiT-B	ViT-Adapter	48.4	43.1	53.3	79.9	60.7
Properties	DeiT-B	ViT-Adapter	**49.0**(+0.6)	**45.4**(+2.3)	**58.2**(+4.9)	**81.4**(+1.5)	**65.7**(+5.0)

Table 1: Categories v.s. Properties. We adopt the state-of-the-art segmentation method ViT-Adapter [12, 74] as our baseline.Our ProLab consistently shows stronger performance on three natural scene datasets and two self-driving datasets.

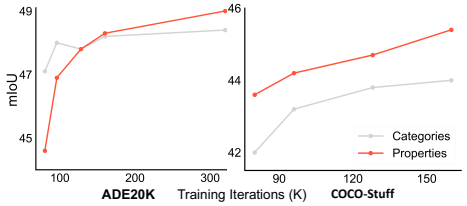


Fig. 6: Better scalability. The demonstrated results validate that our property-level semantic space effectively reduces the tendency of models to overfit, even with extended training durations.

Method	Backbone	ADE20K	Citysc.	BDD
DeepLabv3+	ResNet50	42.7	79.5	63.6
w/ ProLab		**43.6**	**79.9**	**64.7**
Segformer	MIT-B1	41.5	78.8	61.7
w/ ProLab		**42.3**	**79.9**	**62.1**

Table 2: Generalizable to other methods. ProLab demonstrates generalizable effectiveness by making consistent improvements segmentation frameworks (*i.e.*, DeepLabv3+ [10], SegFormer [75]).

Better scalability. Our proposed property-label based method demonstrates significant potential for scalability with extended training steps. The evidence supporting this is illustrated in Fig. 6, where we observe performance improvements in our model (depicted by the red lines) that surpass those of the baseline model (represented by the gray lines) when the models trained. This trend is evident on both the ADE20K and COCO-Stuff datasets.

4.4 Generalizablity to Other Segmentation Models

To evaluate the generalizability of our property-level label space, we arm two other classic segmentation methods (*i.e.*, DeepLabv3+ [10] and Segformer [75]) with ProLab. As detailed in Table 2, ProLab consistently improves the performance with both DeepLabv3+ and Segformer on three classic datasets (*i.e.*, ADE, Cityscapes and BDD), demonstrating strong generalizability with ProLab.

4.5 Comparison with State-of-the-Art Methods

To validate the versatility of our approach across different backbone architectures, especially those with advanced pretraining, we evaluated our method using larger backbones pretrained with state-of-the-art methods [4, 54]. As indicated in Table 3, our approach enhances the performance of the ViT-Adapter-L model to achieve a new state-of-the-art performance of 58.2 mIoU on the ADE20K validation set using BeiT-Large, which further increases to 58.7 mIoU when using a higher input resolution of 896 with BeiTv2. It's worth noting that while

Method	Framework	Backbone Pre-train	Crop Size	ADE20K
Swin-L [46]	Mask2Former	IN-22K, sup	640	56.1
Swin-L-FaPN [30]	Mask2Former	IN-22K, sup	640	56.4
SeMask-Swin-L [31]	Mask2Former	IN-22K, sup	640	57.0
HorNet-L [57]	Mask2Former	IN-22K, sup	640	57.5
ViT-Adapter-L [12]	Mask2Former	IN-22K, sup	640	56.8
BEiT-L [4]	UperNet	IN-22K, BEiT	640	56.7
ViT-Adapter-L [12]	UperNet	IN-22K, BEiT	640	58.0
BEiTv2-L [54]	UperNet	IN-22K, BEiTv2	512	57.5
ViT-Adapter-L [12]	UperNet	IN-22K, BEiTv2	512	58.0
ProLab (Ours)	UperNet	IN-22K, BEiT	640	58.2
ProLab (Ours)	UperNet	IN-22K, BEiTv2	896	**58.7**
SwinV2-G [45]	UperNet	IN-22K, sup	896	59.3

Table 3: Comparison with state-of-the-art methods on the ADE20K *val* set. ProLab achieves state-of-the-art performance. We report **mIoU** with single-scale inference. SwinV2-G [45] is grayed as its backbone is in a much larger scale (*i.e.*, 3B).

SwinV2-G [45], implemented in the UperNet framework, achieves slightly higher performance, its model size is approximately five times larger than ours.

4.6 Ablation Studies

Description embedding model. In our study, we explore two widely used language embedding models—Sentence TR [58] and BGE [84]—to encode descriptive properties into a language semantic representation space. As detailed in Table 4a, we experiment with two variants of these models, featuring embedding dimensions of 384 and 768. Our findings indicate a clear preference for BGE language embedding models, which seem to more effectively capture the essence of the sentences. Notably, larger embedding models outperform their smaller counterparts for both BGEs and Sentence TRs. Based on these results, we choose BGE as our language embedding model for subsequent experiments, aiming to leverage its enhanced performance capabilities.

Number of clusters. We also ablate the number of clusters for clustering embedding descriptions in Tab. 4b. Clustering has shown to be a critical component in our methodology, as evidenced by the markedly poor performance in the first row (without clustering). Among different cluster numbers, models with 64 and 512 clusters show slight performance drops, hovering around 47.8 mIoU and 47.6 mIoU. The model using 128 clusters performs well, with only a small gap compared to the best cluster number of 256 (48.0 *vs*. 48.3). These suggest that ProLab is not highly sensitive to cluster numbers, but they should be within a reasonable range. The optimal number of clusters can vary for different datasets. 1/6 to 1/8 of the number of descriptions usually works well based on our experiments in Tab. 1.

Prompts with text encoders. We conduct a comprehensive ablation study of the embedding space and text encoders. As outlined in Tab. 4c, we experiment

Model.	Embed. Len.	mIoU
Sent. TR-Small	384	47.7
Sent. TR-Base	768	47.8
BGE-Small	384	47.9
BGE-Base	768	**48.3**

(a) Description embedding model.

# Cluster	mIoU
None	30.2
64	47.8
128	48.0
256	**48.3**
512	47.6

(b) Description cluster number.

Prompts	Text Encoder	Embed. Space	mIoU
CLIP Official	CLIP-B	Category	48.6
Ours Descr.	CLIP-B	Category	48.6
Ours Descr.	T5-B	Category	42.0
Ours Descr.	BERT-B	Category	40.8
Ours Descr.	BGE-B	Category	47.7
Ours Descr.	BGE-B	Property	**49.0**

(c) Prompts with text encoders.

Logit Temp.	mIoU
0.02	47.4
0.04	**47.7**
0.07	47.2

(d) Description logit temperature.

Loss Function	mIoU
Binary Cross Entropy	47.4
Cosine Simi. (w/o Sigmoid)	47.3
Cosine Simi. (w/. Sigmoid)	**47.7**

(e) Loss function for multi-hot labels.

LLM	Desc. Format	mIoU
LLAMA2-7B	Naive [50]	47.0
GPT-3.5	Naive [50]	47.5
GPT-3.5	Aligned	**47.7**

(f) Description format.

Table 4: Ablation studies on the ADE20k dataset, mIoU scores are reported. Best performance settings are marked in gray.

with different kinds of language models: the CLIP text encoder, classic LMs (such as BERT and T5), and sentence embedding model (BGE). The first two rows of our experiment aim to determine if our prompt results in improvements when used with CLIP text encoders. The 3rd and 4th rows reveal that pure language models like BERT and T5 are less effective at encoding descriptions into embeddings compared to the CLIP text encoder and also underperform relative to sentence embedding models. This is possibly because general-purpose language models like BERT and T5 struggle to capture semantic relationships between sentences due to the absence of contrastive training. In contrast, the BGE sentence embedding model demonstrates reasonable performance when encoding descriptions into a category-level embedding space and achieves the best result of 49.0 when the embedding space is constructed at the property level.

Description logit temperature. We explore the impact of sigmoid temperature on model performance, considering its role in rescaling the cosine similarity scores in property-level (ranging from -1 to 1) and smoothing gradients for better convergence. In our experiments detailed in Tab. 4d, we test three different temperatures: 0.02, 0.04, and 0.07. The results indicate that a temperature setting of 0.04 leads to the highest performance, achieving 47.6 mIoU. However, the temperatures of 0.02 and 0.07 also yield strong results (47.4 mIoU and 47.2 mIoU, respectively), showing only slight variations in performance. Based on these results, we use a temperature of 0.04 by default.

Loss function for multi-hot labels. We explore both binary cross-entropy loss and cosine similarity loss for our multi-hot property-level labels. Interestingly, utilizing the cosine similarity loss allows the model to achieve comparable

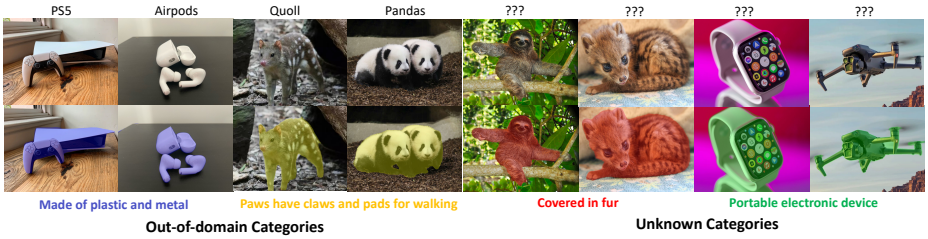


Fig. 7: Emerged generalization ability. Our model demonstrates the ability to segment categories that are outside its training domain, including those it has never encountered, as well as unknown categories, using significant descriptive properties. These properties draw on human-interpretable, common sense knowledge, showcasing the model's adaptability and depth of understanding.

performance to binary cross-entropy loss (47.3 vs. 47.4). Moreover, when we apply the sigmoid function to the output logits, the cosine similarity loss outperforms binary cross-entropy loss (47.6 vs. 47.4). This improvement is attributed to multi-hot labels being binary (0 or 1), making sigmoid-rescaled logits more suitable for model optimization than directly output logits.

Description format. We evaluate the effectiveness of different description prompts with various Large Language Models (LLMs). Our baseline prompt, inspired by [50], poses a question to the LLM: *"What are useful features for distinguishing a category name in a photo?"* As detailed in Tab. 4f, we experiment with GPT-3.5 and LLAMA2-7B for generating description prompts. The first two rows show that GPT-3.5 delivers common sense knowledge of higher quality compared to LLAMA2-7B. Using our specifically crafted prompt (termed as "Aligned") led to a further increase of 0.2 mIoU compared to the models trained with "Naive" prompts. These findings indicate that our tailored prompts are more effective at extracting property-level common sense knowledge.

5 Discussion

Emerged Generalization Ability

As highlighted in Section 4.2, our approach enables segmentation based on descriptive properties, such as shape, texture, and other common-sense knowledge. This capability leads to an emergent generalization ability, allowing our model to segment objects even if their categories are not included in the training set, by identifying their properties. Figure 7 presents four examples: "PS5", "Airpods", "Quoll", and "Pandas". Despite these categories not being part of the training dataset, our model successfully segments them based on properties like "made of plastic and metal" and "paws with claws and pads for walking".

Additionally, in instances where category names are unknown, while humans can still segment objects using common sense properties, most deep learning models fail without specific category information. In contrast, our method

Method	PT. Data (Scale)	Training Data	A-847	A-150	PC-59	PAS-20
SPNet [70]	ImageNet (1.28M)	Pascal VOC	-	-	24.3	18.3
ZS3Net [7]	ImageNet (1.28M)	Pascal VOC	-	-	19.4	38.3
LSeg [38]	ImageNet (1.28M)	Pascal VOC	-	-	-	47.4
ZegFormer [18]	CLIP-WIT (400M)	COCO-Stuff	5.6	18.0	45.5	89.5
LSeg+ [24]	ALIGN (1.8B)	COCO-Stuff	3.8	18.0	46.5	-
SimBaseline [78]	CLIP-WIT (400M)	COCO-Stuff	7.0	20.5	47.7	88.4
MaskCLIP [85]	CLIP-WIT (400M)	COCO-Stuff	8.2	23.7	45.9	-
ODISE [77]	CLIP-WIT (400M)	COCO-Stuff	**11.1**	**29.9**	**57.3**	-
ProLab	ImageNet (1.28M)	COCO-Stuff	2.1	15.6	44.9	**90.6**
ProLab†	ImageNet (1.28M)	COCO-Stuff	3.5	23.1	57.7	92.5

Table 5: Open-vocabulary semantic segmentation. †: Linear probing with property-level logits (only 1 fully-connected layer is trainable). Without large-scale pre-training on image-text pairs for alignment, our method still shows strong open-vocabulary capability. Moreover, with a minimal linear projection (property-level → class-level), ProLab could beat a lot of methods with a large margin.

demonstrates the ability to generalize to unknown categories using properties such as "covered in fur" and "electronic device". This somehow mirrors human reasoning processes, showcasing our model's advanced capability to recognize and segment objects beyond its trained categories.

In Table 5, we quantitatively evaluate our method with the open vocabulary setting. Without pre-alignment on image-text pairs, our method still shows comparable performance on 4 benchmarks (*i.e.*, ADE-847, ADE-150, Pascal Context, and Pascal VOC). Moreover, with simple linear probing on property-level logits, ProLab significantly outperforms competing methods by a wide margin, showcasing the efficacy of our property-level label space. This underscores ProLab's promising open-vocabulary capabilities and is worthy for further exploration.

6 Conclusion

The quest for an interpretable semantic space modeling inter-class semantic correlations has been a long-standing goal in computer vision. Previous methods, such as manual label merging [37], hierarchical label spaces [40], or CLIP text encoders [38, 89], have fallen short. Addressing this, we propose **ProLab**, a semantic segmentation method leveraging a **pro**perty-level **lab**el space, drawing from the common sense knowledge base of Large Language Models (LLMs), aligning with human reasoning to enhance interpretability and relevance.

Empirically, ProLab shows superior performance and scalability across classic benchmarks, adeptly segmenting out-of-domain and unknown categories based on in-domain descriptive properties. This paves the way for future research to improve segmentation models beyond traditional category-level supervision, aiming for a holistic understanding of scenes and objects that mirrors human perception.

Acknowledgments

This work was supported by ONR grant N00014-23-1-2641. We thank Chen Wei, Zongwei Zhou, and Yaoyao Liu for their valuable discussions.

References

1. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for attribute-based classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 819–826 (2013) 4
2. Amit, Y., Fink, M., Srebro, N., Ullman, S.: Uncovering shared structures in multiclass classification. In: Proceedings of the 24th international conference on Machine learning. pp. 17–24 (2007) 4
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 **1**(2), 3 (2023) 4
4. Bao, H., Dong, L., Piao, S., Wei, F.: BEiT: BERT pre-training of image transformers. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=p-BhZSz59o4 10, 11
5. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2846–2854 (2016) 3
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 8
7. Bucher, M., Vu, T.H., Cord, M., Pérez, P.: Zero-shot semantic segmentation. Advances in Neural Information Processing Systems **32** (2019) 14
8. Chen, J., Yang, Z., Zhang, L.: Semantic segment anything. https://github.com/fudan-zvg/Semantic-Segment-Anything (2023) 3
9. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023) 4
10. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017) 1, 10
11. Chen, X., Li, S., Lim, S.N., Torralba, A., Zhao, H.: Open-vocabulary panoptic segmentation with embedding modulation. In: Proceedings of the IEEE/CVF international conference on computer vision (2023) 3
12. Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., Qiao, Y.: Vision transformer adapter for dense predictions. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=plKu2GByCNW 1, 8, 10, 11
13. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016) 1, 3, 9

14. Davis, E., Marcus, G.: Commonsense reasoning and commonsense knowledge in artificial intelligence. Communications of the ACM **58**(9), 92–103 (2015) 2
15. Dekel, O., Keshet, J., Singer, Y.: Large margin hierarchical classification. In: Proceedings of the twenty-first international conference on Machine learning. p. 27 (2004) 4
16. Deng, J., Ding, N., Jia, Y., Frome, A., Murphy, K., Bengio, S., Li, Y., Neven, H., Adam, H.: Large-scale object classification using label relation graphs. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. pp. 48–64. Springer (2014) 4
17. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 4, 8
18. Ding, J., Xue, N., Xia, G.S., Dai, D.: Decoupling zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11583–11592 (2022) 14
19. Durand, T., Mordan, T., Thome, N., Cord, M.: Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 642–651 (2017) 3
20. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**, 303–338 (2010) 1
21. Farhadi, A., Endres, I., Hoiem, D.: Attribute-centric recognition for cross-category generalization. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 2352–2359. IEEE (2010) 4
22. Fergus, R., Bernal, H., Weiss, Y., Torralba, A.: Semantic label sharing for learning with many categories. In: Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11. pp. 762–775. Springer (2010) 4
23. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013) 1
24. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Open-vocabulary image segmentation. In: ECCV (2022) 14
25. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: European Conference on Computer Vision. pp. 540–557. Springer (2022) 3
26. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021) 3
27. Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H.: Brain tumor segmentation with deep neural networks. Medical image analysis **35**, 18–31 (2017) 1
28. Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 108–124. Springer (2016) 4
29. Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural language object retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4555–4564 (2016) 4

30. Huang, S., Lu, Z., Cheng, R., He, C.: Fapn: Feature-aligned pyramid network for dense image prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 864–873 (2021) 11
31. Jain, J., Singh, A., Orlov, N., Huang, Z., Li, J., Walton, S., Shi, H.: Semask: Semantically masked transformers for semantic segmentation. arXiv preprint arXiv:2112.12782 (2021) 11
32. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1780–1790 (2021) 4
33. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 787–798 (2014) 4
34. Kim, D., Tsai, Y.H., Suh, Y., Faraki, M., Garg, S., Chandraker, M., Han, B.: Learning semantic segmentation from multiple datasets with label shifts. In: European Conference on Computer Vision. pp. 20–36. Springer (2022) 4
35. Knowlton, B.J., Squire, L.R.: The learning of categories: Parallel brain systems for item memory and category knowledge. Science **262**(5140), 1747–1749 (1993) 2
36. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. arXiv preprint arXiv:2308.00692 (2023) 4
37. Lambert, J., Liu, Z., Sener, O., Hays, J., Koltun, V.: Mseg: A composite dataset for multi-domain semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2879–2888 (2020) 1, 4, 14
38. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: International Conference on Learning Representations (2022) 1, 3, 14
39. Li, F., Zhang, H., Sun, P., Zou, X., Liu, S., Yang, J., Li, C., Zhang, L., Gao, J.: Semantic-sam: Segment and recognize anything at any granularity. arXiv preprint arXiv:2307.04767 (2023) 3
40. Li, L., Zhou, T., Wang, W., Li, J., Yang, Y.: Deep hierarchical semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1246–1257 (2022) 1, 4, 14
41. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022) 4
42. Liang, X., Zhou, H., Xing, E.: Dynamic-structured semantic propagation network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 752–761 (2018) 1
43. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 1, 3, 9
44. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023) 4
45. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12009–12019 (2022) 11

46. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021) 11
47. Lloyd, S.: Least squares quantization in pcm. IEEE transactions on information theory **28**(2), 129–137 (1982) 7
48. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015) 1
49. Marszalek, M., Schmid, C.: Semantic hierarchies for visual object recognition. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–7. IEEE (2007) 4
50. Menon, S., Vondrick, C.: Visual classification via description from large language models. In: International Conference on Learning Representations (2023) 12, 13
51. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 891–898 (2014) 1
52. Mukhoti, J., Lin, T.Y., Poursaeed, O., Wang, R., Shah, A., Torr, P.H., Lim, S.N.: Open vocabulary semantic segmentation with patch aligned contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19413–19423 (2023) 3
53. Palatucci, M., Pomerleau, D., Hinton, G.E., Mitchell, T.M.: Zero-shot learning with semantic output codes. Advances in neural information processing systems **22** (2009) 4
54. Peng, Z., Dong, L., Bao, H., Ye, Q., Wei, F.: Beit v2: Masked image modeling with vector-quantized visual tokenizers. arXiv preprint arXiv:2208.06366 (2022) 10, 11
55. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023) 4
56. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 2, 3
57. Rao, Y., Zhao, W., Tang, Y., Zhou, J., Lim, S.N., Lu, J.: Hornet: Efficient high-order spatial interactions with recursive gated convolutions. Advances in Neural Information Processing Systems **35**, 10353–10366 (2022) 11
58. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019) 6, 8, 11
59. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015) 1
60. Roth, H.R., Lu, L., Farag, A., Shin, H.C., Liu, J., Turkbey, E.B., Summers, R.M.: Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18. pp. 556–564. Springer (2015) 1
61. Russakovsky, O., Fei-Fei, L.: Attribute learning in large-scale datasets. In: Trends and Topics in Computer Vision: ECCV 2010 Workshops, Heraklion, Crete, Greece,

September 10-11, 2010, Revised Selected Papers, Part I 11. pp. 1–14. Springer (2012) 4
62. Sharmanska, V., Quadrianto, N., Lampert, C.H.: Augmented attribute representations. In: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12. pp. 242–255. Springer (2012) 4
63. Tousch, A.M., Herbin, S., Audibert, J.Y.: Semantic hierarchies for image annotation: A survey. Pattern Recognition **45**(1), 333–345 (2012) 4
64. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021) 8
65. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 8
66. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. arXiv preprint arXiv:2305.11175 (2023) 4
67. Wang, X., Ye, Y., Gupta, A.: Zero-shot recognition via semantic embeddings and knowledge graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6857–6866 (2018) 4
68. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019) 8
69. Wu, C., Lin, Z., Cohen, S., Bui, T., Maji, S.: Phrasecut: Language-based image segmentation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10216–10225 (2020) 4
70. Xian, Y., Choudhury, S., He, Y., Schiele, B., Akata, Z.: Semantic projection network for zero-and few-label semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8256–8265 (2019) 14
71. Xiao, J., Xu, Z., Yuille, A., Yan, S., Wang, B.: Palm2-vadapter: Progressively aligned language model makes a strong vision-language adapter. arXiv preprint arXiv:2402.10896 (2024) 4
72. Xiao, J., Xu, Z., Lan, S., Yu, Z., Yuille, A., Anandkumar, A.: 1st place solution of the robust vision challenge 2022 semantic segmentation track. arXiv preprint arXiv:2210.12852 (2022) 4
73. Xiao, S., Liu, Z., Zhang, P., Muennighoff, N.: C-pack: Packaged resources to advance general chinese embedding (2023) 6, 8
74. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV). pp. 418–434 (2018) 1, 8, 10
75. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021) 1, 10
76. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18134–18144 (2022) 3
77. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models. arXiv preprint arXiv:2303.04803 (2023) 3, 14

78. Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., Bai, X.: A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. In: ECCV. pp. 736–753 (2022) 14
79. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18155–18165 (2022) 4
80. Ye, L., Rochan, M., Liu, Z., Wang, Y.: Cross-modal self-attention network for referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10502–10511 (2019) 4
81. Yu, F.X., Cao, L., Feris, R.S., Smith, J.R., Chang, S.F.: Designing category-level attributes for discriminative visual recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 771–778 (2013) 4
82. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2636–2645 (2020) 1, 3, 9
83. Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Yang, J., Zhang, L.: A simple framework for open-vocabulary segmentation and detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1020–1031 (2023) 3
84. Zhang, P., Xiao, S., Liu, Z., Dou, Z., Nie, J.Y.: Retrieve anything to augment large language models. arXiv preprint arXiv:2310.07554 (2023) 11
85. Zheng Ding, Jieke Wang, Z.T.: Open-vocabulary universal image segmentation with maskclip. In: International Conference on Machine Learning (2023) 3, 14
86. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016) 3
87. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017) 1, 3, 9
88. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European Conference on Computer Vision (ECCV) (2022) 3
89. Zhou, Q., Liu, Y., Yu, C., Li, J., Wang, Z., Wang, F.: Lmseg: Language-guided multi-dataset segmentation. In: International Conference on Learning Representations (2023) 1, 4, 14
90. Zhu, C., Chen, F., Ahmed, U., Shen, Z., Savvides, M.: Semantic relation reasoning for shot-stable few-shot object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8782–8791 (2021) 4
91. Zou, X., Dou, Z.Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., et al.: Generalized decoding for pixel, image, and language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15116–15127 (2023) 3
92. Zweig, A., Weinshall, D.: Exploiting object hierarchy: Combining models from different category levels. In: 2007 IEEE 11th international conference on computer vision. pp. 1–8. IEEE (2007) 4