S3A: Towards Realistic Zero-Shot Classification via Self Structural Semantic Alignment

Sheng Zhang[1], **Muzammal Naseer**[1], **Guangyi Chen**[1, 2], **Zhiqiang Shen**[1], **Salman Khan**[1, 3], **Kun Zhang**[1, 2], **Fahad Shahbaz Khan**[1, 4]

[1]Mohamed bin Zayed University of Artificial Intelligence
[2]Carnegie Mellon University
[3]Australian National University
[4]Linköping University
{sheng.zhang, muzammal.naseer, guangyi.chen, zhiqiang.shen, salman.khan}@mbzuai.ac.ae
kunz1@cmu.edu, fahad.khan@liu.se

Abstract

Large-scale pre-trained Vision Language Models (VLMs) have proven effective for zero-shot classification. Despite the success, most traditional VLMs-based methods are restricted by the assumption of partial source supervision or ideal target vocabularies, which rarely satisfy the open-world scenario. In this paper, we aim at a more challenging setting, *Realistic* Zero-Shot Classification, which assumes no annotation but instead a broad vocabulary. To address the new problem, we propose the Self Structural Semantic Alignment (**S^{3}A**) framework, which extracts the structural semantic information from unlabeled data while simultaneously self-learning. Our **S^{3}A** framework adopts a unique Cluster-Vote-Prompt-Realign (CVPR) algorithm, which iteratively groups unlabeled data to derive structural semantics for pseudo-supervision. Our CVPR algorithm includes iterative clustering on images, voting within each cluster to identify initial class candidates from the vocabulary, generating discriminative prompts with large language models to discern confusing candidates, and realigning images and the vocabulary as structural semantic alignment. Finally, we propose to self-train the CLIP image encoder with both individual and structural semantic alignment through a teacher-student learning strategy. Our comprehensive experiments across various generic and fine-grained benchmarks demonstrate that the **S^{3}A** method substantially improves over existing VLMs-based approaches, achieving a more than 15% accuracy improvement over CLIP on average. Our codes, models, and prompts are publicly released at https://github.com/sheng-eatamath/S3A.

Introduction

In recent years, large-scale pre-trained Vision Language Models (VLMs) such as CLIP (Radford et al. 2021; Ren et al. 2022), ALIGN (Li et al. 2021), and BLIP (Li et al. 2022, 2023) have garnered significant attention for their remarkable zero-shot generalization ability on multifarious downstream tasks, particularly in recognizing unseen categories (Zhang et al. 2023a). The common practice to leverage this ability is packing category names into a textual prompt (e.g., "A photo of a `[CLS]`") and aligning image embeddings with text embeddings of filled prompts in VLM joint embedding space for classification. To adapt pre-trained VLMs to downstream unseen data, existing prevailing methods (Wu et al. 2023; Zang et al. 2022; Ghiasi et al. 2021) usually assume the access to source labeled data (Chen et al. 2022; Khattak et al. 2023; Zhou et al. 2022a) (*e.g.*, in zero-shot learning (Zhou et al. 2022b; Gao et al. 2021)), target label distribution (*e.g.*, in unsupervised prompt tuning (Kahana, Cohen, and Hoshen 2022)), or an *ideal vocabulary* that exactly matches the ground-truth label set or with very few open words (*e.g.*, in open-vocabulary learning (Wu et al. 2023; Zang et al. 2022; Ghiasi et al. 2021)). However, this ideal vocabulary is unattainable without exhaustive annotation of all unseen data; whereas, human annotations are exorbitant and difficult to scale. Therefore, both assumptions are restrictive and impractical in open-world scenarios with diverse and dynamic nature.

In this paper, we embark on a journey towards *Realistic* Zero-Shot Classification (RZSC), a more challenging yet practical problem compared with conventional zero-shot learning due to its realistic conditions. Here, we term *Realistic* as the realistic nature of RZSC which aims to recognize categories on unseen datasets without annotation and ideal vocabulary, but with a vast, comprehensive vocabulary with more than 20K category names encompassing all common classes (Sariyildiz et al. 2021; Ridnik et al. 2021). However, it is challenging since the vast vocabulary can lead to alignment confusion among fine-grained options; as we witness the consistent and dramatic CLIP (Radford et al. 2021) performance drops and reduced neighborhood ranges in Fig. 1.

To confront this challenge, we introduce the Self Structural Semantic Alignment (**S^{3}A**) framework, which iteratively discovers structural semantic alignment from unlabeled data for joint self-learning. This is orchestrated through our unique Cluster-Vote-Prompt-Realign (CVPR) algorithm, a principled process comprising four key steps: `(1)` **Clustering** unearths inherent grouping structures of image embeddings, producing meaningful image semantics. `(2)` **Voting** associates each cluster with initial category candidates, representing potential structural semantic alignments. These two steps can be executed iteratively to obtain more reliable candidates. `(3)` **Prompting** leverages the power of large language models (LLMs) to discern

Setting	Vocab.	Anno.	Train
Zero-Shot Transfer	$\mathcal{Y}_{tgt}$	✗	$\mathcal{Y}_{tgt}$
Zero-Shot Classification	$\mathcal{Y}_{tgt}$	✓	$\mathcal{Y}_{base}$
Open-Vocabulary Learning	<2K	✓	$\mathcal{Y}_{base}$
Unsupervised Fine-tuning	$\mathcal{Y}_{tgt}$	✗	$\mathcal{Y}_{tgt}$
RZSC	>20K	✗	$\mathcal{Y}_{tgt}$

Table 1: Our realistic zero-shot classification and other related settings. Here, following (Wu et al. 2023), we denote $\mathcal{Y}_{base}$ and $\mathcal{Y}_{tgt}$ as sets of base training classes and target testing classes, which satisfies $\mathcal{Y}_{base} \bigcap \mathcal{Y}_{tgt} = \phi$. The learning goal of all settings is to recognize $\mathcal{Y}_{tgt}$ in test data.

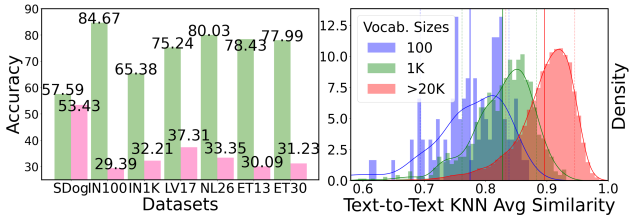

Figure 1: **(a)** Performance comparison between CLIP w/ an ideal vocabulary (Green) and w/ a large vocabulary of 20K categories (Pink). **(b)** Distribution plot of text-to-text average 3-Nearest Neighbors cosine similarity of each text embedding for three types of vocabulary: with ImageNet-100, ImageNet-1K, and 20K category names.

nuanced candidates by augmenting prompts with discriminative attributes. (4) **Re-alignment** represents calibrating the cluster-vocabulary alignment with LLM-augmented prompts as pseudo structural semantic alignment labels. Incorporating our CVPR algorithm, our **S^{3}A** framework self-trains a student model based on derived individual and structural semantic alignment labels from a stable teacher. Simultaneously, the teacher is updated by student weights to produce more reliable pseudo semantic alignments.

We extensively evaluate our **S^{3}A** framework across multiple setups, spanning various generic and fine-grained benchmarks. The results show that **S^{3}A** not only consistently outperforms previous adapted State-of-The-Arts (SOTAs) under the RZSC setting on all benchmarks, but excels in out-of-vocabulary evaluation, where category names can fall outside the **S^{3}A** vocabulary. Comprehensive evaluations evidence the effectiveness of our **S^{3}A** framework in RZSC.

Our contributions include: (1) We propose a Self Structural Semantic Alignment (**S^{3}A**) framework, to address the challenging *Realistic Zero-Shot Classification* problem, which jointly extracts and self-learns on the individual and structural semantic alignment. (2) We propose a Cluster-Vote-Prompt-Realign algorithm to reliably derive reliable structural semantic alignments between images and the large vocabulary. (3) **S^{3}A** achieves SOTA performance on various generic and fine-grained benchmarks, remarkably boosting CLIP by over 15% accuracy, and even in the out-of-vocabulary scenarios.

Related Work

Zero-Shot Learning/Open-Vocabulary Learning with VLMs. Traditional *(Generalized) Zero-Shot Classification* (ZSC) aims to categorize novel classes in unseen test data with training on annotated base/seen classes with or without unlabeled target novel classes (Wang et al. 2019; Pourpanah et al. 2022). However, they usually assume auxiliary semantic information of both seen and unseen classes, *e.g.*, category attributes (Lampert, Nickisch, and Harmeling 2009), knowledge graph (Akata et al. 2015), textual keywords (Lei Ba et al. 2015; Cappallo, Mensink, and Snoek 2016). Recently, large-scale pre-trained VLMs have been introduced to alleviate these assumptions (Jia et al. 2021; Radford et al. 2021; Zhang et al. 2023a). Furthermore, *Open-Vocabulary Learning* (OVL) (Wu et al. 2023; Zhou et al. 2023; Zhou, Loy, and Dai 2022; Karazija et al. 2023) aims to train the models with some annotated data, *i.e.*, base classes, or large-scale image-text pairs, and to test them on target novel classes (Xu et al. 2023; Shin, Albanie, and Xie 2023). Our RZSC setting differs from conventional ZSC and OVL in not requiring any labeled training data, and not assuming an *ideal vocabulary* with a ground-truth target label set or one with few open words (Wu et al. 2023; Xu et al. 2023; Karazija et al. 2023).

Zero-Shot Transfer/Unsupervised Fine-tuning of VLMs. Both *Zero-Shot Transfer* (ZST) and *Unsupervised Fine-tuning* (UF) assume no annotations of target datasets, which are essentially visual concept discovery problems (Vaze et al. 2022; Wen, Zhao, and Qi 2023; Zhang et al. 2023b) with vocabulary prior. ZST (Radford et al. 2021; Ren et al. 2022) directly uses the pre-trained VLMs for zero-shot prediction without fine-tuning. UF further transductively adapts pre-trained models with task-specific training, *e.g.*, with self-training or prompt tuning (Li, Savarese, and Hoi 2022; Kahana, Cohen, and Hoshen 2022; Shin, Albanie, and Xie 2023). However, both ZST&UF assume known ground-truth target label sets or distribution (Kahana, Cohen, and Hoshen 2022; Li, Savarese, and Hoi 2022). In this paper, we aim to alleviate the reliance on these assumptions and propose a new setting, RZSC. Besides, an extended ZST work, SCD (Han et al. 2023), iteratively refines CLIP zero-shot inference predictions on a WordNet vocabulary (Miller 1995) with a heuristic clustering algorithm. However, they have limited adaptability (Li, Savarese, and Hoi 2022), a mismatched linguistic vocabulary still based on the closed-world assumption.

Discussion on Zero-Shot Settings. Here, we summarize the main differences between our RZSC setting and others in Table 1. Previous related settings adopt restrictive assumptions including an ideal vocabulary, the target label distribution, and labeled base classes. By contrast, our RZSC aims to learn to categorize an unlabeled dataset with a huge vocabulary based on a large visual taxonomy with over 20K classes. An expanded vocabulary presents significant challenges for RZSC problem, as evidenced by the consistent and substantial CLIP performance drop (Fig. 1a) on all datasets when the vocabulary scales up. The primary challenge arises from

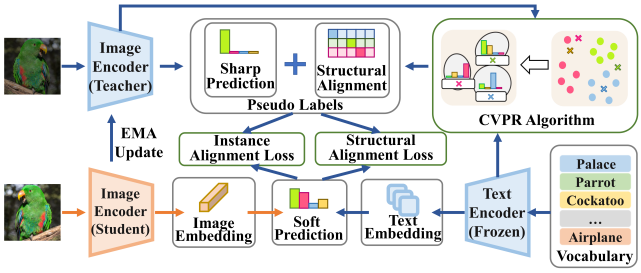

Figure 2: Illustration of our Self Structural Semantic Alignment ($\mathbf{S^3A}$) framework, which fine-tunes pre-trained CLIP encoder with a teacher-student architecture. The teacher is updated by the student in an exponentially moving average manner. The student is guided by on-the-fly one-hot instance alignment predicted by the teacher, and self-trains with structural semantic alignment labels derived by our per-epoch CVPR algorithm on all teacher image embeddings.

increased confusing open words, complicating fine-grained category discrimination for pre-trained VLMs. As displayed in Fig. 1b, the averaged cosine similarity between a query text embedding and its 3-nearest text neighbors grows with the vocabulary size.

Methodology

Problem: Realistic Zero-Shot Classification

Existing methods that adapt pre-trained VLMs to unseen data usually rely on specific knowledge of target datasets, such as prior distributions or an ideal vocabulary. These conditions are often challenging to fulfill in real-world environments. In this paper, we explore a more practical task, Realistic Zero-Shot Classification, abbreviated as RZSC.

RZSC is formally defined as follows: Consider an unlabeled dataset $\mathcal{D}_u = \{(\mathbf{x}_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times \mathcal{Y}$ with N images, where $\mathcal{Y}$ is the underlying category set, and a pre-trained VLM such as CLIP, equipped with image and text encoders f_I and f_T, respectively. Then, we assume no information of $\mathcal{Y}$ and instead with a comprehensive vocabulary that contains more than 20,000 distinct category names, *i.e.*, $|\mathcal{Y}| \ll |\mathcal{W}|$. We build our vocabulary from all visual categories from ImageNet1K (Deng et al. 2009) and ImageNet21K (Ridnik et al. 2021) datasets since they are annotated with expert taxonomic knowledge (Miller 1995) and encompasses most commonly-seen visual categories in the real world (Sariyildiz et al. 2021). The goal of the RZSC task is to adapt the pre-trained VLM, *i.e.*, f_I, f_T to predict the correct category of an unseen dataset:

$$\hat{y}_i = \arg\max_{w_j \in \mathcal{W}} \mathbf{z_i} \cdot \mathbf{h}_j, \tag{1}$$

where $\mathbf{z}_i = f_I(\mathbf{x}_i)$ denotes the image embedding while text embedding $\mathbf{h}_j = f_T(w_j)$ are obtained with a text prompt, *e.g.*, "a photo of a {category}", and the category name w_j. Here, we denote $\cdot$ as cosine similarity.

Overview: Self Structural Semantic Alignment

RZSC presents a more formidable challenge than previous tasks, primarily owing to the absence of label information and an increased vocabulary size. As illustrated in part (a) of Fig. 1, the performance of CLIP declines sharply as the vocabulary size increases. This decline can be attributed to the inclusion of confusing open words as hard negative samples, which introduces noise to pre-trained CLIP, hindering its ability to accurately identify image-category alignments.

We are motivated to propose our Self Structural Semantic Alignment ($\mathbf{S^3A}$) framework, which discovers the structural semantics through iterative self-alignment between visual images and textual vocabulary. As shown in Fig. 2, our $\mathbf{S^3A}$ incorporates a Cluster-Vote-Prompt-Realign (CVPR) algorithm to derive structural semantics as alignment labels, and both models and pseudo alignments are iteratively refined during self-training. Our CVPR algorithm and $\mathbf{S^3A}$ self-training procedure can achieve a synergistic effect: as training progresses in adapting representations, the teacher model can provide increasingly reliable pseudo alignments in subsequent iterations. Concurrently, the CVPR algorithm contributes structural semantics as a refined supervisory signal for subsequent self-training. We elaborate on all components in the sections that follow.

Cluster-Vote-Prompt-Realign (CVPR)

Our Cluster-Vote-Prompt-Realign algorithm lies at the heart of the $\mathbf{S^3A}$ framework, representing an innovative approach to uncovering structural semantics in data. As illustrated in Fig. 3, our CVPR algorithm consists of four key stages, each contributing to the alignment and identification of structural relationships between visual images and textual vocabulary, including discovering semantic clusters, voting category names on large vocabulary, prompting LLM to discriminate the nuanced candidates, and refine the cluster-vocabulary alignment. Each step is explained in detail in the subsequent paragraphs. Below we delineate these stages and their functions within the algorithm.

Clustering. Based on existing evidence (Radford et al. 2021) and our observation, the pre-trained CLIP excels at grouping instances with the same or similar semantic labels in the image embedding space. We thus produce the pseudo supervision by semantic clustering and aligning the clusters with vocabulary. Specifically, given image embeddings $\mathbf{z_i}$ in $\mathcal{D}_u$, we apply KMeans (Arthur and Vassilvitskii 2007) to obtain the K clusters, $\Gamma = \{\Gamma_k\}_{k=1}^K$, where Γ_k denotes the k-th set of image embeddings.

Voting. Given the semantic cluster results Γ, we compute a vocabulary voting distribution matrix $M \in \mathbf{R}^{K \times |\mathcal{W}|}$, where $M_{k,j}$ represents the normalized frequency of the prototype of category w_j being the nearest neighbor to all instances in the k-th cluster. Specifically, it is computed as

$$M_{k,j} = \frac{1}{K|\Gamma_k|} \sum_{\mathbf{z} \in \Gamma_k} \mathbb{I}(w_j = \arg\max_w \mathbf{z} \cdot \mathbf{h}) \tag{2}$$

where $\mathbb{I}$ is an indicator function, and $|\Gamma_k|$ denotes the size of the k-th cluster. M is cluster-wise and vocabulary-wise normalized, with $||M||_1 = 1$. Rather than naively assigning each cluster to the argmax prototype in the vocabulary, we keep the top-m frequent words for each cluster as potential candidates which are treated equally. For each row $M_k = (M_{k,j})_{j=1}^{|\mathcal{W}|}$, we set all entries but the highest m ones as 0.

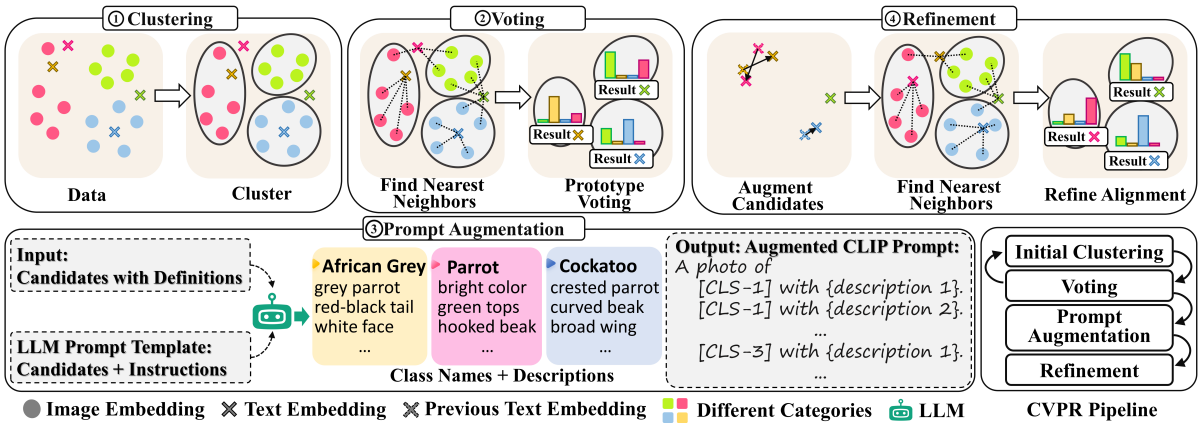

Figure 3: An illustrative toy example for our CVPR algorithm, comprising four steps: (1) We cluster all image embeddings. (2) We conduct 1-nearest neighbor voting on all text prototypes of the large vocabulary for each cluster. Since the results of the naive assignment in this step are susceptible to the noise of text embeddings, we generate cluster-wise candidate categories instead. (3) We augment CLIP text prompts with visual discriminative descriptions from the large language model to discern nuanced candidates. (4) With augmented prompts, the cluster-vocabulary alignment is calibrated and refined.

Nonetheless, the initial clustering and voting may introduce noise, leading to low-quality pseudo-labels. To mitigate this issue, we iteratively refine the previous clusters based on the current voting outcomes. In particular, we utilize the Hungarian matching (Kuhn 1955) for textual embeddings and clusters to align each cluster with a single prototype. Subsequently, we reassign the image embeddings, using these prototypes as the updated cluster centers (Han et al. 2023). We iterate this process three times.

Prompting. Through our empirical studies, we observed that CLIP representation struggles to differentiate nuanced candidates effectively. This observation spurred our efforts to refine the embeddings of textual candidates. We speculate that the challenge in distinguishing fine-grained options arises from the presence of noisy or ambiguous image-caption alignments during CLIP pre-training.

To address this challenge, our approach is to enhance the conventional CLIP prompts by accentuating the subtle semantic differences. We achieve this by integrating auxiliary textual knowledge drawn from LLMs, which are effective in knowledge retrieval (Dale 2021; Yang et al. 2021). Specifically, we feed m candidate category words of the k-th cluster into a single LLM prompt template, each accompanied by their specific definition. Then, we add an instruction to the prompt to extract nuanced visual attributes of each category from the LLM. Our prompt template is structured as:

Prompt: Given visual concepts: [CLS-1]: [DEF-1], ..., [CLS-m]: [DEF-m].

Instruction: To discriminate these visual concepts in a photo. Please list all possible visual descriptive phrases for each visual concept.

In this template, `[CLS]` represents the category name, and `[DEF]` stands for its definition from WordNet (Miller 1995). The LLM then generates a list of distinctive attributes for each category, such as 'red-and-black tail'. To avoid linguistic ambiguity arising from the polysemy phenomenon, we utilize all possible synset-definition pairs in WordNet (Miller 1995) for a single category as the input visual concepts for the LLM prompt. Finally, each (category, attribute) pair is filled into a CLIP prompt for augmentation, *e.g.*, "A photo of a {category} with {attribute}.". An ensemble of augmented text embeddings for each category name is constituted.

Re-alignment. During the re-alignment phase, our goal is to enhance the structural semantic alignments in Eq. 2. The refined re-alignment matrix, $\tilde{M} \in \mathbf{R}^{K\times|\mathcal{W}|}$, is derived by casting votes on all augmented text embeddings generated in the previous prompting stage. Specifically, the re-alignment probability between the k-th cluster and word w_j is determined by the frequencies of augmented embeddings of the word w_j in $\mathcal{A}_k$ being the top-3 nearest neighbors of $\mathbf{z} \in \Gamma_k$. We denote $\mathcal{A}_k$ as the set of augmented embeddings of all candidate category words of Γ_k. It can be formulated as:

$$\tilde{M}_{k,j} = \frac{\alpha_{w_j}}{3K|\Gamma_k|} \sum_{\mathbf{z}\in\Gamma_k} \mathbb{I}\left(w_j \in \arg\operatorname*{top3}_{w}(\mathbf{z}, \mathcal{A}_k)\right) \quad (3)$$

where arg extracts the category name linked with the augmented text embedding in $\mathcal{A}_k$. To avoid the imbalance issue raised by varied numbers of augmented embeddings of different category names, we consider the weight factor $\alpha_{w_j} = \frac{1}{|\mathcal{A}_k(w_j)|}$, which uniformly distributes total mass 1 to all augmented embeddings of w_j. Therefore, each row of $\tilde{M}$ sums to be $\frac{1}{K}$, and $||\tilde{M}||_1 = 1$. We again employ the maximum Hungarian matching (Kuhn 1955) on a bipartite graph between clusters and category words, with the cost matrix $\tilde{M}$. Consequently, the structural alignment is obtained from the solution, which enforces a one-to-one mapping between clusters and category names.

Self-training with Semantic Alignment

In this section, we present our S^{3}A self-training framework, as depicted in Fig. 2. The self-training process leverages

both instance-wise and structural alignment pseudo labels which are derived by our CVPR algorithm with an exponentially moving averaged (EMA) teacher model (Grill et al. 2020). Throughout this process, we adapt CLIP image encoder to enhance its representation and fix its text encoder.

Structural Semantic Alignment. To incorporate the structural semantic alignments into online learning, one challenge needs to be addressed. Obtaining high-quality structural alignment pseudo-labels requires consistent model embeddings from the entire dataset, which is computationally costly; while determining the optimal execution interval of CVPR across datasets is challenging. To mitigate these issues, we introduce a slowly updated EMA teacher model. It provides stably refined embeddings and executes the CVPR algorithm once per epoch to yield stable and reliable structural pseudo alignments, which then guides the self-training of the student model.

We define the structural semantic alignment loss as the cross-entropy between the predictions of the student model and the pseudo structural alignments generated by the teacher model. Formally, this loss for the i-th instance can be expressed as:

$$L_{str}(\mathbf{x}_i) = -\hat{\mathbf{p}}_T^T(i) \log \mathbf{p}_S(\mathbf{x}_i). \tag{4}$$

In this equation, $\hat{\mathbf{p}}_T(i)$ represents the one-hot pseudo structural alignment for the i-th instance, which is inferred from the teacher CVPR results during the last epoch. On the other hand, $\mathbf{p}_S$ denotes the softmax prediction of the student model over the entire vocabulary, computed for the input $\mathbf{x}_i$. As a result, the sharpened pseudo labels can cluster images with the same semantics as well as align clusters.

Individual Semantic Alignment. In addition to the structural semantic alignment loss, we also guide our model with instance-wise pseudo alignments, which are generated on-the-fly by the EMA teacher model. Without this guidance, our model would likely converge to suboptimal solutions rapidly. We formulate the individual semantic alignment loss for the i-th instance as follows:

$$L_{in}(\mathbf{x}_i) = -\mathbb{I}(\tilde{\mathbf{p}}_T(\mathbf{x}_i) > \tau) \tilde{\mathbf{p}}_T^T(\mathbf{x}_i) \log \mathbf{p}_S(\mathbf{x}_i). \tag{5}$$

In this equation, $\tilde{\mathbf{p}}_T$ represents the one-hot sharpened pseudo label produced by the teacher model at each iteration. The symbol τ denotes a confidence threshold, which ensures that the loss is computed only for samples for which the teacher model has a high level of confidence.

To strike a balance between the structural and instance alignment losses, we introduce a weighted combination of both. In this way, individual alignment retains original instance alignment information, while structural alignment groups and aligns similar semantics. Consequently, our total loss function for the i-th instance is formulated as:

$$L(\mathbf{x}_i) = L_{str}(\mathbf{x}_i) + \gamma L_{in}(\mathbf{x}_i). \tag{6}$$

Here, γ represents a balancing factor that weights the contribution of the instance alignment loss relative to the structural alignment loss. This total loss is computed at each iteration, based on our CVPR algorithm which is executed once per epoch on the teacher model.

Experiments

Evaluation

Datasets. We evaluate **S^{3}A** on two generic and five fine-grained benchmarks, *i.e.*, the generic benchmarks of sampled ImageNet-100 (IN100) and ImageNet-1K (IN1K) (Deng et al. 2009), and fine-grained benchmarks of StanfordDogs-120 (SDogs) (Khosla et al. 2011), Living17-68 (LV17), Nonliving26-104 (NL26), Entity13-260 (ET13), and Entity30-240 (ET30) in BREEDS (Santurkar, Tsipras, and Madry 2020)). Furthermore, we evaluate our **S^{3}A** on three benchmarks for the out-of-vocabulary evaluation (containing categories out of our vocabulary), *i.e.*, Oxford-IIIT Pet (Pet) (Parkhi et al. 2012), CIFAR100 (Krizhevsky, Hinton et al. 2009), and Caltech101(Clatech) (Fei-Fei, Fergus, and Perona 2004).

Metrics. We adopt the top-1 classification accuracy and clustering accuracy (following SCD (Han et al. 2023) and defined below) for the evaluation.

$$\text{ACC}_{clu} = \frac{1}{N} \sum_{i=0}^{N} \max_{\rho} \mathbb{I}(y_i = \rho(\hat{y}_i)), \tag{7}$$

where ρ is a permutation assignment of cluster indices. y_i and $\hat{y}_i$ are ground-truth predicted categories. Meanwhile, we adopt Intersection-over-Union (IoU) score as an auxiliary metric in ablations to inspect the overlap between our predictions $\mathcal{Y}_{pred}$ and the ground-truth label set $\mathcal{Y}_{gt}$, *i.e.*, $\frac{|\mathcal{Y}_{pred} \bigcap \mathcal{Y}_{gt}|}{|\mathcal{Y}_{pred} \bigcup \mathcal{Y}_{gt}|}$. In the out-of-vocabulary experiments, some class names cannot be found in the vocabulary. Thus, we instead apply a soft accuracy score, defined as the similarity between the predicted word (in vocabulary) and the ground truth label. Inspired by BertScore (Zhang et al. 2019), we adopt a language model, Sentence-Bert (Reimers and Gurevych 2019), to calculate the similarity.

Baselines. RZSC is a new setting in which few baselines are ready-to-use. Thus, we evaluate the baseline methods by reproducing them with officially released codes in our setting. Specifically, we consider CLIP as the naive baseline, and two state-of-the-art methods in ZST and UF, *i.e.*, SCD (Han et al. 2023) and MUST (Li, Savarese, and Hoi 2022). In summary, the following baselines are included for performance comparisons:

- **DINO+KMeans** (Caron et al. 2021): DINO is an contrastive self-supervised learning method. We include it here for clustering quality comparisons. We only report its clustering accuracy as it cannot classify.
- **CLIP** (Radford et al. 2021): a large-scale VLM pre-trained with massive image-caption pairs conducts zero-shot prediction given our vocabulary.
- **CLIP (Group)** (Radford et al. 2021): We sequentially conduct clustering, voting, and Hungarian matching on CLIP image embeddings for structural zero-shot transfer, using **S^{3}A** vocabulary.
- **CLIP (Ideal)** (Radford et al. 2021): it denotes zero-shot transfer with pre-trained CLIP but given an ideal vocabulary, showcasing the upper bound performance of CLIP representation.

Methods	SDogs	IN100	IN1K	LV17	NL26	ET13	ET30	Avg
CLIP (Ideal)	57.59/58.07	84.67/84.90	65.38/65.53	75.24/75.53	80.03/80.05	78.43/78.50	77.99/78.03	74.19/74.37
DINO+KMeans	-/45.99	-/75.16	-/55.27	-/72.52	-/62.81	-/67.37	-/64.69	-/63.40
CLIP	53.43/55.43	29.39/38.54	32.21/39.77	37.31/47.24	33.35/38.96	30.09/40.00	31.23/39.90	35.29/42.83
CLIP (Group)	19.37/55.92	40.62/77.68	26.41/56.92	38.33/68.81	41.09/70.51	32.85/71.08	36.36/70.78	33.57/67.38
MUST	57.20/60.61	33.37/52.56	28.97/37.00	31.71/49.35	35.30/48.68	38.46/58.25	33.41/47.08	36.92/50.50
SCD	52.63/55.93	48.89/77.39	37.06/57.00	43.33/68.81	52.18/71.84	40.46/71.25	46.29/70.89	45.83/67.59
S^{3}A (Our)	**58.94/62.19**	**52.08/82.76**	**42.43/63.15**	**48.34/75.57**	**56.20/75.97**	**45.21/76.92**	**50.41/76.14**	**50.51/73.24**

Table 2: Transductive evaluation on seven benchmarks. Top-1 classification accuracy scores (left of '/') and clustering accuracy scores (right of '/') are reported in percentage. We highlight the highest scores except for the upper bound.

#Row	Prompt	S.T.	L_{str}	ImageNet-100 Acc	ImageNet-100 Cluster	Living17 Acc	Living17 Cluster
1	✗	✗	✗	48.89	77.39	43.33	68.81
2	✓	✗	✗	51.81	79.38	44.60	68.69
3	✗	✓	✗	46.23	81.49	46.28	74.60
4	✗	✓	✓	49.00	82.08	46.55	73.04
5	✓	✓	✓	**52.08**	**82.76**	**48.34**	**75.57**

Table 3: Top-1 accuracy and Clustering results for our method ablations on IN-100 and LV17. We conduct ablations on our discriminative prompt augmentation (Prompt), self-training stage (S.T.), and structural semantic alignment loss (L_{str}).

- **MUST** (Li, Savarese, and Hoi 2022): it is an unsupervised ZSC method leveraging instance-wise unsupervised self-training jointly with self-supervised masked-image prediction. We adapt it with our huge vocabulary.
- **SCD** (Han et al. 2023): it is an unsupervised/semi-supervised zero-shot transfer method with WordNet vocabulary. Its iterative algorithm aligns each cluster with one category name. We adapt it with our **S^{3}A** vocabulary.

Implementation Details. In our method, we fix $m = 3$ and $\gamma = 0.25$ on all datasets. Considering efficiency, we only compute prompting at the first epoch. We adopt ViT-B/16 (Dosovitskiy et al. 2020) as our CLIP backbone. Our data augmentations and optimizer follow MUST (Li, Savarese, and Hoi 2022). We train on all datasets for up to 30K iterations, with 60 epochs for Pet and 30 epochs for other datasets. Besides, we linearly warmup the EMA decay parameter to 0.9998 within specified iterations. We set the initial EMA weight decay of Pet and other datasets as 0.99 and 0.999, respectively. The warmup iterations are 500 for CIFAR, 100 for Pet, and 2000 for other datasets. The threshold τ are 0.3 for CIFAR, 0.7 for Pet, and 0.5 for other datasets. During inference, we adopt the teacher model for prediction on entire **S^{3}A** vocabulary. Experiments are conducted on a single A6000 GPU.

Main Results

To validate the effectiveness of our proposed method, **S^{3}A**, we conducted an extensive evaluation under RZSC setting. We compared our **S^{3}A** with various baselines on both fine-grained and generic datasets. The results are in Table 2.

Our method, **S^{3}A**, consistently achieves SOTA results, outperforming CLIP by a substantial margin, *i.e.*, $+15\%$ in top-1 accuracy. Furthermore, **S^{3}A** notably excels over our adapted SOTA baselines, with nearly $+5\%$ top-1 accuracy and $+6\%$ clustering accuracy. Generally, we can observe that more classes introduce challenges, and fine-grainedness decreases clustering quality but improves alignment accuracy, *e.g.*, IN100, NL26. Besides, CLIP (Group) encounters alignment issues though with quality clustering, as seen on IN1K and SDogs. We argue that our **S^{3}A** can dynamically calibrate noisy clustering during self-training. Note that the existing UF SOTA, MUST, sometimes degenerates its initial representation when on **S^{3}A**vocabulary. This underlines the significance of structural alignment learning for RZSC.

Ablations and Analysis

Method Ablations. To validate the contribution of **S^{3}A** components, we conduct method ablations on one generic and fine-grained dataset, *i.e.*, IN100 and LV17. We present the results in Table 3. The last row represents our full method. When we only keep the initial iterative clustering in our CVPR (the 1^{st} row), our method is equivalent to SCD (Han et al. 2023). The 2^{nd} row denotes our CVPR without all self-training-related components; while, the 3^{rd} row conducts self-training only with instance-wise semantic alignment. The 4^{th} row indicates our **S^{3}A**without LLM knowledge guidance. Based on the results, we can conclude that: `(1)` From Row 4&5, although the clustering quality remains comparable without our discriminative prompt augmentation, the semantic alignment degrades, as witnessed by the drop in top-1 accuracy. `(2)` From Row 1&2&3, self-training with structural alignment dominates the contribution in representation adaptation, witnessed by the cluster performance boosts. `(3)` From Row 3&4, we observe that the structural alignment w/o prompt augmentation yields great improvements on generic datasets, while its effect is less pronounced on fine-grained datasets due to the lack of language signals to discriminate among similar visual categories. All components contribute to the performance.

Performance on Estimated K. In Table 4, we present performance with estimated K instead of the ground-truth in Table 2. We implement an iterative estimation approach with three passes: first, we scan the range [LB_0, UB_0], then [LB_0,

Methods	LV17 (73)	NL26 (101)	ET13 (252)	ET30 (206)
DINO+KMeans	-/72.68	-/62.93	-/67.41	-/63.22
CLIP	37.31/47.24	33.35/38.96	30.09/40.00	31.23/39.90
MUST	31.71/49.35	35.30/48.68	38.46/58.25	33.41/47.08
SCD	40.70/69.17	52.63/70.21	40.12/71.37	45.03/69.14
S^{3}A (Est-K)	**49.83/76.23**	**57.10**/75.66	**45.54/77.23**	47.86/72.75
S^{3}A (GT-K)	48.34/75.57	56.20/**75.97**	45.21/76.92	**50.41/76.14**

Table 4: Transductive evaluation on four fine-grained benchmarks with estimated cluster numbers (Acc/Cluster). The estimated number is behind the dataset title.

Methods	Caltech (0.34)	CIFAR100 (0.12)	Pet (0.62)
CLIP (Ideal)	91.25/90.96	81.54/81.12	90.87/92.39
CLIP	50.59/49.66	41.62/41.65	55.60/57.96
MUST	51.20/50.80	42.93/42.96	58.32/55.83
SCD	54.08/54.46	42.62/41.64	58.57/57.58
S^{3}A (Our)	**55.29/55.55**	**46.10/46.40**	**59.00/60.57**

Table 5: Tranductive and inductive evaluation on out-of-vocabulary benchmarks (Train/Test Acc). The OOV ratios for each dataset are provided alongside their respective names. Performance is reported by cosine similarity of generic pre-trained Sentence-BERT, upscaled $\times 100$.

S_1], and finally [S_2, $\frac{S_2+S_1}{2}$], each time applying elbow algorithm optimized with Silhouette score (Rousseeuw 1987). Here, S_1 and S_2 denote the solution of the first and second pass. We consistently set $LB_0 = 50$ and $UB_0 = 2000$ for all datasets. Consequently, we obtain minor $(+1.0, +1.2, -0.9)$ train/cluster/test accuracy differences when K is overestimated by 30% w.r.t. the performance with ground-truth K on NL26, which exhibits robustness.

On Out-Of-Vocabulary (OOV) Scenarios. Considering the scenarios in which target datasets have category names out of our **S^{3}A** vocabulary, we further conduct an out-of-vocabulary evaluation on three benchmarks, *i.e.*, Caltech101 (Fei-Fei, Fergus, and Perona 2004), CIFAR100 (Krizhevsky, Hinton et al. 2009), and Oxford-IIIT Pet (Parkhi et al. 2012). The out-of-vocabulary ratios of datasets and results are presented in Table 5. We can conclude that **S^{3}A** still achieves SOTA performance in this challenging setup on both inductive and transductive evaluation.

On Effectiveness of S^{3}A Prompt Augmentation. In this ablation experiment, we analyze the effect of the proposed LLM-guided discriminative prompt augmentation in our CVPR algorithm. We compare with four augmentation setups in Table 6: `(1)` using WordNet definition for augmentation (5^{th} row); `(2)` reduce prompt semantic discriminativeness by requesting visual attributes for only a single category name in each LLM prompt (6^{th} row); `(3)` our prompt augmentation guided by ChatGPT (7^{th} row); `(4)` our prompt augmentation guided by GPT-4. Besides, we also compare with a recent SOTA, CHiLS (Novack et al. 2023), in prompt augmentation for zero-shot prediction. We use their prompt to generate ten subcategories for each class. We

#Row	Methods	IN1K	ET13	ET30
1	CLIP (Ideal)	65.38/96.58	78.43/99.61	77.99/99.58
2	CLIP	32.21/**96.49**	30.09/**97.31**	31.23/**95.83**
3	SCD	37.06/35.09	40.46/32.31	46.29/39.94
4	CHiLS*	36.23/34.46	41.13/33.33	46.09/39.94
5	Our (WordNet)	18.69/18.42	21.82/16.85	20.38/15.94
6	Our (Single)	37.40/35.74	41.13/33.00	47.09/40.76
7	Our (ChatGPT)	37.69/36.11	42.65/36.84	47.43/41.18
8	Our (GPT-4)	**37.95**/36.48	**44.98**/37.56	**48.37**/42.01

Table 6: Ablations on prompt augmentation techniques (Acc/IoU). Performance is reported by cosine similarity of generic pre-trained Sentence-BERT, upscaled $\times 100$.

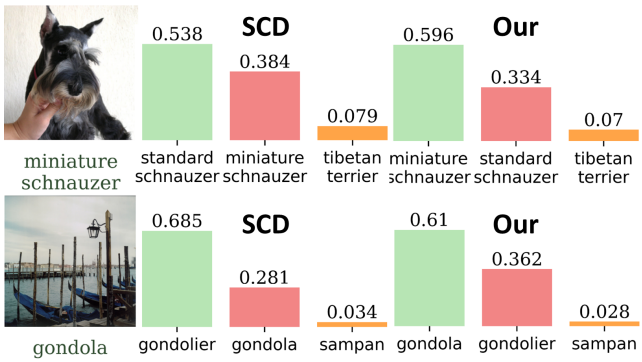

Figure 4: Qualitative results in IN100 without finetuning (SCD (Han et al. 2023) and our CVPR).

can draw the following conclusions: `(1)` Semantic distinctiveness in prompts aids fine-grained differentiation; `(2)` Incorporating WordNet linguistic knowledge hinders semantic discriminativeness; `(3)` Our approach outperforms CHiLS, thus is more tailored to RZSC tasks; `(4)` CLIP focuses on instance alignment and leads to low ACC but high IoU; `(5)` Our method benefits from advanced LLMs.

Qualitative Examples. We present qualitative examples from IN100 in Fig. 4, which demonstrate that our CVPR algorithm can effectively correct category misrecognitions and precisely focus on salient objects.

Conclusion

In this work, we address the challenging task of Realistic Zero-Shot Classification, without assuming partial source supervision or ideal vocabularies. We propose a Self Structural Semantic Alignment (**S^{3}A**) framework, anchored by an innovative Cluster-Vote-Prompt-Realign (CVPR) algorithm for structural semantic relationship mining and a self-training process for iterative semantic alignment. Our experiments demonstrate the effectiveness of **S^{3}A**, consistently achieving significant accuracy improvements over baseline methods on all generic and fine-grained benchmarks, with unknown class numbers, and in out-of-vocabulary scenarios.

References

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2015. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7): 1425–1438.

Arthur, D.; and Vassilvitskii, S. 2007. K-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 1027–1035.

Cappallo, S.; Mensink, T.; and Snoek, C. G. 2016. Video stream retrieval of unseen queries using semantic memory. *arXiv preprint arXiv:1612.06753*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, G.; Yao, W.; Song, X.; Li, X.; Rao, Y.; and Zhang, K. 2022. Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*.

Dale, R. 2021. GPT-3: What's it good for? *Natural Language Engineering*, 27(1): 113–118.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. *2004 Conference on Computer Vision and Pattern Recognition Workshop*, 178–178.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.

Ghiasi, G.; Gu, X.; Cui, Y.; and Lin, T.-Y. 2021. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284.

Han, K.; Li, Y.; Vaze, S.; Li, J.; and Jia, X. 2023. What's in a Name? Beyond Class Indices for Image Recognition. *arXiv preprint arXiv:2304.02364*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Kahana, J.; Cohen, N.; and Hoshen, Y. 2022. Improving Zero-Shot Models with Label Distribution Priors. *arXiv preprint arXiv:2212.00784*.

Karazija, L.; Laina, I.; Vedaldi, A.; and Rupprecht, C. 2023. Diffusion Models for Zero-Shot Open-Vocabulary Segmentation. *arXiv preprint arXiv:2306.09316*.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. MaPLe: Multi-Modal Prompt Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19113–19122.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; and Li, F.-F. 2011. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*. Citeseer.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, 951–958. IEEE.

Lei Ba, J.; Swersky, K.; Fidler, S.; et al. 2015. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE international conference on computer vision*, 4247–4255.

Li, D.; Li, J.; Li, H.; Niebles, J. C.; and Hoi, S. C. H. 2021. Align and Prompt: Video-and-Language Pre-training with Entity Prompts. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4943–4953.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Li, J.; Savarese, S.; and Hoi, S. C. 2022. Masked unsupervised self-training for zero-shot image classification. *arXiv preprint arXiv:2206.02967*.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41.

Novack, Z.; McAuley, J.; Lipton, Z. C.; and Garg, S. 2023. Chils: Zero-shot image classification with hierarchical label sets. In *International Conference on Machine Learning*, 26342–26362. PMLR.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. V. 2012. Cats and dogs. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3498–3505.

Pourpanah, F.; Abdar, M.; Luo, Y.; Zhou, X.; Wang, R.; Lim, C. P.; Wang, X.-Z.; and Wu, Q. J. 2022. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ren, S.; Li, L.; Ren, X.; Zhao, G.; and Sun, X. 2022. Rethinking the Openness of CLIP. *arXiv preprint arXiv:2206.01986*.

Ridnik, T.; Ben-Baruch, E.; Noy, A.; and Zelnik-Manor, L. 2021. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20: 53–65.

Santurkar, S.; Tsipras, D.; and Madry, A. 2020. Breeds: Benchmarks for subpopulation shift. *arXiv preprint arXiv:2008.04859*.

Sariyildiz, M. B.; Kalantidis, Y.; Larlus, D.; and Alahari, K. 2021. Concept generalization in visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9629–9639.

Shin, G.; Albanie, S.; and Xie, W. 2023. Zero-shot Unsupervised Transfer Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4847–4857.

Vaze, S.; Han, K.; Vedaldi, A.; and Zisserman, A. 2022. Generalized Category Discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7492–7501.

Wang, W.; Zheng, V. W.; Yu, H.; and Miao, C. 2019. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2): 1–37.

Wen, X.; Zhao, B.; and Qi, X. 2023. Parametric Classification for Generalized Category Discovery: A Baseline Study. arXiv:2211.11727.

Wu, J.; Li, X.; Yuan, S. X. H.; Ding, H.; Yang, Y.; Li, X.; Zhang, J.; Tong, Y.; Jiang, X.; Ghanem, B.; et al. 2023. Towards Open Vocabulary Learning: A Survey. *arXiv preprint arXiv:2306.15880*.

Xu, J.; Liu, S.; Vahdat, A.; Byeon, W.; Wang, X.; and De Mello, S. 2023. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2955–2966.

Yang, Z.; Gan, Z.; Wang, J.; Hu, X.; Lu, Y.; Liu, Z.; and Wang, L. 2021. An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA. *ArXiv*, abs/2109.05014.

Zang, Y.; Li, W.; Zhou, K.; Huang, C.; and Loy, C. C. 2022. Open-vocabulary detr with conditional matching. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, 106–122. Springer.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2023a. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*.

Zhang, S.; Khan, S.; Shen, Z.; Naseer, M.; Chen, G.; and Khan, F. 2023b. PromptCAL: Contrastive Affinity Learning via Auxiliary Prompts for Generalized Novel Category Discovery. arXiv:2212.05590.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhou, C.; Loy, C. C.; and Dai, B. 2022. Extract free dense labels from clip. In *European Conference on Computer Vision*, 696–712. Springer.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, Z.; Lei, Y.; Zhang, B.; Liu, L.; and Liu, Y. 2023. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11175–11185.