Image-Caption Encoding for Improving Zero-Shot Generalization

Eric Yu [*][†]
UCSD

Christopher Liao [*]
Boston University

Sathvik Ravi [†]
University of Maryland

Theodoros Tsiligkaridis
MIT Lincoln Laboratory

Brian Kulis
Boston University

Abstract

Recent advances in vision-language models have combined contrastive approaches with generative methods to achieve state-of-the-art (SOTA) on downstream inference tasks like zero-shot image classification. However, a persistent issue of these models for image classification is their out-of-distribution (OOD) generalization capabilities. We first show that when an OOD datapoint is misclassified, the correct class can be typically found in the Top-K predicted classes. In order to steer the model prediction toward the correct class within the top predicted classes, we propose the Image-Caption Encoding (ICE) *method, a straightforward approach that directly enforces consistency between the image-conditioned and caption-conditioned predictions at evaluation time only. Intuitively, we take advantage of unique properties of the generated captions to guide our local search for the correct class label within the Top-K predicted classes. We show that our method can be easily combined with other SOTA methods to enhance Top-1 OOD accuracies by* 0.5% *on average and up to* 3% *on challenging datasets. Our code:* `https://github.com/Chris210634/ice`

1. Introduction

There has been rapid progress in zero-shot image classification over the past two years, thanks to advancements in vision-language (VL) pre-training such as CLIP, ALIGN, and BLIP [9,13,20]. At a high level, these models use a pair of encoders that project visual and textual inputs into a joint latent embedding space. As described in the CLIP framework [20], zero-shot classification can be reformulated as an image-to-text retrieval problem, where the class name closest to the image in embedding space is predicted as the label. However, state-of-the-art (SOTA) zero-shot classification lags behind in-distribution supervised fine-tuning on
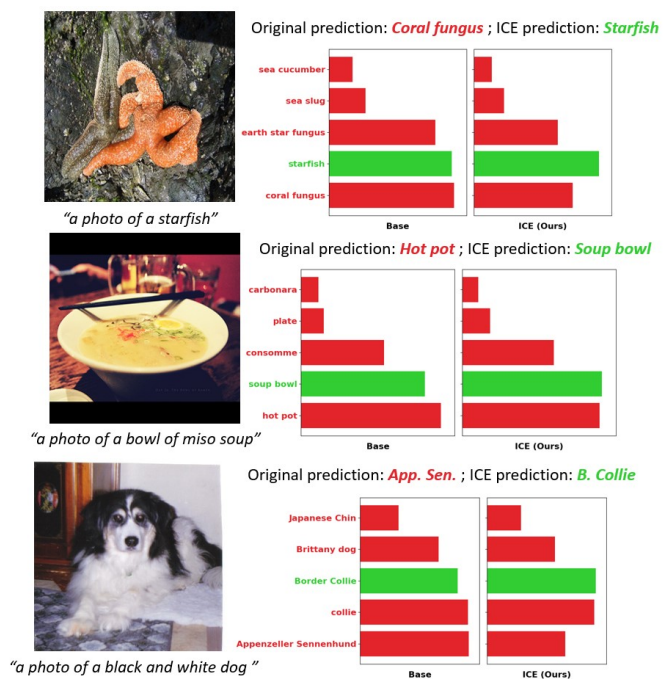

Figure 1. A demonstration for how our ICE method can be used to reclassify correctly. In these examples, ICE is applied directly to a frozen pre-trained CLIP-based model for zero-shot classification. Using the contexts given from the generated captions, ICE is able to successfully influence the pretrained model into predicting the correct classes.

all benchmarks. In many applications, in-distribution data is not available during training, so fine-tuning on out-of-distribution (OOD) source data in a way that generalizes to unseen data and labels remains a challenging problem. Prior works in this area follow two general directions: (1) *Few-shot OOD* methods such as CoOp [34], CoCoOp [33], and MaPLe [10] fine-tune the VL model on generic few-shot source data (e.g. ImageNet). The fine-tuning process is constrained to a carefully selected subset of parameters to ensure generalization to target datasets. (2) *Zero-shot* meth-

---

[*]Equal contribution.
[†]Based on work done at MIT Lincoln Laboratory.

ods, such as [16] and manual prompt ensembling [20], focus on refining the zero-shot prediction without additional fine-tuning. These methods do not require additional data, but they typically either require a large closed-source LLM or human-engineered prompts.
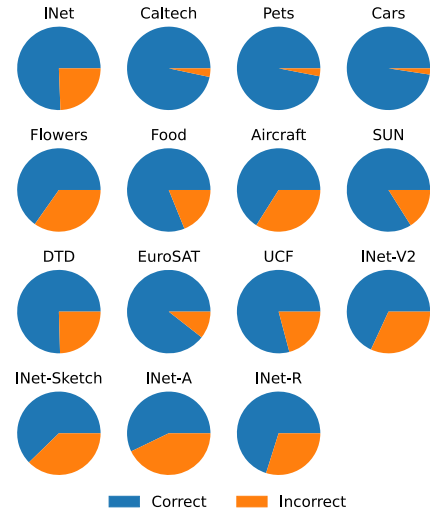
Our goal is improve zero-shot classification by leveraging captioners, which is previously under-explored. Towards this goal, we first observe in Figure 2a that the Top-K accuracy (the percentage of samples where the correct label is within the K classes with highest predicted scores, $K > 1$) is consistently higher than the Top-1 accuracy. The reason is that the Top-K predicted classes form a strict superset of the Top-1 predicted classes when $K > 1$, and thus characterize a wider range of potentially correct classes. We observe from Figure 2a that when an image is misclassified, the correct class can usually be found within the Top-5 predicted classes. Thus, our motivating question is: in order to improve Top-1 accuracy, *how can we steer the model prediction toward the correct prediction within the Top-K predicted classes using generated captions?*

Before we address this question, we first note that current SOTA zero-shot image classification methods perform a nearest-neighbor search between image and text CLIP embeddings [20]. Recently, many works such as CoCa [31], BLIP-2 [12] and LLaVA [14] extend CLIP with an additional text decoder. This text decoder is trained to output a description of the image by cross-attending to *all* image tokens outputted by the image encoder. Consequently, the decoder output captures fine-grained spatial information that may be absent from the image cls token. Furthermore, the caption verbalizes the content of the image as discrete text tokens, which can oftentimes be used to directly infer the image label. These advantages are illustrated in Figure 2b, where we use a spider plot to compare the Top-1 zero-shot accuracies achieved by CLIP image embeddings, CoCa image embeddings, and CoCa caption embeddings, across 15 datasets. We notice that while the caption-only CoCa underperforms compared to standard CoCa, it is still competitive and even surpasses CLIP on many datasets. This suggests that the captions likely contain enough information about the image to supplement the standard zero-shot prediction.

To leverage this additional information, we propose a novel zero-shot method called *Image-Caption Encoding (ICE)*, where we combine the information encoded by both image embeddings and caption embeddings in order to make a more informed decision at test time. As illustrated in Figure 3a, ICE is a zero-shot method with no training component and can be easily paired with existing SOTA methods for improved downstream classification performance.

Although ICE draws inspiration from traditional ensembling techniques, there are several key differences. First, ICE leverages the predictions obtained from a *single* model rather than those from several *different* models. Second, in-

(a) A visualization of the Top-5 accuracies on misclassified Top-1 datapoints in each test dataset. Recall that correct Top-5 classifications form a strict superset over the correct Top-1 classifications. We observe that across all datasets, the true correct class can be found within the Top-5 predicted classes for most misclassified datapoints.
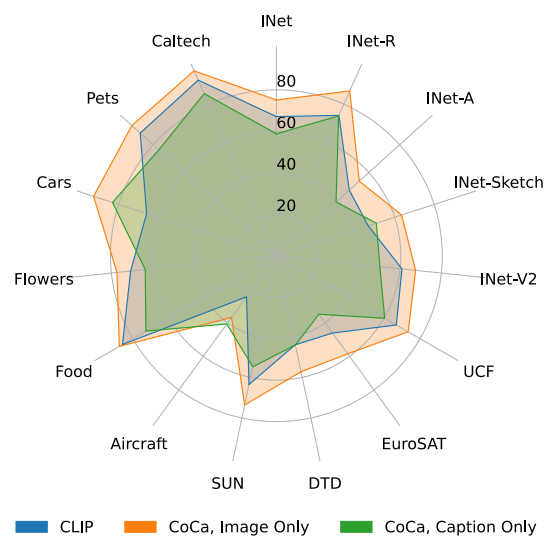

(b) A visualization of Top-1 accuracies between CLIP, CoCa using image embeddings only, and CoCa using caption embeddings only. We observe that while caption embeddings generally underperform compared to standard CoCa, they still retain competitive performance. We include more details on datasets and experiments in Section 4.

Figure 2. Quantitative motivation and results visualization.

stead of aggregating predictions over all classes, we only consider the Top-K image predicted classes. Third, we incorporate a lightweight confidence selection mechanism that sets the weight on the caption prediction dynamically. Finally and most importantly, we exploit specific properties induced within captions that are not present in the image
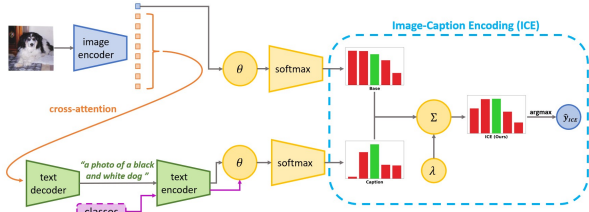
(a) An overview of our *Image-Caption Encoding (ICE)* method. Here, we query a captioner and obtain the caption embedding using the text encoder. We calculate the image and caption probability distributions over the classes by passing the image embeddings, caption embeddings, and class embeddings through the cosine similarity function θ and softmax operation. Then, we select the Top-K classes and perform a weighted sum of the image and caption probabilities. The weight on the caption prediction λ is adaptively selected based on the relative confidence of the image and caption predictions.
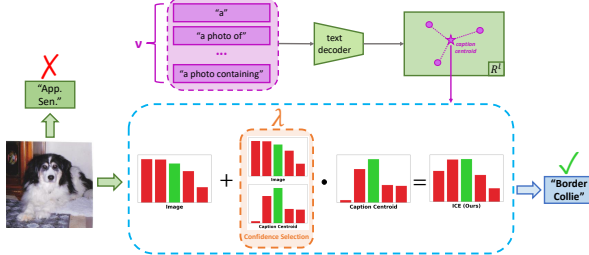
(b) A more detailed look at how our *Image-Caption Encoding (ICE)* method works. In practice, instead of using a single caption for ICE, we use the centroid of υ differently-prompted caption embeddings. Then, using the centroid caption embedding, we adaptively select the λ weight by comparing the standard deviations of the image prediction probabilities and caption prediction probabilities, over the Top-5 classes. The final ICE scores are then a λ-weighted sum between the two probability distributions.

Figure 3. Illustrations of our ICE method.

embeddings for standard zero-shot classification. We discuss these properties in-depth in Section 3.4 and demonstrate specific examples of their non-trivial impact in Section 4.3.

Our contributions are as follows:

1. We extend the zero-shot classification literature by leveraging captioners, which is a direction that is previously under-explored.

2. We propose *Image-Caption Encoding (ICE)*, a novel zero-shot classification method that utilizes information from both images and captions to make a decision at evaluation-time only.

3. We provide experimental results for ICE paired with several different SOTA baselines across 15 different OOD datasets. We show consistent improvements of 0.5% on average and up to 3% on several challenging datasets.

4. We analyze the benefits and drawbacks of using ICE, and provide ablation studies to analyze the effects of changing different parameters in our ICE method.

2. Related Works

Multimodal foundational models. Many VL foundational models have emerged over the past two years, including CLIP [20], ALIGN [9], BLIP [13], CoCa [31], BLIP-2 [12], and LLaVA [14]. These models achieve SOTA on VL tasks by using vast quantities of un-curated image-text data from the web. CLIP uses an image encoder and a text encoder to project the two modalities into a joint latent embedding space. Popular downstream applications include zero or few-shot classification and image-text retrieval. The more recent models such as CoCa, BLIP-2 and LLaVA improve CLIP by additionally training a text decoder to explain the embedding space with a caption. In the current work, we leverage this additional captioning capability to improve the CLIP zero-shot accuracy.

Robust fine-tuning. There is growing interest in fine-tuning multimodal foundational models on limited training data such that the resulting model achieves high accuracy even on domain-shifted data and data with labels not seen during training. Many modern approaches rely on prompt tuning and ensembling. CoOp [34] is a seminal work which treats the prompt preceding the label names as soft learnable tokens. CoCoOp [33] trains a meta-network to condition the prompt tokens on the image embedding. MaPLe [10] shows that learning a conditional visual prompt jointly with textual prompts improves target accuracy. All three works achieve impressive results on a diverse set of test datasets despite only being trained on few-shot ImageNet data. ClipOOD [24] uses an adaptive margin loss to optimize the visual encoder only, attaining good results on domain generalization benchmarks. Our proposed method ICE is a training-free approach that can be readily combined with the above fine-tuning methods to yield higher accuracy on most target datasets.

Ensembling for robust classification. Ensembling methods leverage multiple diverse predictions to form a robust final prediction; many recent works, including ours, focus on discovering new sources of diversification. WiSE-FT [29] calculates a weight space ensemble of the fine-tuned and pre-trained models to increase robustness under distribution shifts in target data, while inference time remains the same. [27] demonstrate that combining cross-entropy fine-tuning with stochastic weight averaging improves domain generalization. [16] use GPT descriptions to generate a more diverse set of text prototypes for zero-shot classification. [20] use an ensemble of 80 handcrafted prompts to achieve the same goal. In our paper, while drawing inspiration from ensembling, we introduce the novel use of captions as a unique source of diversification, an avenue not yet explored by previous studies. Importantly, our approach is designed to seamlessly integrate with other zero-shot methods, as demonstrated in our experimental results. We emphasize that, contrary to ensembling methods which use a

separately trained LLM [16, 32], our method uses the CoCa model [31], which trains a pair of aligned image and text encoders jointly with a text decoder on the same dataset. Under this constraint, our method does not use more pre-training data than baselines.

3. Methodology

3.1. Preliminaries

Consider a dataset $\mathcal{D} \subset \mathcal{I} \times \mathcal{T}$ where $\mathcal{I}$ is the image domain and $\mathcal{T}$ is the text domain, and (I_i, T_i) forms a corresponding image-text pair (i.e. T_i is a caption that describes I_i). In the CLIP framework [20], there are two main architectural components: an image encoder $f_{\mathcal{I}} : \mathcal{I} \to \mathbb{R}^l$ that maps images to a shared latent space, and a text encoder $f_{\mathcal{T}} : \mathcal{T} \to \mathbb{R}^l$ that maps text to the same shared latent space. Both encoders are pre-trained using a contrastive loss that pulls corresponding image-text embeddings close together in latent space, and pushes non-corresponding image-text embeddings away from each other. In our framework, we require an additional text decoder, pre-trained using a next-token-prediction loss, to provide captioning functionality.

In image classification following the CLIP framework [20], for an image vector I and class labels vector $y := [y_1, y_2, \ldots, y_m]^{\mathsf{T}}$, we first feed each class label through a prompt skeleton to obtain class prompts (e.g. for class "cat" and a prompt skeleton "A photo of a {}", the resulting prompt is "A photo of a cat"). Then, both the image and class prompt vectors pass through the image and text encoders to obtain latent embeddings $\widetilde{I}$ and $\widetilde{y}$, respectively. The predicted class label $\hat{y}$ is then given by $\arg\max_i \theta(\widetilde{I}, \widetilde{y}_i), i \in \{1, 2, \ldots, m\}$, where $\theta : \mathbb{R}^l \times \mathbb{R}^l \to \mathbb{R}$ is the cosine similarity function.

3.2. Image-Caption Encoding

Consider a text decoder $f_{\phi} : \mathcal{P} \times \mathcal{I} \to \mathcal{T}$ that maps a prompt $p \in \mathcal{P}$ (e.g. "a photo of") and an image I to a caption $c \in \mathcal{T}$. We can feed caption c back through the text encoder to obtain $\widetilde{c} := f_{\mathcal{T}}(c)$.

Using the Softmax function [2], we obtain the class probabilities for image I and caption c, respectively, as

$$S^I := \mathrm{Softmax}\left(\left[\theta(\widetilde{I}, \widetilde{y}_1), \theta(\widetilde{I}, \widetilde{y}_2), \ldots, \theta(\widetilde{I}, \widetilde{y}_m) \right]^{\mathsf{T}} \right)$$
$$S^c := \mathrm{Softmax}\left(\left[\theta(\widetilde{c}, \widetilde{y}_1), \theta(\widetilde{c}, \widetilde{y}_2), \ldots, \theta(\widetilde{c}, \widetilde{y}_m) \right]^{\mathsf{T}} \right).$$
$$(1)$$

The indices corresponding to the K classes with highest image-predicted probability, and the final ICE prediction, are computed respectively as:

$$\Omega_K^I := \underset{J \subset M, |J|=K}{\arg\max} \sum_{j \in J} S_j^I \quad \underset{\omega \in \Omega_K^I}{\arg\max} \, S_\omega^I + \lambda S_\omega^c \quad (2)$$

where $M = \{1, 2, \ldots, m\}$ and λ is a scalar variable. In essence, the ICE prediction remains anchored within the primary Top-K classifications as determined by the image class probabilities. When we incorporate the caption scores tied to these Top-K predictions, it reshapes the probability landscape over the initial Top-K image-determined classes. By selecting the class with the highest probability from this refined distribution, we aim to align closer with the true class while retaining the accuracy of previous classifications. Intuitively, the caption probability distribution should provide information about the image that is not clear or fully captured by the image probability distribution, as detailed in Section 3.4, and their aggregated prediction should provide a more reasonable downstream prediction. Figures 3a and 3b contain a high-level overview and an in-depth visual interpretation of our method, respectively.

3.3. Additional Modifications

As visualized in Figure 3b, in our experiments, we use the centroid of a diverse set of captions rather than a single caption for increased robustness. That is, for a set of prompts $P \in \mathcal{P}^v$, we generate a set of corresponding captions $C \in \mathcal{T}^v$, obtain their caption embeddings $\widetilde{C} := \{f_{\mathcal{T}}(c_1), f_{\mathcal{T}}(c_2), \ldots f_{\mathcal{T}}(c_v)\}$, and finally, their centroid $\widetilde{\overline{c}} := \frac{1}{v} \sum_i^v \widetilde{c}_i$. The centroid $\widetilde{\overline{c}}$ is then used in place of c in Equations 1 and 2.

In addition, we dynamically compute the caption scalar variable $\lambda \in \mathbb{R}_+$ as a function of the standard deviation of the captions. That is, given some image I and caption c, we compute λ as

$$\lambda = \xi \frac{\sigma(S_K^c)}{\max(||[\sigma(S_K^I), \sigma(S_K^c)]||_2, \epsilon)} \quad (3)$$

where ξ is a constant, ϵ is a small constant, S_K^I and S_K^c are the Top-K probabilities for S^I and S^c respectively (i.e. the probabilities whose indices are specified by Ω_K^I), and σ is the standard deviation operator. Intuitively, the standard deviation of the Top-K highest probabilities of the image and caption distributions correspond to the model confidence about their respective predictions. On one hand, when the model confidence for the image prediction is high and the caption confidence is low, then the caption probabilities should not influence the image probabilities as much. On the other hand, when the image confidence is low but the caption confidence is high, the caption probabilities should more heavily influence the image probabilities. In the event when both image and caption confidences are high or low, the default weighting would be relatively equal. The constant ξ can specify how much the caption probabilities should affect the image probabilities overall.

3.4. Caption Properties

In general, the zero-shot accuracy obtained using only the caption embeddings is significantly (on average about 5 %, see Table 1) lower than the zero-shot accuracy obtained using image embeddings, with the notable exception

of aircraft fine-grained classification. Caption-only zero-shot classification is often unreliable, since the caption does not always correspond to one of the label choices. For example, a picture containing a teddy bear on top of a bed might be captioned as "a picture of a teddy bear", ignoring the bed in the background. However, if teddy-bear is not one of the labels, and the correct label is "bed", the caption does not provide useful information for the classification problem. For this reason, the optimal hyperparameters for ICE place a greater emphasis on the prediction from the image embedding $\widetilde{I}$. Nonetheless, for the caption embedding to contribute to a higher aggregate accuracy, we only require that the caption-predicted probabilities S_K^c be not completely correlated with S_K^I. In other words, *the caption sometimes contains extra information that nudges the prediction in the correct direction*. We list here a few intuitions for why using captions can improve overall classification accuracy:

1. The text decoder cross-attends to all output image tokens from the vision encoder, while the image prediction only uses the output cls token. The image token matrix contains spatial information that may be pertinent to the target task. Specifically for the CoCa model we use, every other transformer layer in the text decoder stack attends to the full spatial grid of output image tokens.

2. The text decoder was pretrained to caption images with a language-modeling loss. Consequently, it exhibits some rudimentary reasoning ability based on learning relationships between certain concepts. For example, the text decoder has learned that the painting "the starry night" is authored by Vincent van Gogh. This correspondence is learnt by the weights of the decoder and may be useful for some classification tasks. In our experiments, we found that the caption prediction is much better than the image prediction on aircraft classification. This is likely because the correspondence between fine-grained visual concepts and the aircraft model name is learned by the text decoder. This intuition is experimentally confirmed in Table 6 of the Appendix, where textual cues from the image caption correctly steer the prediction on 5.9 % of aircraft images.

3. The caption effectively isolates visual concepts; this is an inherent property of textual data. For example, a caption that reads "a rough red blanket" effectively isolates the texture, color and content of the image. In our experiments, we found that captions on the EuroSAT dataset often isolate the land-use information from the geographical information, e.g. "a photo of agricultural land in China". The caption explicitly separates useful information (agricultural land) from information that is irrelevant to the classification problem (China). Consequently, a classifier trained on captions is less likely to learn domain-specific spurious correlations, especially in the few-shot setting. This intuition is experimentally confirmed in Table 6 of the Appendix, where

textual cues from the image caption correctly steer the prediction on 3.4 % of satellite images.

We provide concrete examples of these discussed intuitions in our empirical analysis in Section 4.3.

4. Experiments

In our experiments in Table 1, we analyze the impact of combining ICE with different zero-shot (ZS) baselines across a suite of benchmarks. We show that our method consistently improves ZS baselines without requiring additional training. In addition, we analyze several data points to show how ICE improves over the base method that it is paired with. All implementation details can be found in Appendix A. To demonstrate the general applicability of our ICE method, we also present results using three different recent multimodal models that contain a decoder component: CoCa [31], BLIP-2 [12], and LLaVA [14]. These results are included in Table 3. For the CoCa setup, all model components are pretrained using the LAION dataset [23], similar to baselines. We emphasize that the captioner/text decoder component does not see more data than the contrastively trained text and image encoders. This is important for fair comparison with prior work. In the Appendix, we show that the LAION dataset is not contaminated with the test images by showing nearest neighbors to select test samples. BLIP-2 and LLaVA are VLMs instruction-tuned with additional supervised image-text data; we included results on these larger models for readers who might be less familiar with CoCa.

Datasets. In line with prior work, our datasets are split into two categories: *cross-dataset evaluation* and *domain generalization*. For cross-dataset generalization, each evaluation dataset has mostly non-overlapping classes and unrelated data distributions for zero-shot classification. For domain generalization, the evaluation datasets are domain-shifted variations of the ImageNet dataset and share the same classes as ImageNet. We evaluate our method on 11 cross-dataset generalization datasets covering a wide range of image recognition tasks. These include two generic objects datasets, ImageNet [22] and Caltech101 [4]; five fine-grained datasets, OxfordPets [18], StanfordCars [11], Flowers102 [17]), Food101 [1], and FGVCAircraft [15]; a scene categorization dataset, SUN397 [30]; an action recognition dataset, UCF101 [25]; a describable textures dataset, DTD [3], and a satellite images dataset, EuroSAT [5]. In addition, we consider four domain generalization datasets, each applying a different distribution shift to the source ImageNet dataset. These include an extension of the ImageNet dataset, ImageNetV2 [21]; a black and white sketches dataset, ImageNet-Sketch [28], a naturally adversarial dataset, ImageNet-A [8], and a dataset containing different renditions (e.g. cartoons, graffiti, plush objects, etc.) of the ImageNet classes, ImageNet-R [6].

4.1. Zero-Shot Classification

Baselines. We consider four existing SOTA methods as zero-shot classification baselines for ICE: (1) a pre-trained CoCa model [31] with class embeddings generated using the prompt "a photo of a {}", where "{}"is replaced by the corresponding class (2) the pre-trained model with class embeddings calculated from the centroids of 80 intermediate embeddings generated using hand-crafted prompts from [20] (3) the pre-trained model using large language model (LLM) generated descriptors from [16] (4) a variation of the previous baseline, where we take the centroid of each descriptor embedding for standard zero-shot classification.

Results. In Table 1, we observe that stacking ICE achieves consistent improvements of around 0.5% on average across cross-dataset evaluation and domain-generalization evaluation benchmarks, with improvements of up to 3% on datasets like FGVCAircraft and EuroSAT. We emphasize that these ZS accuracy improvements are consistent across all datasets (Table 1) and three different pretrained model architectures (CoCa, BLIP-2 and LLaVA in Table 3).

4.2. Few-Shot Classification

Although our ICE method is a zero-shot method, it can be trivially extended to the few-shot learning setting by applying it to the few-shot finetuned model. Accordingly, we consider SOTA methods where a pre-trained model is fine-tuned on 16-shot ImageNet training data. Specifically, the ImageNet training dataset contains 1000 classes with 16 images per class, for a total of 16,000 images in the train dataset. These few-shot results are included in Table 2.

Baselines. We consider two prompt learning methods, CoOp [34] and MaPLe [10], and a fine-tuning method CLIPood [24] as our ICE baselines. We use each method for few-shot fine-tuning on the CLIP components of the CoCa [31] architecture, and perform standard zero-shot classification by following the CLIP framework [20].

Results. As seen in Table 2, applying ICE to each baseline provides improvements of 0.5% on average across cross-dataset generalization evaluation datasets, and smaller improvements for the domain generalization datasets. For the methods such as CLIPood where we do not see improvements in domain generalization on average, we find that ICE at least maintains the average performance of the classification backbone.

4.3. Understanding ICE Improvements

To better understand why ICE improves the base methods paired with it, we analyze examples in Figure 4 where ICE correctly reclassifies a previously incorrectly classified datapoint, where ICE preserves a previously correct classification, where ICE fails to correctly reclassify a previously
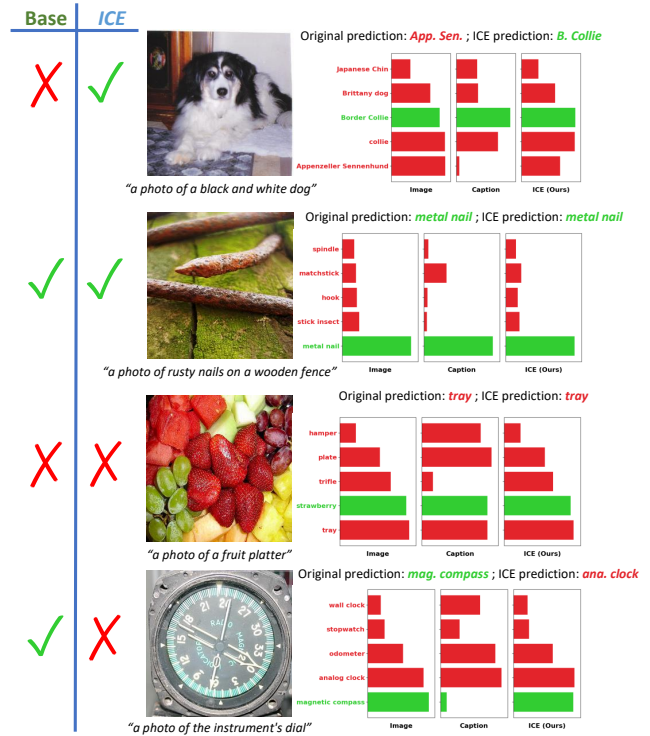

Figure 4. A qualitative analysis on various ways ICE can affect the downstream classification performance.

incorrect classification, and where ICE accidentally incorrectly reclassifies a previously correct classification.

In the first example of Figure 4, we observe that ICE reclassifies the incorrectly predicted *Appenzeller Sennenhund* class to the correct *Border Collie* class. One intuition for why this occurs is that Border Collies most commonly have black and white fur, whereas Appenzeller Sennenhund dogs typically have unique tricolor fur coats. Thus, the caption "a photo of a black and white dog" would correspond more with the Border Collie purely based on its bicolor fur. Additional examples of correct ICE reclassifications can be found in Figure 1.

In the second example, we observe that ICE is able to successfully preserve an originally correct classification. Here, since the caption agrees with the image predicted class, ICE is able to predict the same class as before.

In the third image, ICE fails to correctly reclassify an initially incorrect prediction. In this case, the original prediction was a *tray*, which would make sense since this image can technically be described as a "tray of fruits". Similarly, the captions describe the image as "a photo of a fruit platter", which can correspond more to a *plate* or *tray* than a *strawberry*. Thus, the ICE predicted class still matches that of the original class.

Finally, the fourth image is an example for how ICE can accidentally incorrectly reclassify an initially correct pre-

		INet	Cross-dataset Evaluation Targets											Domain Generalization Targets				
			Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN	DTD	EuroSAT	UCF	Average	INet-V2	INet-Sketch	INet-A	INet-R	Average
Zero-shot(Image)		75.1	**97.6**	**93.8**	92.7	77.3	87.5	36.8	73.6	57.2	58.5	73.4	74.8	67.5	63.5	53.8	87.0	68.0
Zero-shot (Caption)		58.8	85.6	76.3	83.1	63.6	72.7	**40.7**	54.8	44.0	34.9	60.3	61.6	50.2	50.7	38.7	73.8	53.3
+ ICE (ours)		**75.6**	97.1	**93.8**	93.0	78.0	87.7	38.3	**74.0**	59.3	61.3	74.3	75.7	67.8	64.0	54.4	87.5	68.4
Manual Prompts [20]	ICML '21	75.5	97.1	93.7	92.7	77.5	87.5	37.5	74.0	60.7	60.2	73.8	75.5	67.8	**64.7**	53.2	88.1	68.4
+ ICE (ours)		**75.9**	97.1	93.8	93.1	78.7	87.6	39.7	74.1	61.6	61.2	73.9	76.1	68.2	64.7	54.4	88.4	68.9
GPT Centroids [16]	ICLR '23	74.8	**97.8**	93.3	92.4	75.8	87.4	36.4	73.9	58.8	63.7	73.2	75.3	67.3	63.3	52.6	86.7	67.5
+ ICE (ours)		**75.2**	97.4	93.0	92.8	76.3	87.7	39.1	74.2	60.2	64.2	73.7	75.9	67.4	63.5	53.5	87.1	67.9
GPT Score Mean [16]	ICLR '23	74.9	**97.6**	**93.7**	92.4	76.2	87.3	36.2	73.9	58.9	64.9	73.6	75.5	67.6	63.5	52.8	86.8	67.7
+ ICE (ours)		**75.4**	97.4	93.5	92.8	77.0	87.6	39.2	74.2	60.2	65.5	73.9	76.1	67.9	63.7	53.4	87.2	68.1
CuPL [19]	ICCV '23	75.6	-	93.0	92.1	77.0	87.1	39.9	74.2	67.1	-	76.0	-	-	-	-	-	-
+ ICE (ours)		**75.9**	-	93.0	92.6	78.4	87.2	42.2	74.4	67.4	-	76.1	-	-	-	-	-	-

Table 1. Comparison with zero-shot baselines on 15 test datasets. We always observe that stacking our ICE method on top of baseline methods provides consistent improvements. All methods are zero-shot and use CoCa ViT-L/14. The caption zero-shot accuracy is reported using one caption embedding prompted by "a photo of". ImageNet is abbreviated INet.
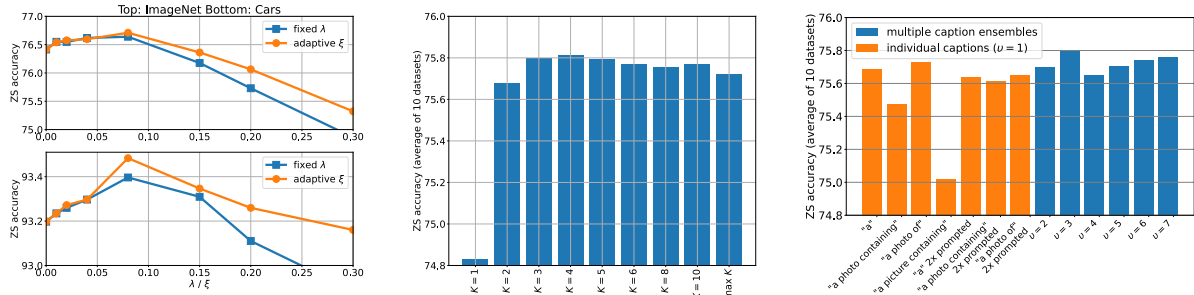

Figure 5. Left: Ablation results on varying ξ. Comparison between fixed caption weight λ and adaptive λ using Eq. 3. Middle: Ablation results on varying caption prompting and number of captions in ensemble (υ). $\upsilon = 3$ captions is optimal. Right: Ablation results on varying K. $K = 4$ is optimal and clearly superior than bypassing Top-K selection step (denoted as max K in bar plot).

diction. Here, the correct class is a *magnetic compass*, as evident by the blue text written on the instrument. While the caption "a photo of the instrument's dial" technically describes the image, without additional visual context, it is easy to believe that the caption is describing a clock rather than a magnetic compass since dials tend to be associated with clocks. Due to this ambiguity, ICE was able to convince the initial predicted class to reclassify to the runner-up class of *analog clock*. This example highlights the importance of why we need to consider both the image and caption information for making the most informed decisions.

5. Ablation Studies

A comprehensive ablation study on the parameters of ICE is presented in Figure 5. First, we evaluate the contribution of the adaptive λ mechanism proposed in Eq. 3 on the left of the figure. On ImageNet and cars, the zero-shot accuracy of the adaptive λ is clearly superior to the fixed λ for varying values of ξ (in the adaptive case) and λ (in the fixed case). In the middle bar plot, we examine the contribution of the Top-K selection procedure outlined in Eq. 2. We compare multiple values of K (where $K = 1$ is equiva-

lent to ignoring caption embeddings, and "max K" denotes score averaging over all classes without Top-K selection). Values of K around $K = 4$ are clearly superior to both ignoring caption embeddings and score averaging without Top-K selection. Finally, in the bar plot on the right, we examine the contribution of ensembling multiple captions. Using $\upsilon = 3$ captions is approximately optimal. Note that the 7 individual captions exhibit high variance in zero-shot accuracy (under the ICE framework); this variance in performance is greatly reduced by ensembling a small number of captions.

6. Limitations

While our results show consistent improvements when evaluating on several different baseline methods across a diverse collection of datasets, we note several limitations of our method. First, as elaborated in Section 3.4, ICE heavily relies on the assumption that the captions can provide useful information about the image that is not fully encoded by the image embeddings. As seen in Section 4.3, when the captions provide unhelpful or adversarial infor-

	Source	Cross-dataset Evaluation Targets											Domain Generalization Targets				
	INet	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN	DTD	EuroSAT	UCF	Average	INet-V2	INet-Sketch	INet-A	INet-R	Average
CLIPood	76.6	**97.2**	**94.3**	92.7	77.5	87.3	37.4	**74.3**	60.3	59.5	75.2	75.6	**69.5**	64.9	56.6	88.7	**69.9**
+ ICE	**76.7**	97.2	93.9	**93.1**	77.8	87.4	39.6	74.3	60.8	61.8	75.7	76.2	69.2	64.9	56.7	88.8	69.9
CoOp	76.4	**97.4**	**93.9**	93.2	77.0	**87.6**	39.0	73.4	59.2	61.2	74.7	75.7	**69.0**	63.3	55.2	87.7	68.8
+ ICE	**76.7**	97.2	93.8	**93.5**	77.7	87.6	40.7	73.6	60.6	62.0	75.1	76.2	69.0	63.4	55.9	88.1	69.1
MaPLe	77.3	**96.7**	**94.2**	92.8	76.9	**87.2**	39.5	73.9	61.1	58.7	76.0	75.7	**70.2**	64.7	54.8	88.2	69.5
+ ICE	**77.5**	96.6	94.1	**93.0**	77.1	87.2	41.4	74.3	61.3	59.6	76.8	76.1	70.1	64.6	55.1	88.6	69.6

Table 2. Comparison with *few-shot baselines* in the *cross-dataset evaluation* setting and the *domain generalization* setting. The model is fine-tuned on ImageNet with three different methods and tested on 15 total target datasets. The average accuracies are calculated separately for the two settings following prior work. In all cases, we observe that evaluating with our ICE method provides consistent improvements.

		Cross-dataset Evaluation Targets											Domain Generalization Targets				
	INet	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN	DTD	EuroSAT	UCF	Average	INet-V2	INet-Sketch	INet-A	INet-R	Average
CoCa Zero-shot (Image)	75.1	**97.6**	**93.8**	92.7	77.3	87.5	36.8	73.6	57.2	58.5	73.4	74.8	67.5	63.5	53.8	87.0	68.0
CoCa Zero-shot (Caption)	58.8	85.6	76.3	83.1	63.6	72.7	**40.7**	54.8	44.0	34.9	60.3	61.6	50.2	50.7	38.7	73.8	53.3
+ ICE	**75.6**	97.1	**93.8**	**93.0**	**78.0**	**87.7**	38.3	**74.0**	**59.3**	**61.3**	**74.3**	**75.7**	**67.8**	**64.0**	**54.4**	**87.5**	**68.4**
BLIP-2 Zero-shot (Image)	73.8	94.6	**93.6**	76.9	**79.4**	**90.9**	**32.8**	68.0	52.7	56.2	74.7	72.0	68.0	57.9	68.3	85.5	69.9
BLIP-2 Zero-shot (Caption)	44.8	87.3	29.8	51.3	49.9	56.2	7.5	50.0	46.3	61.9	61.9	48.0	42.2	47.0	42.5	75.6	51.8
+ ICE	**74.8**	**95.5**	93.1	**78.6**	79.4	90.4	32.6	**70.3**	**54.7**	56.3	**75.4**	**72.6**	**68.8**	**60.0**	**69.6**	**88.0**	**71.6**
LLaVA Zero-shot (Image)	73.8	94.6	**93.6**	76.9	79.4	90.9	**32.8**	68.0	52.7	56.2	**74.7**	72.0	68.0	57.9	68.3	85.5	69.9
LLaVA Zero-shot (Caption)	42.4	85.4	19.8	17.8	16.8	43.8	4.9	49.0	41.5	52.7	56.8	38.8	37.5	38.2	38.7	66.8	45.3
+ ICE	**74.1**	**95.1**	93.0	76.0	78.2	90.6	32.3	**69.0**	**53.1**	57.4	74.6	71.9	**68.3**	**58.4**	**69.0**	86.5	**70.5**

Table 3. Experiments using a variety of multimodal captioners, showing that using ICE to combine the image and caption embeddings improves the overall zero-shot accuracy. CoCa ViT-L/14 uses its own pretrained image encoder, while BLIP-2 and LLaVA use the frozen CLIP ViT-L/14 model from [20].

mation and there lacks a good selection of caption scores weight λ, ICE could decrease the base classification performance. Second, determining a good choice for λ is non-trivial, as it requires the selector to have an understanding for when to trust the caption or image scores more on a per-datapoint basis. This task is comparable with relevant literature on failure mode prediction using confidence estimation [7, 26, 35], which is known to be challenging. Finally, generating captions can be expensive, since each additionally generated caption requires an additional forward pass on the base model. For image-text foundation models like CoCa [31], which typically contain hundreds of millions of parameters, this strategy can quickly become a time bottleneck for ICE. Thus, it is important to find a balance between the robustness benefits reaped from the number of captions used, and the linearly-increasing time costs for each additional caption generated.

7. Conclusion

We explored the use of captioners which is understudied in the vision-language literature for classification. We proposed a novel method for improved zero-shot classification performance based on combining image and caption embeddings. We showed that our method can be easily paired with existing SOTA methods, and provide improvements of 0.5% on average and up to 3% across a diverse array of datasets in cross-dataset generalization and domain generalization. Several ablation studies are presented to study sensitivity of parameters on performance. We performed an in-depth analysis on why our method can help reclassify previously misclassified points, and also cover cases where it might fail. Future work includes extending this work to better balance the weights between image and caption scores, and considering ways to generate more informative captions for improved downstream classification.

Acknowledgments

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 5

[2] John Bridle. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989. 4

[3] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. *CoRR*, abs/1311.3618, 2013. 5

[4] Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178, 2004. 5

[5] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *CoRR*, abs/1709.00029, 2017. 5

[6] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *CoRR*, abs/2006.16241, 2020. 5

[7] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *CoRR*, abs/1610.02136, 2016. 8

[8] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CoRR*, abs/1907.07174, 2019. 5

[9] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1, 3

[10] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 1, 3, 6

[11] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 5

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 3, 5

[13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1, 3

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2, 3, 5

[15] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013. 5

[16] Sachit Menon and Carl Vondrick. Visual classification via description from large language models, 2022. 2, 3, 4, 6, 7

[17] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 722–729, 2008. 5

[18] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 5

[19] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023. 7

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 2, 3, 4, 6, 7, 8

[21] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? *CoRR*, abs/1902.10811, 2019. 5

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. 5

[23] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 5, 12

[24] Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. Clipood: Generalizing clip to out-of-distributions, 2023. 3, 6

[25] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 5

[26] Theodoros Tsiligkaridis. Failure prediction by confidence estimation of uncertainty-aware dirichlet networks. *CoRR*, abs/2010.09865, 2020. 8

[27] Kevin Vogt-Lowell, Noah Lee, Theodoros Tsiligkaridis, and Marc Vaillant. Robust fine-tuning of vision-language models for domain generalization. In *IEEE High Performance Extreme Computing Conference (HPEC)*, 2023. 3

[28] Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary C. Lipton. Learning robust global representations by penalizing local predictive power. *CoRR*, abs/1905.13549, 2019. 5

[29] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. 3

[30] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010. 5

[31] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. 2, 3, 4, 5, 6, 8

[32] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15211–15222, 2023. 4

[33] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 1, 3

[34] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, jul 2022. 1, 3, 6

[35] Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. Rethinking confidence calibration for failure prediction, 2023. 8