AAPL: Adding Attributes to Prompt Learning for Vision-Language Models

Gahyeon Kim* Sohee Kim* Seokju Lee†
Korea Institute of Energy Technology (KENTECH)
{gahyeon, soheekim, slee}@kentech.ac.kr

Abstract

Recent advances in large pre-trained vision-language models have demonstrated remarkable performance on zero-shot downstream tasks. Building upon this, recent studies, such as CoOp and CoCoOp, have proposed the use of prompt learning, where context within a prompt is replaced with learnable vectors, leading to significant improvements over manually crafted prompts. However, the performance improvement for unseen classes is still marginal, and to tackle this problem, data augmentation has been frequently used in traditional zero-shot learning techniques. Through our experiments, we have identified important issues in CoOp and CoCoOp: the context learned through traditional image augmentation is biased toward seen classes, negatively impacting generalization to unseen classes. To address this problem, we propose adversarial token embedding to disentangle low-level visual augmentation features from high-level class information when inducing bias in learnable prompts. Through our novel mechanism called "Adding Attributes to Prompt Learning", AAPL, we guide the learnable context to effectively extract text features by focusing on high-level features for unseen classes. We have conducted experiments across 11 datasets, and overall, AAPL shows favorable performances compared to the existing methods in few-shot learning, zero-shot learning, cross-dataset, and domain generalization tasks.

1. Introduction

Recent research has shown significant improvements not only in model generalization performance through the use of large-scale vision-language models (VLMs), but also in zero-shot image classification performance [46, 61, 63, 64, 66]. It has been demonstrated that utilizing VLMs such as contrastive language-image pretraining (CLIP) [39], ALIGN [17], Flamingo [1], etc., is effective in extracting image and text information for training classification models. The strengths of these VLMs have proven to be effective in prompt learning and handling both visual and textual information efficiently [31, 45, 67, 68]. CoOp [68] and CoCoOp [67] have effectively produced learnable context vectors for classification weights via a text encoder (*e.g.*, Transformer [50]) along with CLIP. Specifically, CoCoOp [67] has enabled the creation of class-specific classification weights by incorporating additional context information generated from images. In addition, visual prompt tuning (VPT) [18] demonstrated performance improvements in downstream tasks by introducing a small number of learnable parameters into the encoder layer of the Transformer along with image patches, without the need to replace or fine-tune the pretrained transformer.

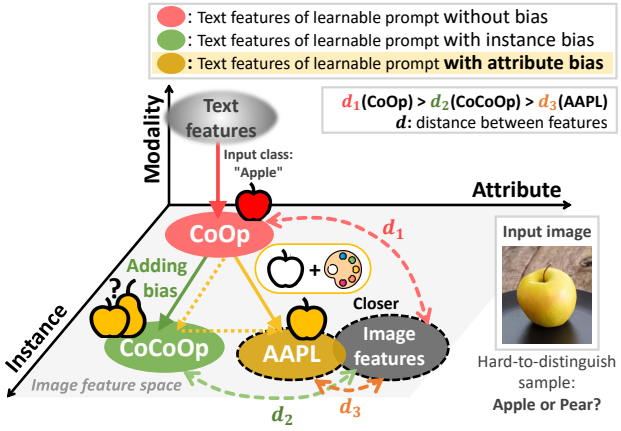


Figure 1. **The illustration of AAPL.** Training the learnable prompt on the class "apple", since the training data mainly consists of red apples, leads to understanding apples as typically red. When a rare "yellow apple" is input, the instance bias may overlook the yellow attribute and incorrectly predict it as a pear. However, AAPL extracts and decomposes attributes from the image, enhancing attribute-specific bias in the semantic features. This enables robustly improved generalization performance across domains.

However, both CoOp [68] and VPT [18] have learnable parameters that are not manageable, especially in the case of CoCoOp [67], where it is unknown how the learnable vector will be shifted by the conditional bias based on particular information taken from the image that is added to the learnable context vector. This lack of management over learnable

*These authors contributed equally.
†Corresponding author.

parameters can lead to unintentional bias in few-shot classification tasks or domain generalization tasks [22, 30, 32]. To address this, we propose a new approach called AAPL, "Adding Attributes to Prompt Learning", as illustrated in Fig. 1. In this context, augmentation generates a learnable bias that can be decomposed, with the augmented image serving as the visual prompt. Subsequent learning with visual-text prompts involves the use of a learnable context vector, which plays an adversarial role and mitigates unintended overfitting in downstream tasks [22, 30, 32]. In summary, our contributions are as follows:

- ◇ We propose using augmented images as a visual prompt and introduce the concept of *"delta meta token,"* which encapsulates attribute-specific information.
- ◇ Employing *delta meta token*, we conduct AdTriplet loss to make the conditional bias include the semantic feature of the class robustly, even in the presence of augmentation added to the learnable prompt through adversarial triplet loss.
- ◇ We demonstrate performance improvements in base-to-new generalization tasks, cross-dataset tasks, and domain generalization tasks.

2. Related Works

Vision-language models Vision-language models (VLMs) using image-text pairs have shown superior capabilities over image-only models, especially in zero-shot transfer tasks for various downstream classification tasks [46, 61, 63, 64, 66]. Prominent models such as CLIP [39] and ALIGN [17], which have advanced through large-scale web data utilization, employ self-supervised learning for enhanced textual and visual alignment. In the embedding space, the contrastive loss draws matched image-text representation pairs closer, while it draws the representation of mismatched pairs farther away. Using this method, CLIP demonstrates exceptional zero-shot image recognition capabilities without the need for further fine-tuning. Our goal is to find efficient methods for applying pretrained vision-language models to downstream applications, especially in prompt learning like CoOp [68] and CoCoOp [67].

Prompt learning in vision-language models The concept of prompt learning was initially proposed in the domain of natural language processing (NLP) [27–29]. Unlike manually designing prompts, prompt learning research focuses on automatically selecting prompts during the fine-tuning stage. Recently, this concept has been extended to the field of computer vision [18, 21, 31, 45, 52, 60, 68, 69]. CoOp [68] introduced continuous prompt learning to the vision domain, applying pretrained vision-language models to various tasks. Instead of using a manual prompt like "a photo of a", they transformed the context word into a learnable context vector to optimize continuous prompts. However, CoOp has limitations in generalizability due to overfitting on few-shot datasets. To address this, CoCoOp [67] adds a conditional bias called *meta token* extracted from image features to the learnable prompt. It shifts the focus from static to dynamic prompts, enabling optimization based on the characteristics of each instance rather than a specific class, consequently enhancing CoOp's domain generalization performance. However, *meta token* obtained from an image sample cannot be claimed to be completely robust against overfitting issues [22, 30, 32], and it is not interpretable because it is extracted from the shallow network, called *metanet*, composed of Linear-ReLU-Linear layers. Therefore, we propose a new prompt learning method using image augmentation to leverage attribute-specific bias added to learnable prompts.

Zero-shot learning Few-shot learning is the process of training on a small number of labeled samples before classifying the new images. In contrast, zero-shot learning (ZSL) aims to distinguish unseen classes by training exclusively on seen classes [5, 57]. This is achieved by exclusively training on a set of base classes and utilizing side information, typically visual attributes like color, shape, and other features, shared with these unseen classes. This auxiliary information helps the machine understand language or concepts in a way humans do, enabling it to recognize unseen classes. The common methods [4, 20, 34, 42, 55] are learning the relation between a class embedding and the image feature, which represents this auxiliary information. However, these methods often exhibit a bias against unseen classes, known as "seen-class bias" [56]. Other research efforts concentrate on enhancing visual-semantic embedding [3, 19, 65], or developing better image feature extractors [16, 59]. However, these methods usually assume a fixed set of auxiliary information, consisting of attributes labeled by humans. This assumption poses challenges, as labeling attributes is expensive, requires expert annotators, and is difficult to scale on large datasets. Diverging from existing ZSL methods, our work focuses on adapting large vision-language models and employs techniques based on prompting.

3. Methodology

3.1. Preliminaries

Prompt learning for CLIP CLIP [39] employs an image encoder based on ResNet [11] or ViT [24] and a text encoder based on Transformer [50] to extract features from images and text, respectively. These features are trained with a contrastive loss in the embedding space, aiming to maximize cosine similarities between paired modality features. When an input image x is processed through the image encoder $f(\cdot)$, it generates an image feature $f(x)$. Using a prompt

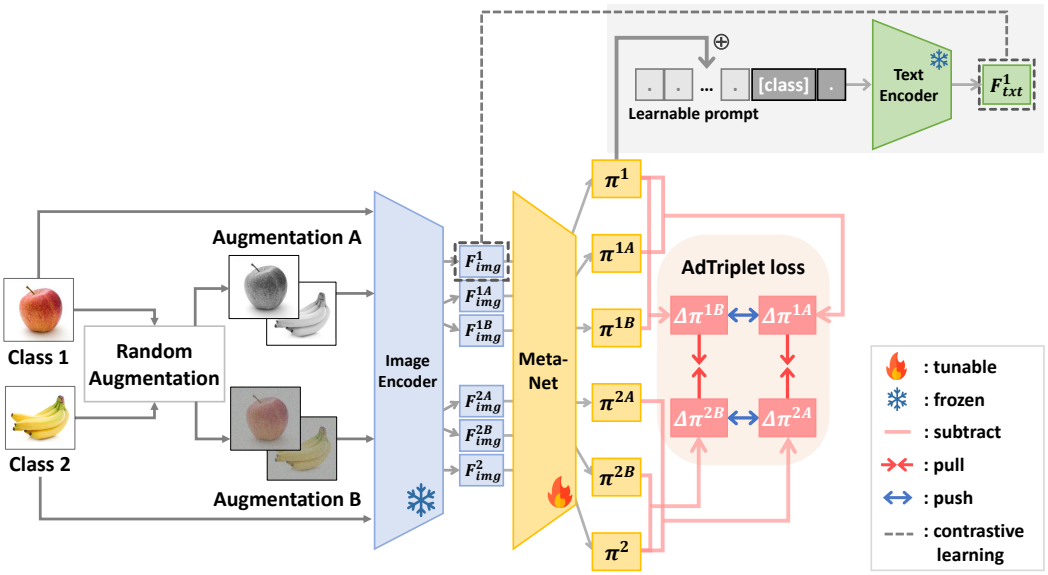


Figure 2. **Overview of AAPL.** We apply two distinct random augmentations to the input images, each with the class labels 1 and 2. Once the image features are extracted from the pretrained CLIP image encoder [39], they are passed through the *metanet* [67] to acquire the *meta token*. These are then utilized to subtract the other meta tokens obtained from the augmented images for each class, resulting in *delta meta tokens*. The goal is to instruct them to use these *delta meta tokens* regardless of their classification. The *delta meta tokens*, which are associated with the same augmentation, approach close within the embedding space using the AdTriplet loss, as shown in Eq. 5. The *delta meta tokens* acquire attribute-specific features, while the *meta token* learns semantic features derived from image features, enabling the use of attribute-specific bias in the learnable prompt through the decomposed features.

template like "a photo of a {class}.", where the {class} token is substituted with the name of the i-th class, yields K text features with corresponding weight vectors $\{w_i\}_{i=1}^{K}$ for the given K class categories. The prediction probability for CoOp is as Eq. 1, where $sim(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature parameter.

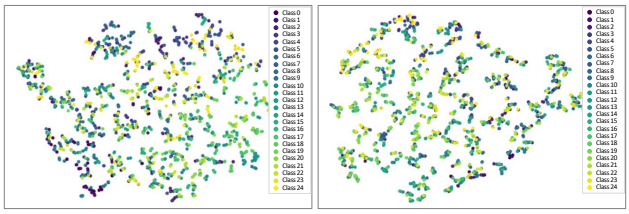
$$p(y|x) = \frac{\exp(sim(f(x), w_y)/\tau)}{\sum_{i=1}^{K} \exp(sim(f(x), w_i)/\tau)} \tag{1}$$

Conditional context optimization in prompt learning CoOp [68] introduces context tokens as trainable vectors, M learnable context, $\{v_1, v_2, ..., v_M\}$, departing from a fixed template like "a photo of a". The i-th class prompt, $t_i = \{v_1, v_2, ..., v_M, c_i\}$, includes these vectors and word embeddings of the class name, c_i. Text features are generated from t_i by CLIP text encoder $g(\cdot)$, which remained frozen throughout training. CoCoOp [67] proposes instance-conditional context to prioritize individual input instances, reducing the overfitting of CoOp. This is done by using a *metanet*, denoted as $h_\theta(\cdot)$ parameterized by θ, to generate a conditional token for each input. Where $\pi = h_\theta(f(x))$ and $m \in \{1, 2, ..., M\}$, each context token is obtained by $v_m(x) = v_m + \pi$. The prompt of the i-th class is conditioned on the input image feature, *i.e.*, $t_i(x) = \{v_1(x), v_2(x), ..., v_M(x), c_i\}$. Jointly updating context vectors $\{v_m(x)\}_{m=1}^{M}$ and *metanet* during training ensures generalizability. The prediction probability for CoCoOp is as follows:

$$p(y|x) = \frac{\exp(\text{sim}(f(x), g(t_y(x)))/\tau)}{\sum_{i=1}^{K} \exp(sim(f(x), g(t_i(x)))/\tau)}. \tag{2}$$

3.2. Delta Meta Token

Effect of augmentation in CoCoOp To investigate the effect of augmentation in prompt learning, we conducted a comparative experiment by adapting augmentation into CoCoOp [67]. We added conditional bias from augmented images to the learnable prompt while maintaining other settings consistent with CoCoOp. As detailed in Table 1, incorporating augmentation leads to a decrease in base-to-new generalization accuracy compared to the original CoCoOp since the *metanet* fails to extract the semantic features from the augmented images; thus extracting arbitrary noise rather than attribute-specific semantics. Additionally, as shown in Fig. 3, it does not show a big difference in class clustering, indicating that the *meta token* fails to capture the crucial semantic features for the classification. Consequently, this suggests that merely using augmentation in prompt learning might not enhance robustness or performance. It potentially leads to detrimental effects due to the *metanet*'s inability to identify meaningful semantic features from the augmented



(a) π of CoCoOp (b) π of CoCoOp with augmentation

Figure 3. The comparison between *meta tokens* of CoCoOp and *meta tokens* of CoCoOp with random augmentation for FGVCAircraft dataset.

Method	Base	New	HM
CoOp [68]	**82.69**	63.22	71.6
CoCoOp [67]	80.47	71.69	75.83
CoCoOp with augmentation	79.25	70.89	74.38
AAPL	80.65	**72.33**	**76.26**

Table 1. The comparison of base-to-new generalization accuracy between AAPL and CoCoOp with augmentation. HM denotes harmonic mean score.

images, focusing on instance-specific features rather than class semantics. To achieve optimal results, augmentation needs to be applied more carefully, ensuring that the conditional biases appropriately capture the semantic information of the class.

Delta meta token: detach attribute feature CoCoOp [67] improves the generalization performance of CoOp [68] by introducing *metanet*, which outputs *meta token* from image samples, then adds it to the learnable prompt. It focuses on learning about individual instance information rather than class information. However, it's still unclear what information the *meta token* contains, as the *metanet* is a black box, and its shallow architecture leads to uncertain feature extraction. As shown in Fig. 3, it fails to demonstrate clear clustering by neither augmentation type nor class. It shows that the *meta token* does not effectively capture the semantic information of the class or the attribute of the input image sample. To address this issue and make it possible to add desired information to the learnable prompt, we propose the concept of a *delta meta token*, the attribute-specific bias. The overview of AAPL is shown in Fig. 2.

To make a *delta meta token*, two images of each of the two different classes are required, *e.g.*, class 1 and class 2, as shown in Fig. 2. Two different augmentation types are randomly selected from 14 augmentations proposed in SimCLR [6] for each pair of input images without any duplication, which is denoted as $Aug_A(\cdot)$ and $Aug_B(\cdot)$. Inspired by TextManiA [62], which demonstrated the extraction of attribute information from text using Word Vector Analogy [9, 35], we generate *delta meta token* by subtracting image features in the same class with different augmentation. *Delta meta token* represents a difference vector from image features that contain augmentation information. They are generated at each iteration. The *delta meta token* from an image x of class 1 and $Aug_A(\cdot)$ can be written as follows:

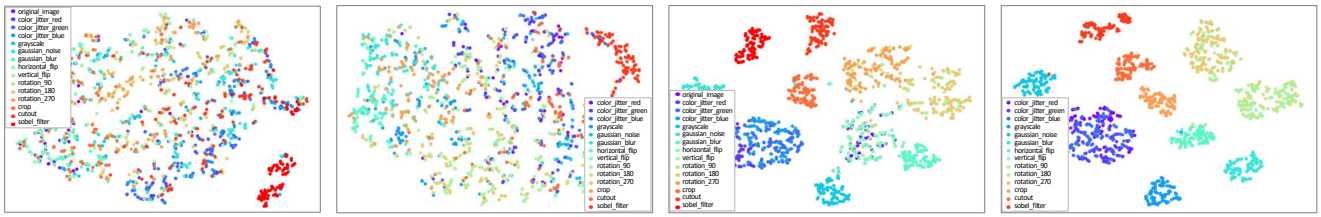
$$\Delta\pi^{1A} = h_\theta(f(Aug_A(x_1))) - h_\theta(f(x_1)). \quad (3)$$

As TextManiA has shown, utilizing attributes containing semantic details derived from class information demonstrates its effectiveness in classification tasks. In other words, while the *meta token* includes both class and attribute information, the *delta meta token* preserves more specific image feature information associated with augmentation. Adding decomposed auxiliary features to the learnable prompts, the *delta meta token* can learn attribute information. We enable the learnable prompt to incorporate semantic features more abundantly, thus making the augmentation more effective. Similar to adversarial prompt learning for natural language processing [37, 54], our method involves the adversarial interaction between class and attribute information, where the *metanet* learns to extract attribute-related information from augmented image features. The more the learnable prompt learns the semantic feature information of the class, the better the classification performance.

Does the delta meta token have exact augmentation information? In Fig. 4, we used t-SNE to compare the validation results of *metanet* of both CoCoOp [67] and AAPL. It shows that CoCoOp fails to distinguish between augmentations compared to AAPL. As comparing Fig. 4 (c) and (d), while *meta token* cannot perfectly discriminate 14 augmentations, *delta meta token* shows almost perfect distinction, except for a few augmentations, *e.g.*, vertical flip and rotations. This clustering result shows that the *delta meta token* extracts more specific information about augmentation than the *meta token*. As demonstrated in TextManiA [62], for the textual case, subtraction between features can retain specific features. In the case of the image, we show that *delta meta token* is more effective in making it contain the exact augmentation information. To the best of our knowledge, we are the first to employ feature decomposition through subtraction using visual features for prompt learning. It is noteworthy that, while the *meta token* still retains information about the class, the *delta meta token* accurately distinguishes between the semantic feature and the attribute feature.

3.3. Adversarial Triplet Loss

Using triplet loss [15, 43, 47, 53], we can eliminate the remaining class-specific information in the *delta meta token* while enhancing information related to augmentations. Training is conducted with 4 *delta meta tokens*, *e.g.*, $\Delta\pi^{1A}$, $\Delta\pi^{1B}$, $\Delta\pi^{2A}$, and $\Delta\pi^{2B}$, in the embedding space, aiming to increase the distance between vectors of the same class while



(a) π of CoCoOp (b) $\Delta\pi$ of CoCoOp (c) π of AAPL (d) $\Delta\pi$ of AAPL

Figure 4. t-SNE visualization of *meta token* and *delta meta token* of CoCoOp [67] and AAPL for FGVCAircraft dataset. The colors of the points represent the 14 different augmentations, and 100 data points from the validation set are used for this. (a) and (c) are the visualization of *meta token*, (b) and (d) are the visualization of *delta meta token*.

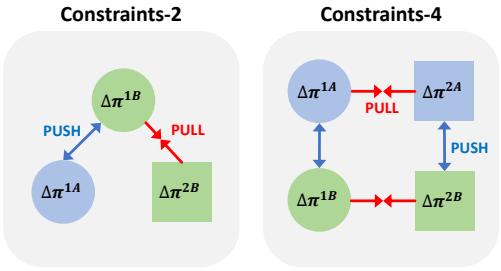


Figure 5. Comparison of the number of constraints of the AdTriplet loss. The constraints-2 setting's anchor is just one, *e.g.*, $\Delta\pi^{1B}$, and the constraints-4 setting has two anchors, *e.g.*, $\Delta\pi^{1A}$ and $\Delta\pi^{2B}$.

minimizing it for the same augmentation. For instance, considering anchor as $\Delta\pi^{1A}$, its positive pair is $\Delta\pi^{2A}$, which has a different class but the same augmentation. In contrast, $\Delta\pi^{1B}$ is considered a negative pair because it has the same class but a different augmentation. The distance between the anchor and the negative pair should be greater than the distance between the anchor and the positive pair. The Euclidean distance is denoted as $\|\cdot\|_2$, and the margin of the triplet loss is denoted as m in Eq. 4.

$$\begin{aligned} &L_{triplet}(x, x^+, x^-; \Delta\pi^{1A}, \Delta\pi^{2A}, \Delta\pi^{1B}) \\ &\quad = max(0,\ \|x - x^+\|_2 - \|x - x^-\|_2 + m) \\ &= max(0,\ \|\Delta\pi^{1A} - \Delta\pi^{2A}\|_2 - \|\Delta\pi^{1A} - \Delta\pi^{1B}\|_2 + m) \end{aligned} \quad (4)$$

Thus, we introduce the Adtriplet loss, which adversarially trains the model to prioritize the alignment of augmentation information over class information. This loss is updated alongside the classification loss, specifically the cross-entropy loss. The AdTriplet loss is used as constraints-4, as illustrated in Fig. 5, to make the connection between the class information domain and augmentation attribute domain more balanced [23].

$$\begin{aligned} L_{AdTriplet} = &L^1_{triplet}(\Delta\pi^{1A}, \Delta\pi^{2A}, \Delta\pi^{1B}) \\ &+ L^2_{triplet}(\Delta\pi^{2B}, \Delta\pi^{1B}, \Delta\pi^{2A}) \end{aligned} \quad (5)$$

Cross-entropy loss is computed following the same method as CoCoOp [67]. To ensure fairness between the training and test phases, only one input image label is used for cross-entropy loss calculation. The final training loss function is as follows:

$$L_{total} = \alpha * L_{AdTriplet} + \beta * L_{CE}, \quad (6)$$

where α and β are hyper-parameters for scaling. In Sec. 4, we provide detailed information on parameter tuning.

4. Experiments

4.1. Experimental Settings

Datasets We use 11 classification datasets based on CLIP [39], CoOp [68], and CoCoOp [67] for base-to-new generalization and cross-dataset transfer: ImageNet [8] and Caltech101 [10] for generic object classification, OxfordPets [38], StanfordCars [26], Flowers102 [36], Food101 [2] and FGVCAircraft [33] for fine-grained image recognition, EuroSAT [12] for satellite image classification, UCF101 [48] for action classification, DTD [7] for texture classification, and SUN397 [58] for scene recognition. For domain generalization experiments, we use ImageNet [8] as the source dataset and 4 other ImageNet-based datasets, *i.e.*, ImageNetV2 [41], ImageNetSketch [51], ImageNet-A [14], and ImageNet-R [13], as the target datasets, which each contain a different kind of domain shift.

Baselines We compare AAPL with 3 baseline methods: the zero-shot CLIP [39], CoOp [68], and CoCoOp [67]. CLIP uses the hand-crafted template "a photo of a {class}" to generate the prompts for knowledge transfer. CoOp learns a static prompt that replaces the hand-crafted prompts with the learnable vectors. CoCoOp generates dynamic prompts by adding the image-conditional prompts to the learnable prompts in CoOp.

Training details Our implementation is based on CoCoOp [67]. We employ the pre-trained ViT-B/16 model from CLIP [39] as the backbone. We fix the context length to 4 and initialize the context vectors randomly. The presented results are the mean values obtained from experiments

conducted with three random seeds. We follow the training epochs, batch sizes, and schedules as prescribed by CoCoOp. In the context of few-shot learning, we confine evaluation to the maximum shot, *i.e.*, 16 shots, considered by CoOp. For evaluation, we use the model from the last epoch. The parameter size of AAPL is the same as CoCoOp, and the hyper-parameter m in Eq. 4 is set to 0.2.

Dataset		CLIP [39]	CoOp [68]	CoCoOp [67]	**AAPL** (Ours)	Δ
Average on 11 datasets	Base	69.34	**82.69**	80.47	80.27	-0.20
	Novel	**74.22**	63.22	71.69	72.17	+0.48
	HM	71.70	71.66	75.83	**76.01**	+0.18
ImageNet	Base	72.43	76.47	75.98	**76.53**	+0.55
	Novel	68.14	67.88	70.43	**70.57**	+0.14
	HM	70.22	71.92	73.10	**73.43**	+0.33
Caltech101	Base	96.84	**98.00**	97.96	97.87	-0.09
	Novel	94.00	89.81	93.81	**95.10**	+1.29
	HM	95.40	93.73	95.84	**96.46**	+0.62
OxfordPets	Base	91.17	93.67	95.20	**95.63**	+0.43
	Novel	97.26	95.29	**97.69**	97.40	-0.29
	HM	94.12	94.47	96.43	**96.51**	+0.08
Stanford Cars	Base	63.37	**78.12**	70.49	70.33	-0.16
	Novel	**74.89**	60.40	73.59	73.50	-0.09
	HM	68.65	68.13	**72.01**	71.88	-0.13
Flowers102	Base	72.08	**97.60**	94.87	95.10	+0.23
	Novel	**77.80**	59.67	71.75	70.63	-1.12
	HM	74.83	74.06	**81.71**	81.06	-0.65
Food101	Base	90.10	88.33	**90.70**	**90.70**	+0.00
	Novel	91.22	82.26	91.29	**91.60**	+0.31
	HM	90.66	85.19	90.99	**91.15**	+0.16
FGVC Aircraft	Base	27.19	**40.44**	33.41	34.07	+0.66
	Novel	**36.29**	22.30	23.71	24.17	+0.46
	HM	**31.09**	28.75	27.74	28.28	+0.54
SUN397	Base	69.36	80.60	**79.74**	79.65	-0.09
	Novel	75.35	65.89	76.86	**76.90**	+0.04
	HM	72.23	72.51	**78.27**	78.25	-0.02
DTD	Base	53.24	**79.44**	77.01	73.90	-3.11
	Novel	**59.90**	41.18	56.00	53.43	-2.57
	HM	56.37	54.24	**64.85**	62.02	-2.83
EuroSAT	Base	56.48	**92.19**	87.49	87.00	-0.49
	Novel	64.05	54.74	60.04	**66.30**	+6.26
	HM	60.03	68.69	71.21	**75.25**	+4.04
UCF101	Base	70.53	**84.69**	82.33	82.20	-0.13
	Novel	**77.50**	56.05	73.45	74.27	+0.82
	HM	73.85	67.46	77.64	**78.03**	+0.39

Table 2. **Base-to-new generalization experiment compared to baselines.** The model is trained from the base classes (16 shots) and evaluated in new classes. HM denotes the harmonic mean. Δ is the difference between AAPL and CoCoOp. The **bold** highlighting indicates the highest performance scores.

4.2. Generalization from Base-to-New Classes

We divided the classes equally into two groups, one for the base classes and another for the new classes, *i.e.*, unseen classes, just like in CoCoOp [67]. Learning-based models are trained solely on base classes. In few-shot learning, the model is evaluated with the base classes, whereas in zero-shot learning, it is evaluated with the new classes to test the model's generalizability. In this task, we set hyper-parameters α and β to 0.2 and 1. Table 2 presents the performance results of AAPL compared to the baseline. AAPL outperformed in 7 out of 11 datasets, with the harmonic mean of total dataset accuracy exceeding that of CoCoOp. However, performance on the DTD [7] was significantly lower. The geometrical augmentations, especially flips and rotations, appear to have minimal effect on AAPL, as they do not significantly alter the appearance of the original images in the context of texture. This demonstrates that the effectiveness of AAPL varies across different datasets.

4.3. Cross-Dataset Transfer

To assess the robustness and adaptability of AAPL, we tested its generalization ability across datasets by training it on all 1000 ImageNet classes and then applying it on the other 10 datasets, as shown in Table 3. We assume that the model can learn semantic information about image features by learning precise attributes. To evaluate this, we increased the model's focus on learning augmentation information by setting both hyper-parameters, α and β, to 1 in this experiment and afterward. AAPL achieves higher generalization in 3 datasets: OxfordPets [38], FGVCAircraft [33], and UCF101 [48], compared to CoCoOp [67]. However, the performance on DTD [7] and EuroSAT [12] was noticeably poorer than other datasets. This suggests that these datasets are vulnerable to AAPL's augmentation-based prompt learning. These datasets are not object-centric but rather possess global features, *e.g.*, long-distance satellite images and texture images. Extracting specific attributes from these datasets is challenging due to their unique characteristics.

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
CoOp	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
CoCoOp	71.02	**94.43**	90.14	**65.32**	**71.88**	**86.02**	22.94	**67.36**	**45.73**	**45.37**	68.21	**65.74**
AAPL	71.37	94.17	**90.73**	65.10	71.67	86.00	**23.03**	66.80	44.80	41.83	**69.30**	65.34

Table 3. **Cross-dataset transfer experiment.** The model is trained on the entire class of ImageNet (16 shots) and evaluated on the other 10 datasets.

4.4. Domain Generalization

For domain generalization, we trained our model on the whole ImageNet dataset, same as in Sec. 4.3, and evaluated it on 4 datasets that represent a domain shift from ImageNet (*e.g.*, ImageNetV2 [41], ImageNetSketch [51], ImageNet-A [14], and ImageNet-R [13]). The comparison of these tests

is presented in Table 4. We achieved better performance on all datasets except for ImageNet-A. This demonstrates that attribute-specific bias effectively deals with domain shift.

	Source	Target				
	ImageNet	-V2	-S	-A	-R	*Avg.*
CLIP	66.73	60.83	46.15	47.77	73.96	57.18
CoOp	**71.51**	**64.20**	47.99	49.71	75.21	59.28
CoCoOp	71.02	64.07	48.75	**50.63**	76.18	59.91
AAPL	71.37	**64.20**	**48.80**	50.60	**76.87**	**60.12**

Table 4. **Domain Generalization experiment.** The model is trained on the entire class of ImageNet (16 shots) and evaluated on four different ImageNet-based datasets, including domain shifts.

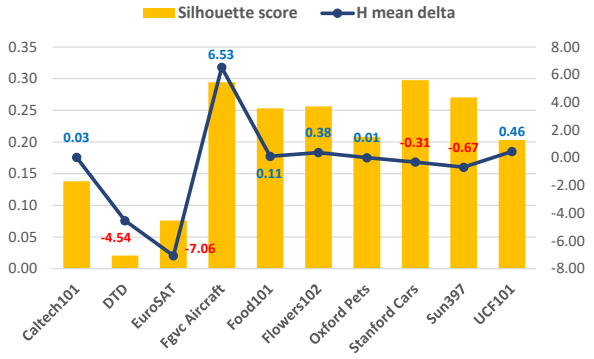


Figure 6. **The correlation between silhouette score and generalization performance.** Silhouette score and the difference in harmonic mean accuracy for zero-shot classification between CoCoOp and AAPL

4.5. Augmentation Profiling

Why should the delta meta token learn about attributes rather than class information? To assess the effectiveness of learning attributes, we compared the silhouette scores [44] based on augmentation types. The silhouette score evaluates how well data points are clustered, considering both cohesion (proximity within the same cluster) and separation (distance from the nearest neighboring cluster). The silhouette score $S(i)$ for data point i, is calculated as follows: $S(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}}$, where $a(i)$ is the average distance of i to all other data points in the same cluster, and $b(i)$ is the average distance of i to the data points in the nearest cluster that i does not belong to. A higher silhouette score indicates better clustering. In other words, datasets that effectively learn information about augmentations from the Adtriplet loss have higher silhouette scores. As shown in Fig. 6, the zero-shot classification performance of AAPL generally improves. However, there is a sharp decrease in performance for DTD [7] and EuroSAT [12]. This suggests that datasets that cannot effectively extract augmentation information do not perform well. Training precise attributes to *delta meta token* is crucial for zero-shot classification, and it's evident that determining what information to add to the learnable prompt is highly important for datasets sensitive to AAPL.

Which dataset is vulnerable for AAPL? To assess the impact of various datasets on the evaluation of learning attribute features, we applied AAPL's proposed AdTriplet loss and the traditional triplet loss method. Unlike the AdTriplet loss, the traditional triplet loss trains the *delta meta token* to cluster classes rather than augmentation types. As shown in Table 5, when utilizing Adtriplet loss across 6 datasets, performance improvement was observed compared to using triplet loss. Particularly, FGVCAircraft [33] exhibited approximately a 7% higher

	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
Triplet	**73.44**	95.81	96.18	**72.22**	80.65	90.70	27.97	**78.34**	**61.73**	64.15	**78.78**	74.54
AdTriplet	73.09	**96.87**	**96.44**	71.70	**82.09**	**91.10**	**34.27**	77.60	60.31	**65.16**	78.10	**75.16**

Table 5. **AAPL with Triplet and AdTriplet loss.** The comparison of harmonic means of base-to-new generalization accuracy between AAPL trained with AdTriplet loss and traditional Triplet loss.

performance improvement with the triplet loss. Utilizing the traditional triplet loss method means that the *delta meta token* is trained to bring the same class together regardless of augmentation type. Consequently, datasets that showed improved performance on AdTriplet loss have a higher dependency on class information. Triplet loss-trained *delta meta token* extracts class-related information, causing *meta token* to contribute noisy features rather than class semantic features when added to prompts. In contrast, AdTriplet loss-trained tokens focus on extracting the class semantic features. Datasets with AdTriplet loss perform well because they rely more on the class information. This highlights the advantage of AAPL based on the dataset's characteristics.

Which augmentation is effective to prompt learning? The t-SNE visualization of the *delta meta token* for 14 augmentations is shown, along with their silhouette scores, in Fig. 7 (a). It turned out that it is difficult to distinguish rotations from flips and between color jitters, while other augmentations are obvious. All datasets exhibit difficulty in distinguishing these augmentations. Following selective augmentation training, when trained only on augmentations whose results are good (shown in Fig. 7 (b)), clustering is greatly enhanced, and silhouette scores are also raised. Also, the average performance for base-to-new generalization improved, as seen in Table 6. But when training solely with the opposite, *i.e.*, bad augs (Fig. 7 (c)), there is neither significant improvement in silhouette scores nor in the average

Method	**AAPL**	Good Augs	Bad Augs
ImageNet	**73.09**	72.91	73.05
Caltech101	95.87	**96.43**	96.00
OxfordPets	96.44	**96.49**	95.96
StanfordCars	71.70	**71.85**	71.67
Flowers102	**82.09**	80.80	81.74
Food101	**91.10**	90.45	90.90
FGVCAircraft	**34.27**	34.02	18.14
SUN397	77.60	77.97	**78.03**
DTD	60.31	61.24	**61.43**
EuroSAT	64.15	66.68	**74.70**
Ucf101	78.10	77.09	**78.11**
Average	74.97	**75.08**	74.52

Table 6. **AAPL with some augmentation types.** The comparison of harmonic means of base-to-new generalization accuracy when conducting AAPL using only good augmentations and bad augmentations.

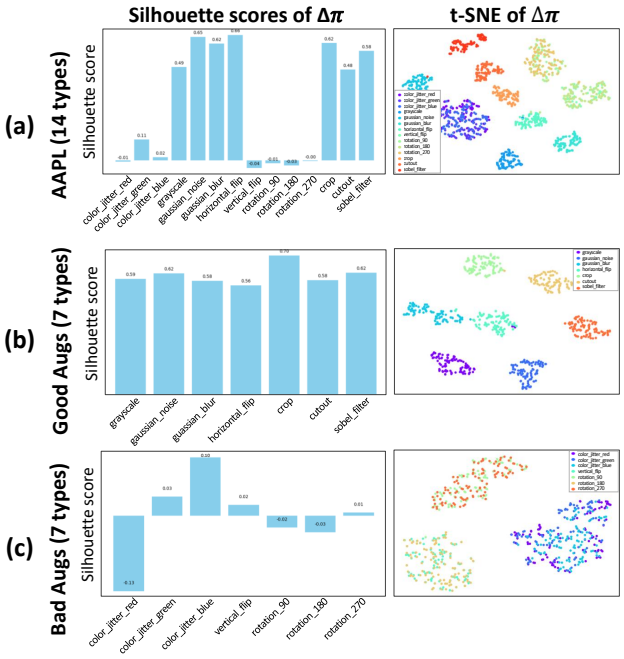


Figure 7. **The comparison of silhouette score and t-SNE** of the base-to-new generalization for each of the specific augmentation types on FGVCAircraft. All results are from the last epoch.

base-to-new generalization results. The ambiguity of augmentations between flips and rotations and between color jitters limits the learning capacity of the *metanet*.

AAPL with weighted random sampling Fig. 6 shows a consistent correlation between lower silhouette scores and worse zero-shot classification performance compared to CoCoOp [67] across several datasets. Insufficient knowledge of semantic features makes classifying unseen classes more difficult. To address this, an active approach [25, 40, 49] was utilized for datasets DTD [7], EuroSAT [12], StanfordCars [26], and SUN397 [58], which have insufficient learning of augmentation type information. For training, silhouette scores were used as thresholds for random sampling weights. As shown in Table 7, this improved the performance of base-to-new generalization across all 4 datasets. Notably, EuroSAT showed a significant 10% improvement, emphasizing the effectiveness of dynamically selecting and emphasizing weaker augmentation types during each epoch. It demonstrates that attribute-specific feature decomposition for challenging augmentations enables more robust learning of semantic features.

	AAPL	WRS	Δ
StanfordCars	71.70	71.82	**+0.12**
SUN397	77.60	78.14	**+0.54**
DTD	60.31	61.39	**+1.08**
EuroSAT	64.15	74.25	**+10.10**

Table 7. **AAPL with weighted random sampling for vulnerable 4 datasets.** The comparison of harmonic means of base-to-new generalization accuracy. WRS is short for weighted random sampled AAPL.

5. Conclusion

Our novel approach efficiently extracts specific semantic features and delta meta tokens by subtracting the augmented image feature from the original image feature. Leveraging AdTriplet loss adversarially enhances classification loss, enabling precise discernment of attribute features through augmentations—a foundational aspect of our approach. By decomposing attribute and semantic features more accurately, we introduce attribute-specific bias into the prompt. Furthermore, our study underscores the indispensability of AAPL in prompt learning with augmentation for zero-shot classification tasks. In summary, our emphasis on attribute decomposition in prompt learning is underscored through augmentation profiling and analysis of dataset correlations, augmentations, and AAPL performance.

Acknowledgments Thanks to Prof. George Kamenos for his invaluable assistance in reviewing and editing this paper. This work was partly supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (IITP-2024-00156287, 40%). This research was supported by the Korea Institute for Advancement of Technology (KIAT) grant funded by the Ministry of Trade, Industry, and Energy (MOTIE), Korea, (P0025331, 30%). This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00252616, 30%).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022. 1

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 5

[3] Yannick Le Cacheux, Herve Le Borgne, and Michel Crucianu. Modeling inter and intra-class relations in the triplet loss for zero-shot learning. In *ICCV*, 2019. 2

[4] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, 2016. 2

[5] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, 2016. 2

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 4

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 5, 6, 7, 8

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[9] Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Towards understanding linear word analogies. In *ACL*, 2019. 4

[10] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshop*, 2004. 5

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[12] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 5, 6, 7, 8

[13] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021. 5, 6

[14] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021. 5, 6

[15] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*. Springer, 2015. 4

[16] Zhong Ji, Yanwei Fu, Jichang Guo, Yanwei Pang, Zhongfei Mark Zhang, et al. Stacked semantics-guided attention model for fine-grained zero-shot learning. *NIPS*, 2018. 2

[17] 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 *ICML*, 2021. 1, 2

[18] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 1, 2

[19] Huajie Jiang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Transferable contrastive network for generalized zero-shot learning. In *ICCV*, 2019. 2

[20] Muhammad Gul Zain Ali Khan, Muhammad Ferjad Naeem, Luc Van Gool, Alain Pagani, Didier Stricker, and Muhammad Zeshan Afzal. Learning attention propagation for compositional zero-shot learning. In *WACV*, 2023. 2

[21] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, 2023. 2

[22] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *ICCV*, 2023. 2

[23] Junsik Kim, Seokju Lee, Tae-Hyun Oh, and In So Kweon. Co-domain embedding using deep quadruplet networks for unseen traffic sign recognition. In *AAAI*, 2018. 5

[24] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[25] Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Learning active learning from data. *NIPS*, 2017. 8

[26] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshop*, 2013. 5, 8

[27] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021. 2

[28] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, 2021.

[29] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *ACL*, 2022. 2

[30] Yajing Liu, Yuning Lu, Hao Liu, Yaozu An, Zhuoran Xu, Zhuokun Yao, Baofeng Zhang, Zhiwei Xiong, and Chenguang Gui. Hierarchical prompt learning for multi-task learning. In *CVPR*, 2023. 2

[31] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *CVPR*, 2022. 1, 2

[32] Chengcheng Ma, Yang Liu, Jiankang Deng, Lingxi Xie, Weiming Dong, and Changsheng Xu. Understanding and mitigating overfitting in prompt tuning for vision-language models. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2

[33] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5, 6, 7

[34] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *CVPR*, 2021. 2

[35] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *NAACL*, 2013. 4

[36] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, 2008. 5

[37] Venkata Prabhakara Sarath Nookala, Gaurav Verma, Subhabrata Mukherjee, and Srijan Kumar. Adversarial robustness of prompt-based few-shot learning for natural language understanding. In *ACL*, 2023. 4

[38] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 5, 6

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 5, 6

[40] Hiranmayi Ranganathan, Hemanth Venkateswara, Shayok Chakraborty, and Sethuraman Panchanathan. Deep active learning for image classification. In *ICIP*. IEEE, 2017. 8

[41] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 5, 6

[42] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015. 2

[43] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 4

[44] Ketan Rajshekhar Shahapure and Charles Nicholas. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*. IEEE, 2020. 7

[45] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *NeurIPS*, 35:14274–14289, 2022. 1, 2

[46] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, 2022. 1, 2

[47] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *NIPS*, 2016. 4

[48] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5, 6

[49] Alex Tamkin, Dat Nguyen, Salil Deshpande, Jesse Mu, and Noah Goodman. Active learning helps pretrained models learn the intended task. *NeurIPS*, 2022. 8

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 2017. 1, 2

[51] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 2019. 5, 6

[52] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *CVPR*, 2022. 2

[53] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009. 4

[54] Hui Wu and Xiaodong Shi. Adversarial soft prompt tuning for cross-domain sentiment analysis. In *ACL*, 2022. 4

[55] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016. 2

[56] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, 2017. 2

[57] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018. 2

[58] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 5, 8

[59] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. *NeurIPS*, 2020. 2

[60] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, 2023. 2

[61] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *ICLR*, 2022. 1, 2

[62] Moon Ye-Bin, Jisoo Kim, Hongyeob Kim, Kilho Son, and Tae-Hyun Oh. Textmania: Enriching visual feature by text-driven manifold augmentation. In *ICCV*, 2023. 4

[63] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 1, 2

[64] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022. 1, 2

[65] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2017. 2

[66] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, 2022. 1, 2

[67] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 8

[68] 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, 2022. 1, 2, 3, 4, 5, 6

[69] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *CVPR*, 2023. 2