Attend and Enrich: Enhanced Visual Prompt for Zero-Shot Learning

Man Liu[2,4], **Huihui Bai**[1,2,4✉], **Feng Li**[3✉], **Chunjie Zhang**[2,4],
Yunchao Wei[2,4], **Tat-Seng Chua**[5], **Yao Zhao**[2,4]

[1] Tangshan Research Institute of Beijing Jiaotong University
[2]Institute of Information Science, Beijing Jiaotong University
[3] Hefei University of Technology
[4] Beijing Key Laboratory of Advanced Information Science and Network Technology
[5] National University of Singapore
{manliu, hhbai, cjzhang, yunchao.wei, yzhao}@bjtu.edu.cn, fengli@hfut.edu.cn, chuats@comp.nus.edu.sg

Abstract

Zero-shot learning (ZSL) endeavors to transfer knowledge from the seen categories to recognize unseen categories, which mostly relies on the semantic-visual interactions between image and attribute tokens. Recently, the prompt learning has emerged in ZSL and demonstrated significant potential as it allows the zero-shot transfer of diverse visual concepts to downstream tasks. However, current methods explore the fixed adaptation of the learnable prompt on the seen domains, which make them over-emphasize the primary visual features observed during training, limiting their generalization capabilities to the unseen domains. In this work, we propose AENet, which endows semantic information into the visual prompt to distill semantic-enhanced prompt for visual representation enrichment, enabling effective knowledge transfer for ZSL. AENet comprises two key steps: 1) exploring the concept-harmonized tokens for the visual and attribute modalities, grounded on the modal-sharing token that represents consistent visual-semantic concepts; and 2) yielding the semantic-enhanced prompt via the visual residual refinement unit with attribute consistency supervision. It is further integrated with primary visual features to attend to semantic-related information for visual enhancement, thus strengthening transferable ability. Experimental results on three benchmarks show that our AENet outperforms existing state-of-the-art ZSL methods.

Code — https://github.com/ManLiuCoder/AENet.

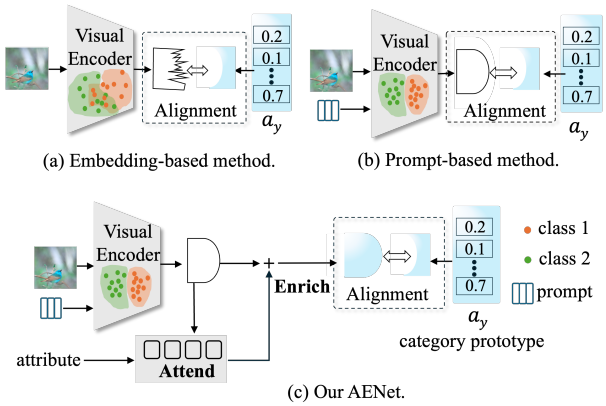

Figure 1: **Motivation of AENet.** (a) Embedding-based methods align visual and semantic features for ZSL, which suffer from vague visual representations due to cross-dataset bias. (b) Prompt-based methods add the learnable prompt to pre-trained vision encoders, adapting them to ZSL scenarios while over-emphasizing primary visual features. (c) Our AENet distills the semantic-enhanced prompt that attends to semantic-related details to enrich visual representations, ensuring more comprehensive knowledge transfer.

Introduction

Humans naturally interact with the world through various channels such as vision and language, enabling them to identify unseen objects based on prior knowledge. Leveraging this cognitive capability, zero-shot learning (ZSL) (Palatucci et al. 2009; Lampert, Nickisch, and Harmeling 2009) aims to classify objects from the unseen domain carrying knowledge from seen categories. One of the mainstream methods in ZSL is the embedding-based approach (Xu et al. 2020; Liu et al. 2021; Chen et al. 2022d,b; Liu et al. 2024b), which learns the semantic-visual alignment in a joint embedding space. Prior to such a cross-modal alignment, as depicted in Figure 1(a), embedding-based models typically commence with a visual encoder, such as ResNet (He et al. 2016) or ViT (Dosovitskiy et al. 2021) pre-trained on ImageNet (Deng et al. 2009), to initialize visual features. However, this can introduce distribution discrepancy when applied to downstream ZSL benchmarks as wider categories are unobserved during pre-training, thereby resulting in cross-dataset bias that produces vague or even misleading visual representations (Torralba and Efros 2011).

To mitigate the distribution discrepancy and cross-dataset bias issues arising from using pre-trained encoders, recent ZSL methods have incorporated attention mechanisms (Xie et al. 2019; Zhu et al. 2019) and vision transformers (Chen et al. 2022b,a, 2024). These approaches refine visual representations by focusing on informative image regions and

✉ Corresponding author.

capturing complex visual-semantic relationships. Recently, motivated by the powerful zero-shot capabilities of prompt engineering (Kojima et al. 2022; Cheng et al. 2023), some works (Jia et al. 2022; Zhou et al. 2022b,a; Wang et al. 2023) propose to incorporate the learnable visual prompt into pre-trained vision encoders to adapt the visual space to downstream datasets (Figure 1(b)), which effectively alleviates the cross-dataset bias. However, in these methods, the learned prompt often tends to overfit the base seen classes and concentrate solely on primary visual features essential for recognizing seen categories (Zhou et al. 2022a). Consequently, they may lack sufficient capacity to capture crucial semantic-related visual features necessitated for a broader range of unseen classes. These features, which are complementary to primary contents, are pivotal for effective cross-domain semantic transfer.

To address these issues, in this work, we propose a novel method, which embraces prompt learning with ZSL. This method follows the existing prompt-based pipeline but investigates concept-harmonized semantic features to provide the semantic-enhanced prompt for visual enhancement, thereby achieving more comprehensive knowledge discovery. It takes the attribute, image, and initialized learnable visual prompt as input, to obtain corresponding embeddings. Given the inherent disparity between the image and attribute, we first devise the concept-aware attention (CAA) to harmonize the concept of visual and attribute tokens. This attention defines a modal-sharing token as the reference that adapts both modalities to concept-harmonized tokens, which are expected to reveal consistent semantic and visual concepts. Then, a visual residual refinement unit (VRRU) operating in a simple yet efficient linear manner is constructed to mine semantic-related details under attribute consistency supervision. In this way, we can form the semantic-enhanced prompt that attends to semantic-related features by combining the predicted residuals with the initial prompt embedding. They are further utilized to enrich the primary visual representations, facilitating the efficacy of visual-semantic alignment for unseen domain recognition. We call this method AENet. Extensive experimental results show that the proposed AENet leads to better performance on ZSL benchmarks.

In summary, our work makes the following contributions: 1) We propose AENet that leverages prompt learning with concept-harmonized semantic features to enhance visual-semantic alignment for ZSL. 2) We introduce the concept-aware attention to harmonize visual and attribute concepts grounded on the modal-sharing token. 3) We propose VRRU, a linear unit which mines semantic-related details under attribute consistency supervision, generating the semantic-enhanced prompt through the combination of predicted residuals.

Related Work

Zero-Shot Learning

ZSL approaches primarily fall into two categories: generative-based and embedding-based methods. Generative methods synthesize visual features for unseen categories using techniques like generative adversarial networks (Han et al. 2021) or variational autoencoders (Chen et al. 2021b,c). Although these methods compensate for the absence of the unseen domain during training, they introduce extra data. The embedding-based method represents the other mainstream branch of GZSL, achieved through the projection and alignment of information from visual and semantic modalities. Early works (Akata et al. 2013; Xian et al. 2016) propose to align images and attributes within a common feature space. However, global visual information falls short of capturing the subtle yet substantial differences between categories, weakening discriminative representations. Thus, part-based methods attempt to highlight the most important parts of the input, better refining the extracted visual representation. For example, some investigations (Xu et al. 2020; Wang et al. 2021; Chen et al. 2022d; Liu et al. 2023a) have pursued semantic-guided attention to localize discriminative attribute-related parts and capture crucial fine-grained features. Distinctive visual features are also emphasized by graph networks (Xie et al. 2020; Hu et al. 2021) that incorporate region-based relational analysis to mine complementary connections among diverse spatial elements. However, they are still confused by the cross-dataset bias issues arising from pre-trained encoders. In contrast to the previous methods, we embrace the idea of prompt learning and enrich the visual features for desirable visual-semantic alignment.

Prompt Learning

Prompt learning emerges as a pivotal methodology within the realm of natural language processing (NLP), which can adapt Large Language Models (LLMs) to various downstream tasks and scenarios (Liu et al. 2023b; Lyu et al. 2024; Tan et al. 2024). Initially, the utilization of prompt learning depends primarily on hand-crafted prompts, which are composed of thoughtfully selected words or phrases. While potentially powerful, it presents non-trivial prompt engineering challenges that require not only extensive domain knowledge but also consume significant time due to its inherently iterative nature of trial and error. Studies such as (Lester, Al-Rfou, and Constant 2021) and (Liu et al. 2023c) attempt to introduce a learnable continuous prompt that is updated dynamically during the learning process, boosting model performance and efficiency. The success of prompt learning in LLMs has sparked interest in computer vision (Liu et al. 2024a; Wang et al. 2024). CoOp (Zhou et al. 2022b) enhances pre-trained vision-language models for downstream image classification through continuous prompt learning. Several methods (Zhou et al. 2022a; Wang et al. 2023) improve CoOp by introducing image-conditioned tokens or synthesized prompts for underrepresented classes, which advance the capabilities of vision-language models in few/zero-shot learning. However, these models mainly focus on primary visual content sufficient for seen classes but ignore crucial visual details that are critical for unseen domains. In this paper, we focus on improving the prompt-based content by supplementing semantic-related information for visual enhancement, allowing it generalize well to the unseen classes.

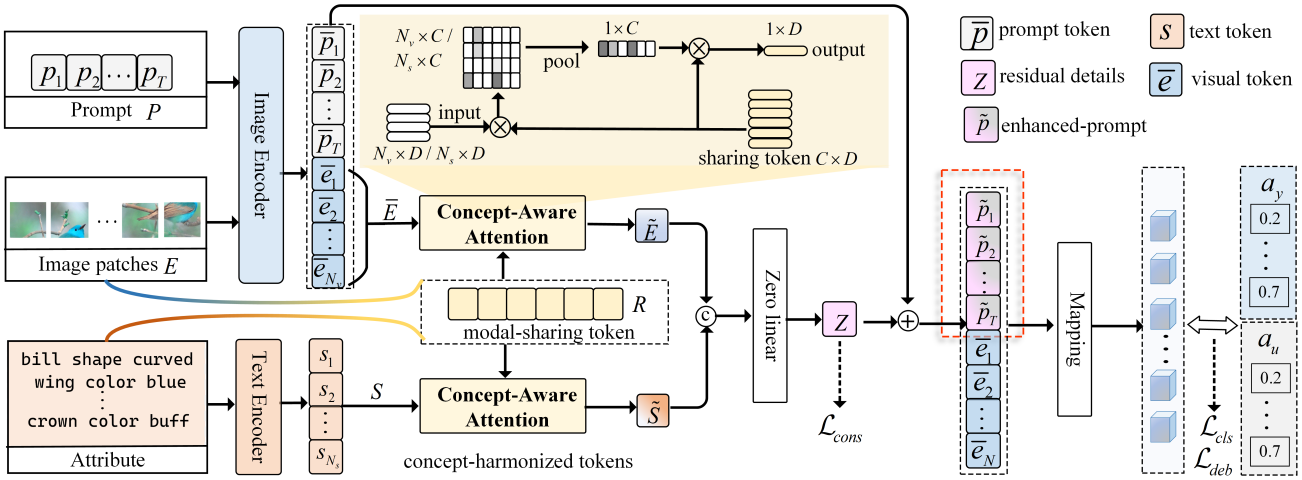

Figure 2: AENet processes attributes, images, and the learnable prompt through the pre-trained encoder. It performs concept-harmonized token exploration using concept-aware attention, followed by semantic-enhanced prompt distillation with attribute consistency. The enhanced prompt is then integrated with visual features to attend to semantic-related information, improving knowledge discovery for unseen classes.

Methodology

The proposed AENet aims to explore the semantic-enhanced prompt, which enriches primary visual representations to conduct accurate semantic-visual alignment for ZSL. As shown in Figure 2, AENet takes the attribute, image, and initialized learnable visual prompt as input to produce corresponding embeddings via the text and image encoders. Upon these, there are two key steps in AENet: 1) Concept-harmonized token exploration for visual and attribute modalities, which is grounded on the modal-sharing token that encodes consistent visual-semantic concepts by using the concept-aware attention; 2) Semantic-enhanced prompt distillation via the visual residual refinement unit with attribute consistency supervision. After that, we integrate the enhanced prompt with visual features to attend to semantic-related information for visual enhancement, achieving more effective and comprehensive knowledge discovery.

Problem Formulation

ZSL aims to discern the novel image categories within the unseen domain $\mathcal{D}^u$ by capturing knowledge derived from the seen domain data $\mathcal{D}^s$. Here, $\mathcal{D}^s = \{(x, y, a_y)|x \in \mathcal{X}^s, y \in \mathcal{Y}^s, a_y \in \mathcal{A}^s\}$ consists of the images x in $\mathcal{X}^s$, corresponding label y, and its associated category prototype a_y from $\mathcal{A}^s$. Similarly, the unseen domain data is defined as $\mathcal{D}^u = \{(x^u, u, a_u)\}$, where $x^u \in \mathcal{X}^u$, $u \in \mathcal{Y}^u$, $a_u \in \mathcal{A}^u$, with $\mathcal{A} = \mathcal{A}^s \cup \mathcal{A}^u$. Note that the category space is disjoint between seen and unseen domains, *i.e.*, $\mathcal{Y}^s \cap \mathcal{Y}^u = \varnothing$, $\mathcal{Y}^s \cup \mathcal{Y}^u = \mathcal{Y}$.

Utilizing $\mathcal{D}^s$ during the training stage, the typical embedding-based ZSL framework learns a mapping function $\mathcal{M}(\cdot)$ that bridges the image space $\mathcal{X}^s$ with the attribute space $\mathcal{A}^s$. This enables the model to generalize its knowledge to $\mathcal{D}^u$ and establish connections between $\mathcal{X}^u$ and $\mathcal{A}^u$ for category inference. In scenarios where the test phase includes both seen and unseen classes, conventional ZSL is often extended to generalized zero-shot learning (GZSL), rendering it more applicable to real-world scenarios. Given an input image representation $f(x)$ during training, the optimization of the embedding-based framework is conducted through the visual-semantic alignment, as follows:

$$\mathcal{L}_{cls} = -\sum_{x \in X^s} \log \frac{\exp\langle \mathcal{M}(f(x)), a_y \rangle}{\sum_{\hat{y} \in \mathcal{Y}^S} \exp\langle \mathcal{M}(f(x)), a_{\hat{y}} \rangle} \quad (1)$$

where the mapping function $\mathcal{M}(\cdot)$ is generally implemented via global average pooling (GAP) and linear projection. $\langle \cdot \rangle$ represents the cosine similarity used for category decision.

Concept-Harmonized Exploration

As shown in Figure 2, taking visual image patches E and the initialized visual prompt P as input, inspired by (Jia et al. 2022), we extract visual tokens $\bar{E} \in \mathbb{R}^{N_v \times D}$ from patches E conditioned on P using the transformer layers of ViT (Dosovitskiy et al. 2021) and produce the corresponding prompt embedding $\bar{P} \in \mathbb{R}^{T \times D}$. Similarly, we can obtain the embedded sharing attribute token $S \in \mathbb{R}^{N_s \times D}$ by encoding the word vectors of each attribute descriptor with GloVe (Pennington, Socher, and Manning 2014).

Concept-Aware Attention As illustrated in Figure 2, the sharing attributes are incorporated with the prompt-based visual features to provide hybrid multimodal information, serving as the foundation for semantic-enhanced prompt learning. However, visual images and textual descriptions inherently differ in terms of semantic levels and granularity (Shao et al. 2022; Chen et al. 2023b). Without explicit visual-semantic alignment, the attribute-related features are prone to sub-optimal representation. Thus, the

concept-aware attention (CAA) is proposed to harmonize the diverse representations across two modalities.

For the text modality, CAA learns attentive semantic representations grounded on a modal-sharing token R. Specifically, the attribute text S is regarded as the query to select distinct sharing tokens related to the attribute and generate the following output:

$$\tilde{S} = \mathrm{softmax}(\mathrm{GMP}(Q_S \cdot K_R^{\mathsf{T}})) \cdot V_R \tag{2}$$

where $Q.$, $K.$, and $V.$ are the linear mapping functions for the query, key, and value. R is the modal-sharing token with a set of learnable parameters. GMP denotes the max-pooling operation posed on the token-level relevance between S and R. This operation helps eliminate the influence of irrelevant noisy tokens that exhibit low relevance to modal-sharing tokens, producing a refined and compact concept-harmonized attribute token $\tilde{S}$.

Similarly, for the visual modality, within CAA, the visual feature $\bar{E}$ serves as the query, and R serves as both the key and value:

$$\tilde{E} = \mathrm{softmax}(\mathrm{GMP}(Q_{\bar{E}} \cdot K_R^{\mathsf{T}})) \cdot V_R \tag{3}$$

As a result, both image and attribute features are represented as combinations of the common modal-sharing token R. This explicit alignment ensures that the concepts of cross-modal information $\tilde{S}$ and $\tilde{E}$ are harmonized, which facilitates the subsequent prompt enhancement with a narrowed cross-modal gap.

Semantic-Enhanced Prompt Distillation

Based on the concept-harmonized outputs $\tilde{S}$ and $\tilde{E}$, these features help reveal a more comprehensive representation from both visual and semantic perspectives. Here, we apply the visual residual refinement to distill the semantic-enhanced prompt by predicting the details specific to attribute clues.

Visual Residual Refinement Unit (VRRU) Instead of relying on the commonly used attention mechanism, which has high computational complexity, VRRU adopts a simpler yet effective approach inspired by LLaVa (Liu et al. 2024a). This unit utilizes a straightforward linear layer to establish a connection between the image embedding and the attribute embedding. Specifically, we first concatenate the outputs of CAA, *i.e.*, $\tilde{S}$ and $\tilde{E}$, to fuse multiple modality features. Then, we employ a specialized linear prediction layer to estimate residual details. As a result, we have the following output:

$$Z = ZLinear\left(\left[\tilde{S}, \tilde{E}\right]\right) \tag{4}$$

where $ZLinear$ function serves as a predictor with its weights initially set to zero (Zhang, Rao, and Agrawala 2023), effectively eliminating harmful noise at the beginning of the training process.

Enhancing Prompt for Visual Enrichment To ensure the predicted residual Z carries meaningful and semantically consistent information, we employ an attribute-guided optimization strategy by aligning Z with category attribute prototypes, denoted as a_y, through a consistency loss $\mathcal{L}_{cons}$:

$$\mathcal{L}_{cons} = \|Z - \mathrm{MLP}(a_y)\| \tag{5}$$

where MLP refers to a multi-layer perceptron that projects the attribute prototypes into the same space as Z. This consistency loss function encourages Z to capture attribute-specific features that are crucial for ZSL tasks.

Then, Z is utilized to enhance the visual prompt by merging it back to the original prompt embedding via a skip connection:

$$\tilde{P} = [\bar{p}_1 + Z, \bar{p}_2 + Z, ..., \bar{p}_T + Z] \tag{6}$$

Z is added to each token ($\bar{p}_i$) of the original prompt, which can augment each prompt token with the attribute-aligned information from Z. The generated semantic-enhanced prompt $\tilde{P}$ incorporates attribute-related details derived from the multimodal information, allowing it to capture fine-grained semantics for distinguishing between similar classes in zero-shot scenarios. Finally, we integrate the enhanced prompt $\tilde{P}$ with the visual token $\bar{E}$ to enrich the visual representation $f(x)$ for x:

$$f(x) = \left[\tilde{P}, \bar{E}\right] \tag{7}$$

Together with the enhanced prompt, AENet is expected to achieve improved comprehensive transferable knowledge discovery, containing complementary information that may be overlooked by conventional prompt-based visual content.

Model Optimization and Inference

Optimization The overall objective loss function of AENet is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{cons}\mathcal{L}_{cons} + \lambda_{deb}\mathcal{L}_{deb} \tag{8}$$

where $\mathcal{L}_{cls}$ is the classification loss (Eq. (1)). λ_{cons} and λ_{deb} are the hyper-parameters controlling the weights of semantic consistency loss $\mathcal{L}_{cons}$ and the debiasing loss $\mathcal{L}_{deb}$, respectively. In addition, as expressed in Eq. (8), we also apply a debiasing loss $\mathcal{L}_{deb}$ to mitigate the seen-unseen bias following (Liu et al. 2023a, 2024b). It aims to balance the score dependency in the seen-unseen domain, pursuing the distribution consistency concerning both mean and variance:

$$\mathcal{L}_{deb} = \|\alpha_s - \alpha_u\|_2^2 + \|\beta_s - \beta_u\|_2^2 \tag{9}$$

where α_s and β_s represent the mean and variance, respectively, of the seen prediction score $\left\langle \mathcal{M}(f(x)), a_{\hat{y}(\hat{y}\in\mathcal{Y}^s)} \right\rangle$. Similarly, α_u and β_u denote the mean and variance, respectively, of the unseen prediction score $\left\langle \mathcal{M}(f(x)), a_{\hat{y}(\hat{y}\in\mathcal{Y}^u)} \right\rangle$.

Inference During training, the model merely learns about the knowledge of seen categories, while unseen categories are inferred at testing time:

$$\tilde{y} = \arg\max_{\hat{y}\in\mathcal{Y}^u} \left(\langle \mathcal{M}(f(x)), a_{\hat{y}} \rangle\right) \tag{10}$$

In the GZSL setting, both seen and unseen categories are encompassed. To jointly define the category, calibrated stacking (CS) (Chao et al. 2016) is applied:

$$\tilde{y} = \arg\max_{\hat{y}\in\mathcal{Y}} \left(\langle \mathcal{M}(f(x)), a_{\hat{y}} \rangle - \gamma \mathbb{I}_{[\hat{y}\in\mathcal{Y}^s]}\right) \tag{11}$$

where $\mathbb{I}_{\mathcal{Y}^S}(\cdot)$ represents an indicator function, yielding a result of 1 when $\hat{y} \in \mathcal{Y}^S$ and 0 otherwise. The calibrated factor γ is employed to trade off the calibration degree on seen categories and determine the category $\tilde{y}$ of an input visual sample x.

Methods	Venue	CUB				SUN				AwA2			
		ZSL	GZSL			ZSL	GZSL			ZSL	GZSL		
		acc	U	S	H	acc	U	S	H	acc	U	S	H
Generative-based Methods													
Composer (Huynh and Elhamifar 2020a)	NeurIPS'20	69.4	56.4	63.8	59.9	62.6	55.1	22.0	31.4	71.5	62.1	77.3	68.8
GCM-CF (Yue et al. 2021)	CVPR'21	–	61.0	59.7	60.3	–	47.9	37.8	42.2	–	60.4	75.1	67.0
SDGZSL (Chen et al. 2021c)	ICCV'21	75.5	59.9	66.4	63.0	–	–	–	–	72.1	64.6	73.6	68.8
CE-GZSL (Han et al. 2021)	CVPR'21	77.5	63.9	66.8	65.3	63.3	48.8	38.6	43.1	70.4	63.1	78.6	70.0
ICCE (Kong et al. 2022)	CVPR'22	78.4	67.3	65.5	66.4	–	–	–	–	72.7	65.3	82.3	72.8
FREE (Chen et al. 2021a)	ICCV'21	–	55.7	59.9	57.7	–	47.4	37.2	41.7	–	60.4	75.4	67.1
HSVA (Chen et al. 2021b)	NeurIPS'21	62.8	52.7	58.3	55.3	63.8	48.6	39.0	43.3	–	59.3	76.6	66.8
LBP (Lu et al. 2021)	TPAMI'21	61.9	42.7	71.6	53.5	63.2	39.2	36.9	38.1	–	–	–	–
f-VAEGAN+DSP (Chen et al. 2023a)	ICML'23	62.8	62.5	73.1	67.4	<u>68.6</u>	<u>57.7</u>	41.3	<u>48.1</u>	71.6	63.7	**88.8**	<u>74.2</u>
SHIP† (Wang et al. 2023)	ICCV'23	–	55.3	58.9	57.1	–	–	–	–	–	–	–	–
Embedding-based Methods													
APN (Xu et al. 2020)	NeurIPS'20	72.0	65.3	69.3	67.2	61.6	41.9	34.0	37.6	68.4	57.1	72.4	63.9
DAZLE (Huynh and Elhamifar 2020b)	CVPR'20	66.0	56.7	59.6	58.1	59.4	52.3	24.3	33.2	67.9	60.3	75.7	67.1
DVBE (Min et al. 2020)	CVPR'20	–	53.2	60.2	56.5	–	45.0	37.2	40.7	–	63.6	70.8	67.0
GEM-ZSL (Liu et al. 2021)	CVPR'21	77.8	64.8	<u>77.1</u>	70.4	62.8	38.1	35.7	36.9	67.3	64.8	77.5	70.6
DPPN (Wang et al. 2021)	NeurIPS'21	77.8	<u>70.2</u>	<u>77.1</u>	73.5	61.5	47.9	35.8	41.0	73.3	63.1	<u>86.8</u>	73.1
GNDAN (Chen et al. 2022c)	TNNLS'22	75.1	69.2	69.6	69.4	65.3	50.0	34.7	41.0	71.0	60.2	80.8	69.0
CLIP (Radford et al. 2021)	ICML'21	–	55.2	54.8	55.0	–	–	–	–	–	–	–	–
CoOP† (Zhou et al. 2022b)	IJCV'22	–	49.2	63.8	55.6	–	–	–	–	–	–	–	–
MSDN (Chen et al. 2022d)	CVPR'22	76.1	68.7	67.5	68.1	65.8	52.2	34.2	41.3	70.1	62.0	74.5	67.7
TransZero (Chen et al. 2022b)	AAAI'22	76.8	69.3	68.3	68.8	65.6	52.6	33.4	40.8	70.1	61.3	82.3	70.2
TransZero++ (Chen et al. 2022a)	TPAMI'22	78.3	67.5	73.6	70.4	67.6	48.6	37.8	42.5	72.6	64.6	82.7	72.5
DUET*† (Chen et al. 2023c)	AAAI'23	72.3	62.9	72.8	67.5	64.4	45.7	<u>45.8</u>	45.8	69.9	63.7	84.7	72.7
I2MVFormer* (Naeem et al. 2023)	CVPR'23	42.1	32.4	63.1	42.8	–	–	–	–	<u>73.6</u>	<u>66.6</u>	82.9	73.8
ZSLViT* (Chen et al. 2024)	CVPR'24	<u>78.9</u>	69.4	**78.2**	<u>73.6</u>	68.3	45.9	**48.3**	47.3	70.2	66.1	84.6	<u>74.2</u>
AENet* (Ours)	–	**80.3**	**73.1**	76.4	**74.7**	**70.4**	**58.6**	45.2	**51.0**	**75.2**	**70.3**	80.1	**74.9**

Table 1: Results (%) of the state-of-the-art ZSL and GZSL models on CUB, SUN, and AwA2, including both generative and embedding-based methods. The symbol "$*$" denotes ViT-based methods. The symbol "†" indicates the prompt-based methods, with results reported in (Chen et al. 2024).

Experiments

Experimental Settings

Datasets We conduct experiments on three standard benchmark datasets: Caltech-UCSD Birds-200-2011 (CUB) (Welinder et al. 2010), SUN Attribute (SUN) (Patterson and Hays 2012), Animals with Attributes2 (AwA2) (Xian et al. 2019). The categorization into seen and unseen categories follows the Proposed Split (PS) (Xian et al. 2019). The CUB dataset consists of 11,788 images illustrating 200 bird classes, with a split of 150/50 for seen/unseen classes and characterized by 312 attributes. SUN is a vast scene dataset that contains 14,340 images spanning 717 classes, divided into seen/unseen classes at 645/72 and annotated with 102 attributes. AwA2 contains 37,322 images of 50 animal classes, with a 40/10 split for seen/unseen classes, and is described by 85 attributes.

Evaluation Metrics We evaluate top-1 accuracy in both the ZSL and GZSL settings. For ZSL, we calculate accuracy solely on unseen classes, denoted as acc. In the GZSL setting, following (Xian et al. 2019), we employ the harmonic mean (as $H = 2 \times S \times U/(S+U)$) to measure performance, where S and U represent the top-1 accuracy of the seen and unseen classes, respectively.

Implementation Details we apply the ViT-Base model (Dosovitskiy et al. 2021) as visual feature extractor. The input image resolution is 224×224, with a patch size of 16×16. Our framework is implemented using PyTorch and executed on an NVIDIA GeForce RTX 3090 GPU.

Comparison with State-of-the-Art Methods

We evaluate our AENet and compare it with recently state-of-the-art methods. The results are presented in Table 1.

Results of ZSL For conventional ZSL, our method demonstrates significant acc improvements of 1.4%, 1.8%, and 1.6% on CUB, SUN, and AwA2 datasets, surpassing the previous state-of-the-art methods (Chen et al. 2024, 2023a; Naeem et al. 2023). AENet facilitates more comprehensive knowledge transference for unseen classification, thus achieving state-of-the-art accuracy performance of 80.3%, 70.4%, and 75.2% on CUB, SUN, and AwA2, respectively. Compared to recent methods (Liu et al. 2021; Chen et al.

Methods	CUB				SUN				AwA2			
	ZSL	GZSL			ZSL	GZSL			ZSL	GZSL		
	acc	U	S	H	acc	U	S	H	acc	U	S	H
AENet w/o prompt P	70.2	67.0	69.1	68.0	64.5	53.8	25.8	34.8	70.7	61.0	88.2	72.1
AENet w/o residual details Z	75.8	72.3	72.7	72.5	68.0	58.1	38.5	46.2	73.2	65.5	85.1	74.0
AENet w/o concept-aware attention	78.8	75.6	72.1	73.8	68.8	43.6	54.9	48.6	73.7	66.8	81.8	73.5
AENet (w/ all)	80.3	73.1	76.4	74.7	70.4	58.6	45.2	51.0	75.2	70.3	80.1	74.9

Table 2: Ablation study of AENet under the ZSL and GZSL settings on CUB, SUN, and AwA2 datasets.

Methods	CUB		AwA2	
	acc	H	acc	H
Linear + Skip	79.8	74.5	74.1	74.5
MLP + Skip	79.5	74.0	74.8	74.3
ZLinear + Gated Skip	79.5	74.4	73.5	74.4
ZLinear + Skip	80.3	74.7	75.2	74.9

Table 3: Ablation study of different implementations for semantic-enhanced prompt distillation.

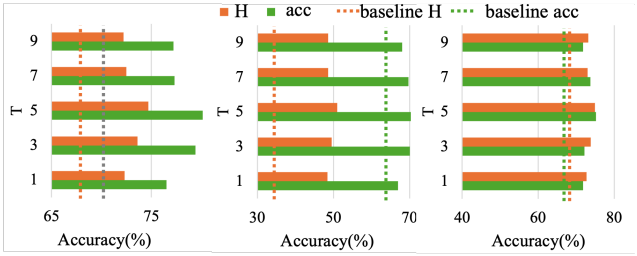

Figure 3: Effect of the length (T) of P on CUB, SUN and AwA2 datasets.

2022b,a, 2024) that refine visual features obtained from pre-trained visual encoders using attention mechanisms, AENet achieves substantial improvements in acc, with gains exceeding 1.4%, 2.1%, and 2.6% on CUB, SUN, and AwA2, respectively. Furthermore, our approach demonstrates superior performance compared to the prompt-based ZSL method (Chen et al. 2023c), yielding obvious improvements of 8.0%, 6.0%, and 5.3% on the CUB, SUN, and AwA2 datasets, respectively.

Results of GZSL Table 1 also reports the results of various methods in the GZSL setting. Results show that AENet can achieve state-of-the-art H performance across all datasets, *i.e.*, 74.7%, 51.0%, and 74.9% on CUB, SUN, and AwA2, respectively. Similar to ZSL, our AENet also significantly outperforms other prompt-based methods (Zhou et al. 2022b; Wang et al. 2023; Chen et al. 2023c) with substantial margins of 7.2%, 5.2%, and 2.2% on CUB, SUN, and AwA2 datasets, respectively. These results demonstrate that AENet significantly enhances cross-domain transferability by enriching the visual representations.

Ablation Study

Effect of components in AENet As shown in Table 2, we evaluate key components in AENet, *i.e.*, the learnable prompt P, semantic-related details Z for the semantic-enhanced prompt, and concept-aware attention for the concept-harmonized token generation. When the original learnable prompt is removed, the model can be regarded as the baseline that directly uses the visual features extracted from vanilla ViT and projects them into semantic space for category inference. AENet outperforms the baseline with large acc/H margins of 20.1%/6.7%, 5.9%/6.2%, and 4.5%/2.8% on CUB, SUN, and AwA2 datasets, respectively. The performance of AENet degrades when residual details are removed, as evidenced by decreases in both ZSL and GZSL across multiple datasets: specifically, a reduction in acc/H of 4.5%/2.2% on CUB, 2.4%/4.8% on SUN, and 2.0%/0.9% on AwA2, compared to the full model implementation. Additionally, concept-aware attention provides concept-harmonized visual and semantic representations, resulting in acc/H improvements of 1.5%/0.9%, 1.6%/2.4%, 1.5%/1.4% on CUB, SUN, and AwA2 datasets, respectively. Additionally, we implement variants by replacing ZLinear in Eq. (4) with MLP and a linear layer, and the skip connection in Eq. (6) with a gated skip. As demonstrated in Table 3, these more complex alternatives do not offer performance improvements as the harmonized tokens are concise and informative. Moreover, ZLinear consistently achieves better performance compared to the linear.

Effect of T T is the length of the learnable visual prompt P. Here, we sweep prompt length $T \in \{1, 3, 5, 7, 9\}$ to investigate the effect of the prompts P on classification performance. Figure 3 shows the values of acc and H as T varies. We can observe that the best performance is achieved when T is approximately 5. Notably, even with a single prompt, AENet demonstrates significant performance improvements over the baseline ($1st$ row in Table 2). Based on these findings, we set T = 5 for CUB, SUN, and AwA2 datasets.

Impact of λ_{cons} and λ_{deb} In this work, we combine the attribute consistency loss $\mathcal{L}_{cons}$ and $\mathcal{L}_{deb}$ and debiasing with the balance hyper-parameters of λ_{cons} and λ_{deb} respectively to train the model. Here, we study the impact of λ_{cons} and λ_{deb} as shown in Figure 4. As λ_{cons} rises from 0.0 to 2.0, *i.e.*, the attribute-supervision $\mathcal{L}_{cons}$ is introduced into AENet for the semantic prediction, H increases on all datasets. The best H is obtained when $\lambda_{cons} = 1.0$. This demon-

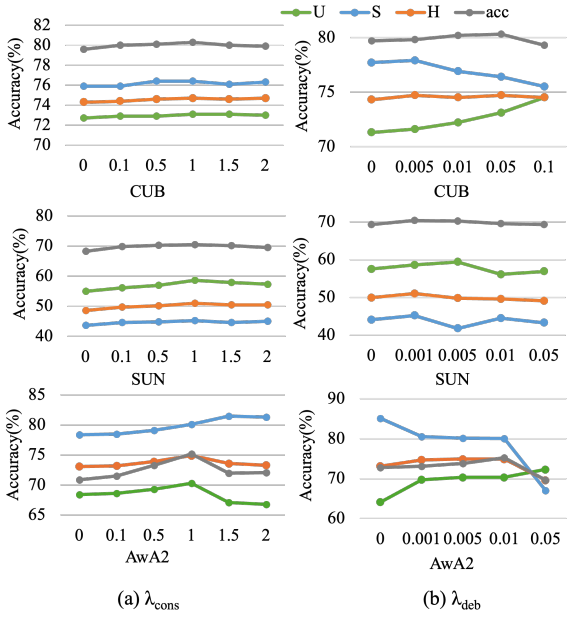

Figure 4: Effect of loss hyper-parameters.

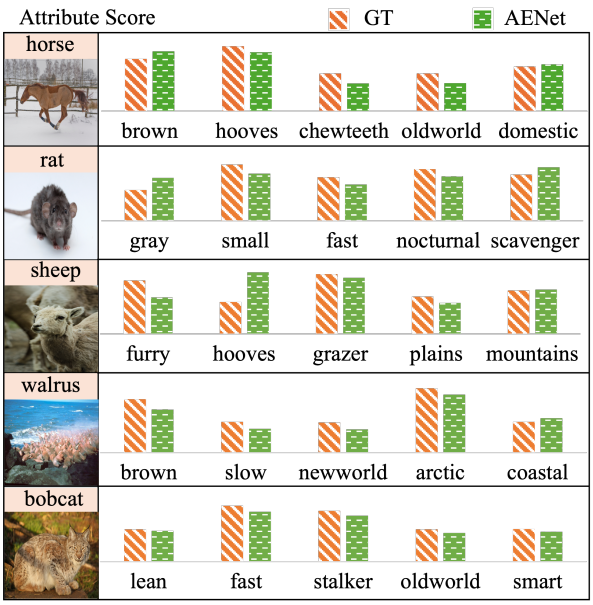

Figure 5: Attribute prediction for visualization.

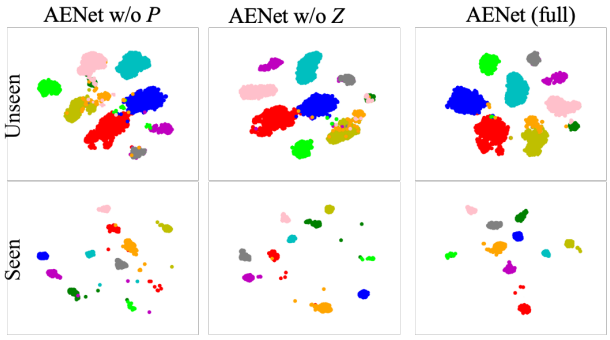

Figure 6: t-SNE visualizations of visual features for seen classes and unseen classes. The 10 colors denote different classes randomly selected from the AwA2 dataset.

strates the effectiveness of semantic consistency in predicting semantically relevant details, closely aligning with category prototypes to focus on attribute-specific features. When $\lambda_{cons} > 1.0$, H decreases. Thus, we set $\lambda_{cons} = 1.0$ for optimal results. Besides, increasing λ_{deb} leads to more consistent distribution between seen and unseen predictions, improving unseen accuracy U and overall H performance.

Visualization

Attribute Prediction To intuitively validate the capability of AENet for capturing enhanced semantic information necessary for unseen classes, we input unseen images into AENet to predict the attribute scores ($\mathcal{M}(f(x))$). As shown in Figure 5, we can observe that AENet demonstrates strong performance in predicting attributes across various animal species in the AwA2 dataset. For example, for the horse, AENet closely matches the ground truth (GT) on "brown", "hooves", and "domestic", with slight discrepancies in "chewteeth" and "oldworld". The rat predicted scores of the rat are remarkably accurate across all attributes, showcasing precision for this species. These results illustrate AENet's robust capability in capturing enhanced semantic information across diverse species, which is crucial for generalizing to unseen classes in ZSL tasks.

T-SNE Visualization As shown in Figure 6, we present the t-SNE visualization of visual features for seen and unseen classes. Compared to full AENet, visual features extracted from the AENet w/o P lack distinctiveness within certain classes. This intuitively suggests that the prompt plays a crucial role in enabling the application of pre-trained ViT to downstream ZSL tasks by obtaining high-quality features for seen classes. Furthermore, compared to AENet w/o Z, the visual features learned by our full AENet showcase desirable distinguishability with higher inter-class discrepancy and clearer decision boundaries.

Conclusion

This paper introduces AENet, which leverages prompt learning and addresses the challenge of over-emphasizing primary visual features from seen domains. AENet comprises two key steps. Firstly, the concept-harmonized tokens are explored for visual and attribute modalities based on the modal-sharing token with consistent concepts. Secondly, by investigating the semantic-related details, the VRRU is proposed to distill the semantic-enhanced prompt integrated with primary visual features to attend to semantic-related information for visual enrichment. Consequently, AENet can achieve superior visual-semantic alignment for generalizing to unseen classes. Extensive experiments across three benchmark datasets show the superiority of AENet.

Acknowledgments

This work was supported in part by Fundamental Research Funds for the Central Universities (2024JBZY001, JZ2024HGTB0255, 2024XKRC082), National Natural Science Foundation of China (No. 62331003, 62120106009, 62302141, 62476021, 72434005, 92470203), Joint Funds of the National Natural Science Foundation of China under Grant U23A20314, Beijing Natural Science Foundation (L223022, L242022), Natural Science Foundation of Hebei Province (F2024105029), the Chinese Association for Artificial Intelligence (CAAI)-Compute Architecture for Neural Networks (CANN) Open Fund, developed on OpenI Community.

References

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2013. Label-embedding for attribute-based classification. In *CVPR*, 819–826.

Chao, W.-L.; Changpinyo, S.; Gong, B.; and Sha, F. 2016. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*.

Chen, S.; Hong, Z.; Hou, W.; Xie, G.-S.; Song, Y.; Zhao, J.; You, X.; Yan, S.; and Shao, L. 2022a. TransZero++: Cross Attribute-Guided Transformer for Zero-Shot Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–17.

Chen, S.; Hong, Z.; Liu, Y.; Xie, G.-S.; Sun, B.; Li, H.; Peng, Q.; Lu, K.; and You, X. 2022b. TransZero: Attribute-guided Transformer for Zero-Shot Learning. In *AAAI*, 330–338.

Chen, S.; Hong, Z.; Xie, G.; Peng, Q.; You, X.; Ding, W.; and Shao, L. 2022c. GNDAN: Graph Navigated Dual Attention Network for Zero-Shot Learning. *IEEE Transactions on neural networks and learning systems*.

Chen, S.; Hong, Z.; Xie, G.; Wang, W.; Peng, Q.; Wang, K.; Zhao, J.; and You, X. 2022d. MSDN: Mutually Semantic Distillation Network for Zero-Shot Learning. In *CVPR*, 7612–7621.

Chen, S.; Hou, W.; Hong, Z.; Ding, X.; Song, Y.; You, X.; Liu, T.; and Zhang, K. 2023a. Evolving Semantic Prototype Improves Generative Zero-Shot Learning. In *ICML*.

Chen, S.; Hou, W.; Khan, S.; and Khan, F. S. 2024. Progressive Semantic-Guided Vision Transformer for Zero-Shot Learning. In *CVPR*.

Chen, S.; Wang, W.; Xia, B.; Peng, Q.; You, X.; Zheng, F.; and Shao, L. 2021a. FREE: Feature Refinement for Generalized Zero-Shot Learning. In *ICCV*, 122–131.

Chen, S.; Xie, G.-S.; Yang Liu, Y.; Peng, Q.; Sun, B.; Li, H.; You, X.; and Shao, L. 2021b. HSVA: Hierarchical Semantic-Visual Adaptation for Zero-Shot Learning. In *NeurIPS*, 16622–16634.

Chen, Y.; Yuan, J.; Tian, Y.; Geng, S.; Li, X.; Zhou, D.; Metaxas, D. N.; and Yang, H. 2023b. Revisiting multimodal representation in contrastive learning: from patch and token embeddings to finite discrete tokens. In *CVPR*, 15095–15104.

Chen, Z.; Huang, Y.; Chen, J.; Geng, Y.; Zhang, W.; Fang, Y.; Pan, J. Z.; Song, W.; and Chen, H. 2023c. DUET: Cross-modal Semantic Grounding for Contrastive Zero-shot Learning. In *AAAI*, 405–413.

Chen, Z.; Luo, Y.; Qiu, R.; Wang, S.; Huang, Z.-Y.; Li, J.; and Zhang, Z. 2021c. Semantics Disentangling for Generalized Zero-Shot Learning. In *ICCV*, 8712–8720.

Cheng, D.; Huang, S.; Bi, J.; Zhan, Y.; Liu, J.; Wang, Y.; Sun, H.; Wei, F.; Deng, D.; and Zhang, Q. 2023. Uprise: Universal prompt retrieval for improving zero-shot evaluation. *arXiv preprint arXiv:2303.08518*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Han, Z.; Fu, Z.; Chen, S.; and Yang, J. 2021. Contrastive Embedding for Generalized Zero-Shot Learning. In *CVPR*, 2371–2381.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hu, Y.; Wen, G.; Chapman, A.; Yang, P.; Luo, M.; Xu, Y.; Dai, D.; and Hall, W. 2021. Graph-based visual-semantic entanglement network for zero-shot image recognition. *IEEE Transactions on Multimedia*, 24: 2473–2487.

Huynh, D.; and Elhamifar, E. 2020a. Compositional Zero-Shot Learning via Fine-Grained Dense Feature Composition. In *NeurIPS*, 19849–19860.

Huynh, D.; and Elhamifar, E. 2020b. Fine-Grained Generalized Zero-Shot Learning via Dense Attribute-Based Attention. In *CVPR*, 4482–4492.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *ECCV*, 709–727. Springer.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *NeurIPS*, 35: 22199–22213.

Kong, X.; Gao, Z.; Li, X.; Hong, M.; Liu, J.; Wang, C.; Xie, Y.; and Qu, Y. 2022. En-Compactness: Self-Distillation Embedding & Contrastive Generation for Generalized Zero-Shot Learning. In *CVPR*, 9306–9315.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 951–958.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024a. Visual instruction tuning. *NeurIPS*, 36.

Liu, M.; Li, F.; Zhang, C.; Wei, Y.; Bai, H.; and Zhao, Y. 2023a. Progressive Semantic-Visual Mutual Adaption for Generalized Zero-Shot Learning. In *CVPR*, 15337–15346.

Liu, M.; Zhang, C.; Bai, H.; and Zhao, Y. 2024b. Part-Object Progressive Refinement Network for Zero-Shot Learning. *IEEE Transactions on Image Processing*.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; and Tang, J. 2023c. GPT understands, too. *AI Open*.

Liu, Y.; Zhou, L.; Bai, X.; Huang, Y.; Gu, L.; Zhou, J.; and Harada, T. 2021. Goal-Oriented Gaze Estimation for Zero-Shot Learning. In *CVPR*, 3794–3803.

Lu, Z.; Guan, J.; Li, A.; Xiang, T.; Zhao, A.; and Wen, J.-R. 2021. Zero and Few Shot Learning With Semantic Feature Synthesis and Competitive Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43: 2510–2523.

Lyu, J.; Lan, X.; Hu, G.; Jiang, H.; Gan, W.; and Xue, J. 2024. ETAU: Towards Emotional Talking Head Generation Via Facial Action Unit. In *ICME*, 1–6.

Min, S.; Yao, H.; Xie, H.; Wang, C.; Zha, Z.; and Zhang, Y. 2020. Domain-Aware Visual Bias Eliminating for Generalized Zero-Shot Learning. In *CVPR*, 12661–12670.

Naeem, M. F.; Khan, M. G. Z. A.; Xian, Y.; Afzal, M. Z.; Stricker, D.; Van Gool, L.; and Tombari, F. 2023. I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification. In *CVPR*, 15169–15179.

Palatucci, M.; Pomerleau, D.; Hinton, G. E.; and Mitchell, T. M. 2009. Zero-shot learning with semantic output codes. In *NeurIPS*.

Patterson, G.; and Hays, J. 2012. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2751–2758.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*, 1532–1543.

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

Shao, Z.; Zhang, X.; Fang, M.; Lin, Z.; Wang, J.; and Ding, C. 2022. Learning granularity-unified representations for text-to-image person re-identification. In *ACM MM*, 5566–5574.

Tan, C.; Tao, R.; Liu, H.; Gu, G.; Wu, B.; Zhao, Y.; and Wei, Y. 2024. C2P-CLIP: Injecting Category Common Prompt in CLIP to Enhance Generalization in Deepfake Detection. *arXiv preprint arXiv:2408.09647*.

Torralba, A.; and Efros, A. A. 2011. Unbiased look at dataset bias. In *CVPR*.

Wang, C.; Min, S.; Chen, X.; Sun, X.; and Li, H. 2021. Dual Progressive Prototype Network for Generalized Zero-Shot Learning. In *NeurIPS*, 2936–2948.

Wang, Y.; Cheng, L.; Fang, C.; Zhang, D.; Duan, M.; and Wang, M. 2024. Revisiting the Power of Prompt for Visual Tuning. In *ICML*.

Wang, Z.; Liang, J.; He, R.; Xu, N.; Wang, Z.; and Tan, T. 2023. Improving zero-shot generalization for clip with synthesized prompts. In *ICCV*, 3032–3042.

Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S. J.; and Perona, P. 2010. Caltech-UCSD Birds 200. *Technical Report CNS-TR-2010-001, Caltech,*.

Xian, Y.; Akata, Z.; Sharma, G.; Nguyen, Q.; Hein, M.; and Schiele, B. 2016. Latent Embeddings for Zero-Shot Classification. In *CVPR*, 69–77.

Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2019. Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41: 2251–2265.

Xie, G.-S.; Liu, L.; Jin, X.; Zhu, F.; Zhang, Z.; Qin, J.; Yao, Y.; and Shao, L. 2019. Attentive Region Embedding Network for Zero-Shot Learning. In *CVPR*, 9376–9385.

Xie, G.-S.; Liu, L.; Jin, X.; Zhu, F.; Zhang, Z.; Yao, Y.; Qin, J.; and Shao, L. 2020. Region Graph Embedding Network for Zero-Shot Learning. In *ECCV*, 562–580.

Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; and Akata, Z. 2020. Attribute Prototype Network for Zero-Shot Learning. In *NeurIPS*, 21969–21980.

Yue, Z.; Wang, T.; Zhang, H.; Sun, Q.; and Hua, X. 2021. Counterfactual Zero-Shot and Open-Set Visual Recognition. In *CVPR*, 15404–15414.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*, 3836–3847.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *CVPR*, 16816–16825.

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

Zhu, Y.; Xie, J.; Tang, Z.; Peng, X.; and Elgammal, A. 2019. Semantic-Guided Multi-Attention Localization for Zero-Shot Learning. In *NeurIPS*.