SAGE: Structured Attribute-Guided Enhancement for GZSL

Zao Zhang, Liguo Sun, Pin Lyu

Institute of Automation, Chinese Academy of Sciences
No. 95 Zhongguancun East Road, Haidian District, Beijing 100190, China
{zao.zhang, liguo.sun, pin.lv}@ia.ac.cn

Abstract

Embedding-based generalized zero-shot learning (GZSL) models often first forge robust latent semantic correlations between visual and attribute features so that knowledge can generalize to unseen categories. Despite leveraging attributes as priors and learning a shared embedding space, current methods exhibit two critical flaws. First, attributes with heterogeneous granularity are treated uniformly, leading to semantic ambiguity. Second, the source of class-level misclassification seldom aligns with attribute-level errors, preventing models from targeting the specific attributes responsible. To overcome these limitations, we introduce Structured Attribute-Guided Enhancement (SAGE), a unified framework for GZSL. Consensus-aware bidirectional attention first synchronizes visual–semantic focus regions via a mutual-distillation scheme. Next, we partition all attributes into pairwise-disjoint subsets—Global, Context, and Local—and couple them with visual features extracted at matching spatial scales. Finally, we design a cross-sample, subset-aware distillation mechanism—when a sample is misclassified, SAGE identifies the culpable attribute subset, retrieves high-confidence prototypes from a memory bank, and applies a Kullback–Leibler (KL) divergence constraint to the corresponding feature branch. Comprehensive experiments and ablations on the challenging AwA2, CUB, and SUN benchmarks demonstrate the contribution of each component, with SAGE achieving a new state-of-the-art throughout. These findings underscore SAGE's robustness and versatility, marking a substantial advance in generalized zero-shot learning and paving the way for broader zero-resource recognition.

Introduction

Generalized zero-shot learning (GZSL) requires a model to recognize both seen and unseen classes at test time, thereby extending the conventional zero-shot setting. It removes the need for exhaustive manual labels by transferring knowledge from a semantic space—typically attribute vectors or textual descriptions—into the visual domain. Despite rapid progress, state-of-the-art (SOTA) GZSL methods still lag behind fully supervised systems, particularly on unseen classes with subtle appearance variations or in visually cluttered scenes. To bridge this visual–semantic gap, recent work employs embedding networks and vision–language transformers that learn to share a latent space. Yet two critical limitations remain largely unexplored, constraining the performance of current GZSL pipelines.

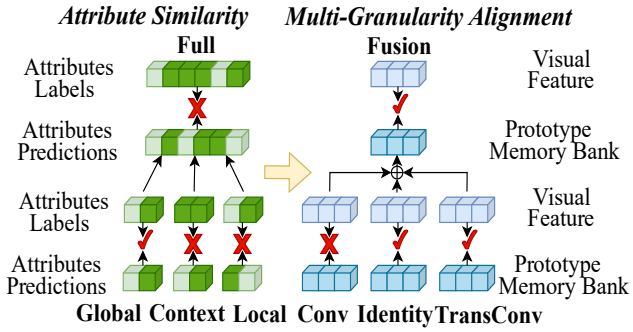


Figure 1: Attributes are partitioned into *Global*, *Context* and *Local* subsets (left), a *check mark* indicates the predicted attributes match the ground-truth label, whereas a *cross mark* denotes a mismatch. Three scale-matched visual branches align with the corresponding attribute subsets, and their fusion predicts the full attribute vector (right), a *check mark* here means a KL-divergence penalty is applied, while a *cross mark* means the opposite.

First, in most methods, the entire attribute set—ranging from abstract descriptors (fast, big) and contextual cues (swims, desert) to fine-grained appearance details (paws, furry)—is collapsed into a single flat vector. Such uniform treatment ignores the underlying semantic hierarchy, letting heterogeneous concepts interfere during training. The resulting entanglement yields ambiguous supervision, weakens attribute discriminability, and ultimately diminishes generalization capacity to unseen classes.

Second, existing models lack targeted guidance for the attributes responsible for misclassification. After a sample is mispredicted, most GZSL models offer no mechanism to localize the semantic portion that caused the error, the constraint remains confined to overall feature layers, hindering the correction of local misalignment or semantic drift. Consequently, hard samples—particularly those near class boundaries—repeatedly receive noisy gradients and fail to converge to discriminative representations.

To address these challenges, we present SAGE, a cross-attention framework that cements integration between visual and semantic representations. Built upon a pretrained Vision Transformer (ViT) (Dosovitskiy et al. 2021), we employ a multi-scale feature extractor (MSFE) to capture both global abstractions and localized fine-grained details. On the semantic side, we initialize attribute embeddings with Word2Vec (Mikolov et al. 2013a,b) to encode prior inter-attribute relationships. These embeddings are then refined by an attribute mixer with contrastive learning (AMCL), which not only enhances inter-class separability but also preserves informative intra-class structures.

To further bridge the semantic–visual gap, we introduce two complementary modules: semantic-guided visual attention (SVA) and vision-guided semantic attention (VSA). This cross-attention scheme enables dynamic information flow between modalities and strengthens mutual alignment. We also apply mutual distillation to synchronize the weights of SVA and VSA, enforcing consistency in both directions.

To decouple entangled attributes and investigate the causes of mispredictions, we propose memory-bank-guided fusion (MBGF). The core idea is illustrated in Fig. 1. We first divide the attribute vocabulary into three semantically coherent subsets—Global, Context, and Local. Each subset is paired with a visual branch operating at a matching spatial scale, obtained from the ViT backbone via convolution, identity mapping, and transposed convolution, respectively. During training, when the fused prediction misclassifies a sample, we inspect each attribute group individually. If a branch's predicted attributes favour an incorrect class over the ground-truth, we regard it as a group-specific failure. We subsequently retrieve a high-confidence prototype for the correct class and attribute group from a class-wise memory bank and impose a KL-divergence loss on that branch. This targeted intervention enhances feature discrimination while avoiding over-regularizing well-aligned branches.

The main contributions are summarised as follows:

- **Multi-granularity semantic–visual alignment.** We explicitly decompose the attribute space into Global, Context, and Local subsets and pair each with scale-matched visual branches, yielding the first GZSL framework that aligns semantic granularity with visual receptive fields end-to-end
- **Cross-attention with mutual distillation.** The cross-attention module lets semantics guide vision and vice-versa while a mutual-distillation loss enforces consensus, markedly reducing cross-modal bias without extra inference cost.
- **Memory-bank-guided error correction.** When a sample is misclassified, SAGE pinpoints the culpable attribute subset and pulls a high-confidence prototype from a class-wise memory bank, applying a targeted KL constraint that corrects only the problematic branch and avoids over-regularization.
- **Comprehensive state-of-the-art validation.** Extensive experiments and ablations on AwA2, CUB, and SUN show consistent gains on the harmonic mean, establishing new SOTA while providing detailed insights into how each component contributes to GZSL robustness.

Related Work

Generative-based ZSL vs Embedding-based ZSL

Zero-Shot Learning (ZSL) now follows two main primary paradigms—*generative-based* and *embedding-based*—both of which seek to classify previously unseen classes by leveraging auxiliary semantic information. While these approaches share a common goal, they differ considerably in the strategies they employ to bridge visual and semantic representations.

Generative approaches, often based on Generative Adversarial Networks (GANs)(Goodfellow et al. 2014) or Variational Autoencoders (VAEs)(Kingma and Welling 2014), synthesize visual features for unseen classes by conditioning on semantic descriptors. By producing pseudo-visual samples for unseen categories, the ZSL problem transforms into a more conventional supervised classification task. For instance, Maunil (Vyas, Venkateswara, and Panchanathan 2020) employ GANs to generate class-specific features guided by attribute vectors, effectively augmenting the feature space for novel classes. While such generative models can be powerful in mitigating data scarcity, they often struggle to produce high-fidelity features in domains with significant visual complexity or high intra-class variation.

Embedding-based methods, on the other hand, learn a shared embedding space that unifies both visual features and semantic representations, such as attribute-level word embeddings. Early work by Xie (Xie et al. 2019) and subsequent efforts by Xu (Xu et al. 2020) focus on projecting visual features into a space defined by semantic vectors, enabling recognition of unseen classes without requiring explicit sample generation. Although these methods offer a more direct approach—bypassing the complexity of synthesizing novel features—they often encounter a pronounced semantic-visual gap. Our approach addresses the limitations of static semantic embeddings by introducing a dynamic, learnable semantic transformation which effectively bridges the semantic-visual gap and yields stronger generalization in zero-shot scenarios.

GZSL vs ZSL

GZSL builds upon the ZSL paradigm by exposing the model to both seen and unseen categories at inference time, making it more representative of real-world scenarios. Unlike conventional ZSL, which presumes that all test instances belong strictly to unseen classes, GZSL introduces the critical challenge of preventing excessive bias toward the seen categories. This bias arises because the model has been trained extensively on seen classes, thereby predisposing it to classify ambiguous samples as one of the seen categories.

Early ZSL approaches (Xian et al. 2018) focus on learning a shared embedding space in which both visual features and semantic descriptors can be mapped. While these methods show promise on standard ZSL benchmarks, they tend to degrade in the GZSL setting due to an inherent preference for classes observed during training. Subsequent solutions (Mancini et al. 2021; Wang et al. 2023) address this is-

sue by introducing specialized loss functions or re-weighting schemes to reduce bias toward the seen classes. However, these methods often face performance degradation when a significant semantic or domain shift exists, particularly if the attribute distributions of unseen classes differ substantially from those of the seen categories. Recent work (Li et al. 2024b) on GZSL endeavors to learn a unified latent space that aligns both visual and semantic representations while mitigating domain shifts. (Chen et al. 2023b) address the issues of attribute imbalance and co-occurrence by contrastive learning. Although these methods can alleviate data scarcity by creating realistic representations for unseen categories, balancing performance across seen and unseen classes remains a difficult challenge.

Visual-Semantic Cross-Modal Learning

Recent developments in visual-semantic cross-modal learning emphasize the construction of shared embedding spaces to establish meaningful correspondences between heterogeneous modalities from two complementary perspectives: cross-modal representation learning and fine-grained alignment mechanisms.

The advent of deep learning has driven substantial progress in multi-modal architectures. Shi (Shi et al. 2019) pioneered a framework that aggregates image scene graphs, subsequently extracting frequently co-occurring concept pairs as generalized semantic units. Beyond deterministic mappings, probabilistic approaches such as PCME (Chun et al. 2021) encode cross-modal embeddings as Gaussian distributions, explicitly modelling uncertainty to enable one-to-many matching. Meanwhile, large-scale contrastive pre-training, exemplified by CLIP-based methods (Lin et al. 2023), demonstrates remarkable few-shot generalization by learning a unified embedding space through image-text alignment at scale.

Moving beyond global feature matching, recent work has explored fine-grained alignment to link local image regions with semantic tokens or linguistic units. Leveraging dynamic attention mechanisms, CHAN (Pan, Wu, and Zhang 2023) formulates an information-theoretic approach to image-text matching, isolating the most salient region-word pairs and discarding weakly informative alignments. TFS (Li et al. 2024a) further extends this paradigm by integrating prompts that effectively capture crucial visual signals in whole-slide images, resulting in enhanced representational strength and significantly improved generalization.

Despite these advancements, challenges persist in ensuring consistent cross-stream alignment and robust fine-grained matching, particularly for complex scenes characterized by rich contextual dependencies in GZSL. Under the GZSL scenario with cross-modal learning, PSMVA+ (Liu et al. 2025) utilizes selective cross-granularity learning to find more reliable granularity, but it still fails to effectively align the attribute with granularity. This limitation motivates our proposed SAGE framework, which integrates disentangling attribute supervision and memory-bank guided prototype distillation to address both coarse-grained and fine-grained alignment gaps.

Method

Problem Definition

In the standard ZSL paradigm, test queries are confined to the unseen label set $\mathcal{Y}^u$. In contrast, GZSL accommodates test samples from both *seen* and *unseen* classes, i.e., $\mathcal{Y}^s \cup \mathcal{Y}^u$.

Formally, let

$$\mathcal{D}^s = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathcal{X},\, y_i \in \mathcal{Y}^s\},$$

denote the training set, where $\mathbf{x}_i$ is the input data and y_i is its associated label from the seen category set $\mathcal{Y}^s$. At inference time, we aim to classify any sample that may originate from either $\mathcal{Y}^s$ or the disjoint unseen set $\mathcal{Y}^u$.

Because labels for $\mathcal{Y}^u$ are not available during training, each class $y \in \mathcal{Y}$ is associated with an auxiliary semantic representation $\mathbf{a}_y \in \mathcal{A}$. Concretely, we let

$$\mathcal{A} = \{\mathbf{a}_y \mid y \in \mathcal{Y}^s \cup \mathcal{Y}^u\}$$

be the collection of all attribute-level embeddings. The goal of the learned model is to utilize these semantic representations $\mathbf{a}_y$ to bridge the gap between seen and unseen categories.

Overall Framework

We propose a bidirectional cross-attention framework to tackle the challenges inherent in GZSL. Our approach integrates the ViT for extracting patch-level representations from each input image, alongside a collection of Word2Vec-based semantic embeddings $\{\mathbf{a}_y \mid y \in \mathcal{Y}\}$ that encode class-level or attribute-level information.

Fig. 2 illustrates the high-level pipeline. First, an input image is divided into multiple patches, which are then projected into a latent space by the ViT encoder. To capture both global context and fine-grained details, these patch embeddings are further processed by the MSFE module across multiple representation levels. Concurrently, the encoded class-attribute matrix undergoes transformation via the AMCL module, optimizing latent semantic representations for both class-level and attribute-level embeddings.

Next, we perform cross-attention between visual and semantic streams with the VSA module and the SVA module. This procedure is repeated three times with successive hidden states from the ViT, thereby accumulating contextual cues at different levels of abstraction. The outputs from each level and scale are subsequently merged in the MBGF, guided by a memory bank-guided fusion strategy. The fused visual feature ultimately serves as a basis for computing similarity with the target class embeddings, enabling classification.

Attribute-Mixer with Contrastive Learning

A critical challenge in ZSL/GZSL tasks lies in effectively modelling semantic representations that remain both *expressive* and *distinct* across various attribute-level and class-level embeddings. To address this, we develop an Attribute-Mixer module that leverages LayerNorm, as well as dimension-expanding and dimension-reducing fully connected layers,

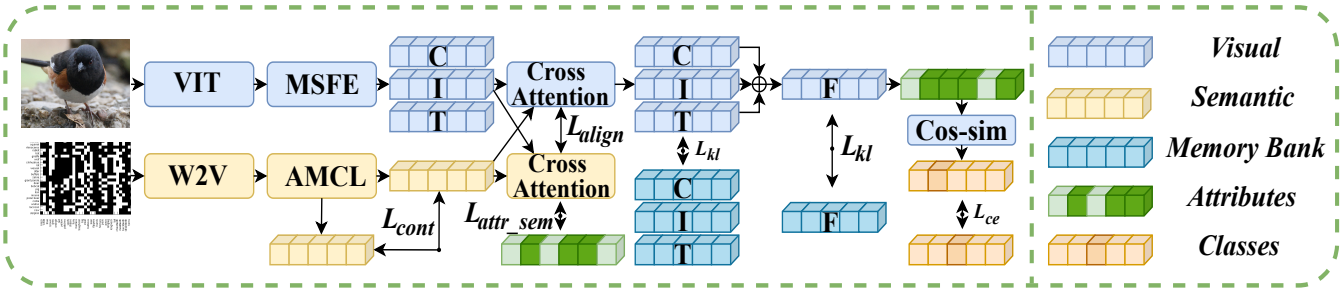


Figure 2: Overview of the Structured Attribute-Guided Enhancement Framework(SAGE).

to enhance the representational capacity of the semantic embeddings. As depicted in Fig. 3 and Fig. 4, the module applies LayerNorm followed by linear projections to learn implicit relationships among attributes, thus establishing semantic bridges that connect shared traits across different classes.

However, semantic overlaps and insufficient separability remain common pitfalls, especially when multiple categories share similar attributes. To mitigate this issue, we introduce a *contrastive learning* objective on the attribute-level embeddings. For a mini-batch $\mathcal{B}$, we define:

$$\mathcal{L}_{\text{cont\_attr}} = -\sum_{(i,j)\in\mathcal{B}} \log \frac{\exp\big(\text{sim}(\mathbf{a}_{y_i}, \mathbf{a}_{y_j})/\tau\big)}{\sum_{k\in\mathcal{B}} \exp\big(\text{sim}(\mathbf{a}_{y_i}, \mathbf{a}_{y_k})/\tau\big)}, \tag{1}$$

where $\mathbf{a}_{y_i}$ denotes the attribute embedding for class y_i, $\text{sim}(\cdot,\cdot)$ is a similarity function, and τ is a temperature parameter. Positive pairs (i,j) satisfy $y_i = y_j$, while negative pairs have $y_i \neq y_j$. This objective compels embeddings sharing the same attribute to be closer while pushing apart embeddings with different attributes, thereby improving attribute-level discrimination.

In parallel, we introduce an additional contrastive loss to ensure sufficient *class-level* separability:

$$\mathcal{L}_{\text{cont\_cls}} = -\sum_{(i,j)\in\mathcal{B}} \log \frac{\exp\big(\text{sim}(\mathbf{c}_{y_i}, \mathbf{c}_{y_j})/\tau\big)}{\sum_{k\in\mathcal{B}} \exp\big(\text{sim}(\mathbf{c}_{y_i}, \mathbf{c}_{y_k})/\tau\big)}, \tag{2}$$

where $\mathbf{c}_{y_i}$ is the class-level embedding for category y_i. This formulation ensures that classes with distinct semantic profiles are well-separated in the latent space.

Multi-scaled Feature Extraction

Given an input image, we first pass it through a pretrained ViT, which partitions the image into patches and generates patch-level embeddings $\{\mathbf{v}_p\}_{p=1}^{P}$. Each patch embedding has dimension d, yielding an overall embedding tensor of shape $P \times d$, where P is the total number of patches.

To obtain a richer set of representations, we further transform the patch embeddings across multiple scales, as illustrated in Figs. 3 and 4. Specifically, we apply (i) standard convolutions, (ii) identity mappings, and (iii) transposed convolutions to capture varying spatial resolutions. This procedure produces three feature maps:

$$\mathbf{F}^{\text{conv}} \in \mathbb{R}^{P^- \times d}, \quad \mathbf{F}^{\text{id}} \in \mathbb{R}^{P \times d}, \quad \mathbf{F}^{\text{trans}} \in \mathbb{R}^{P^+ \times d},$$

where P^-, P, and P^+ respectively denote different patch dimensions resulting from convolution, identity mapping, and transposed convolution.

By integrating multi-scale representations, our approach captures local and global cues in a complementary fashion.

Cross-modal Attention

To enable effective interaction between visual and semantic modalities, we design a dual-stream cross-attention mechanism comprising two complementary modules: VSA and SVA. This design allows visual representations to selectively focus on relevant semantic cues while simultaneously enabling semantic embeddings to be refined according to spatially grounded visual features.

In the VSA stream, as shown in Fig. 3, each semantic embedding $\mathbf{a}_y$ is treated as a query, while multi-scale visual features derived from MSFE serve as keys and values. The attention maps are further constrained by attribute supervision via:

$$\mathcal{L}_{\text{attr\_sem}} = \sum_{a=1}^{A} \big\|\text{MaxPool}\big(\mathbf{W}_s\big)^a - \mathbf{Attr}^a\big\|_2^2, \tag{3}$$

where $\mathbf{W}_s^a$ denotes the attention weight for attribute a, and $\mathbf{Attr}^a$ is the ground-truth attribute label.

In the SVA stream, as shown in Fig. 4, each patch-level visual feature $\mathbf{V}^m$ acts as the query, while the corresponding semantic vector $\mathbf{a}_y$ serves as both key and value. This direction ensures that semantic features selectively emphasize visual content that aligns with specific attributes. The supervision is imposed via:

$$\mathcal{L}_{\text{attr\_vis}} = \sum_{a=1}^{A} \big\|\text{softmax}\big(\mathbf{W}_v\big)^a - \mathbf{Attr}^a\big\|_2^2. \tag{4}$$

To enforce mutual consistency between the two attention flows, we further introduce a symmetric alignment loss:

$$\mathcal{L}_{\text{align}} = \sum_{m=1}^{M} \text{KL}\big(\boldsymbol{\alpha}^m \,\|\, \boldsymbol{\beta}^m\big) + \sum_{m=1}^{M} \text{KL}\big(\boldsymbol{\beta}^m \,\|\, \boldsymbol{\alpha}^m\big), \tag{5}$$

where $\boldsymbol{\alpha}^m$ and $\boldsymbol{\beta}^m$ are the normalized attention distributions from VSA and SVA streams, respectively. By minimizing $\mathcal{L}_{\text{align}}$, we encourage consistent attention patterns across modalities.

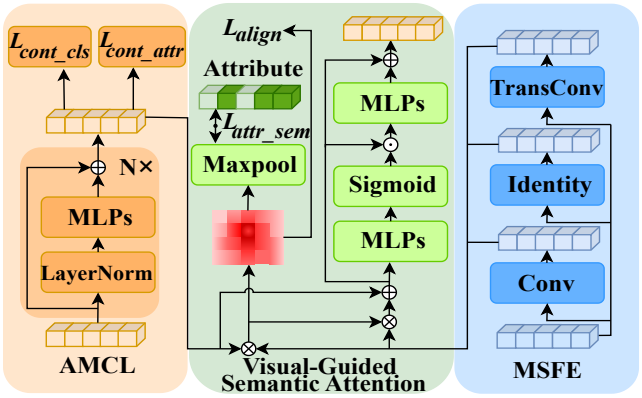


Figure 3: Overview of the visual-guided semantic attention.

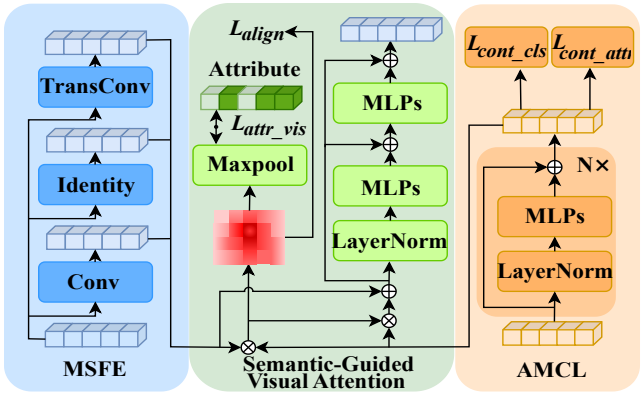


Figure 4: Overview of the semantic-guided visual attention.

Memory Bank-Guided Fusion

To account for the varying relevance and reliability of multi-level representation, we introduce an MBGF module that selectively emphasizes trustworthy branches while correcting unreliable ones through prototype-level guidance.

Let $\{\mathbf{Z}_m\}_{m=1}^{M}$ denote the outputs from M visual-semantic branches, where each $\mathbf{Z}_m \in \mathbb{R}^{1\times d}$ encodes the fused representation at a particular scale. To assess the confidence of each branch with respect to the ground-truth class y, we compute the cosine similarity between $\mathbf{Z}_m$ and the corresponding semantic embedding $\mathbf{a}_y$:

$$s_m = \cos(\mathbf{Z}_m, \mathbf{a}_y) = \frac{\mathbf{Z}_m \cdot \mathbf{a}_y}{\|\mathbf{Z}_m\| \cdot \|\mathbf{a}_y\|}. \tag{6}$$

These scores are normalized to obtain soft fusion weights:

$$w_m = \frac{\exp(\alpha \cdot s_m)}{\sum_{m'=1}^{M} \exp(\alpha \cdot s_{m'})}, \tag{7}$$

where α is a temperature parameter controlling the sharpness of the distribution. The final visual-semantic embedding is given by a weighted sum:

$$\mathbf{Z} = \sum_{m=1}^{M} w_m \cdot \mathbf{Z}_m. \tag{8}$$

If the fused representation $\mathbf{Z}$ leads to a misclassification, we activate a correction mechanism. Specifically, we revisit each branch $\mathbf{Z}_m$ and evaluate whether its similarity to any incorrect class $y' \neq y$ exceeds that of the true class y. For such cases, we retrieve the corresponding *prototype* $\tilde{\mathbf{Z}}_y^m$ from a confidence-aware memory bank $\mathcal{M}_y^m$, which retains only top-k high-confidence representations per class and per scale by dynamically replacing the least reliable ones.

The misaligned branch $\mathbf{Z}_m$ is then refined by minimizing a KL divergence between its predictive distribution and that of the stored prototype:

$$\mathcal{L}_{\text{kl}}^m = \text{KL}\left[\text{softmax}(\mathbf{Z}_m) \parallel \text{softmax}(\tilde{\mathbf{Z}}_y^m)\right]. \tag{9}$$

This mechanism ensures that each branch is not only selectively fused but also retrospectively corrected when semantic confusion occurs. The memory bank is updated dynamically during training: for each class y and scale m, we enqueue new samples only if their prediction confidence exceeds a fixed threshold, and dequeue the oldest entries to maintain a fixed memory size.

Training Objective

Given the fused visual-semantic representation $\mathbf{Z}$ obtained from the MBGF module, we perform classification by computing its similarity with each class-level attribute embedding $\mathbf{a}_y$. This similarity serves as the prediction score for class y, and the model is trained to maximize the score of the correct class.

Formally, the classification loss is defined as:

$$\mathcal{L}_{\text{cls}} = -\log\left(\frac{\exp\big(\text{sim}(\mathbf{Z}, \mathbf{a}_y)\big)}{\sum_{y'\in\mathcal{Y}} \exp\big(\text{sim}(\mathbf{Z}, \mathbf{a}_{y'})\big)}\right), \tag{10}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and $\mathcal{Y}$ is the set of all candidate classes (seen and unseen). This objective encourages the final representation to be most similar to the ground-truth class embedding.

To jointly optimize all components of the framework, we aggregate the classification loss with several auxiliary objectives introduced in previous subsections. These include the attribute-level and class-level contrastive losses ($\mathcal{L}_{\text{cont\_attr}}, \mathcal{L}_{\text{cont\_cls}}$), the cross-attention alignment losses ($\mathcal{L}_{\text{attr\_sem}}, \mathcal{L}_{\text{attr\_vis}}, \mathcal{L}_{\text{align}}$), and the selective memory-bank guided distillation loss ($\mathcal{L}_{\text{kl}}$).

The overall training objective is given by:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{cont}\,\mathcal{L}_{\text{cont}} + \lambda_{attr}\,\mathcal{L}_{\text{attr}} \\ + \lambda_{align}\,\mathcal{L}_{\text{align}} + \lambda_{kl}\,\mathcal{L}_{\text{kl}},\end{aligned} \tag{11}$$

where $\{\lambda_i\}$ are scalar hyperparameters that control the relative importance of each term. This multi-objective optimization ensures that the model not only aligns fused features with the correct class embedding but also maintains semantic consistency across representation levels and improves robustness to hard examples via memory-guided distillation.

Method	AwA2				CUB				SUN			
	A_{zsl}	A_u	A_s	H	A_{zsl}	A_u	A_s	H	A_{zsl}	A_u	A_s	H
Generative-based Method												
CE-GZSL(CVPR'21)	70.4	63.1	78.6	70.0	77.5	63.9	66.8	65.3	63.3	48.8	38.6	43.1
FREE(ICCV'21)	68.9	60.4	75.4	67.1	64.8	55.7	59.9	57.7	65.0	47.4	37.2	41.7
HSVA(NeurIPS'21)	70.6	59.3	76.6	66.8	62.8	52.7	58.3	55.3	63.8	48.6	39.0	43.3
LBP (TPAMI'21)	74.1	–	–	–	61.9	42.7	71.6	53.5	63.2	39.2	36.9	38.1
ICCE(CVPR'22)	72.7	65.3	82.3	72.8	78.4	67.3	65.5	66.4	–	–	–	–
FREE+ESZSL(ICLR'22)	–	51.3	78.0	61.8	–	51.6	60.4	55.7	–	48.2	36.5	41.5
APN+TF-VAEGAN(IJCV'22)	73.5	60.9	79.1	68.8	74.7	65.6	76.3	70.6	66.3	52.6	37.3	43.7
f-VAEGAN+DSP(ICML'23)	71.6	63.7	88.8	74.2	62.8	62.5	73.1	67.4	68.6	57.7	41.3	48.1
VADS (CVPR'24)	82.5	75.4	83.6	79.3	**86.8**	74.1	74.6	74.3	76.3	**64.6**	49.0	55.7
Embedding-based Method												
GNDAN(TNNLS'22)	–	60.2	80.8	69.0	–	69.2	69.6	69.4	–	50.0	34.7	41.0
MSDN(CVPR'22)	70.1	62.0	74.5	67.7	76.1	68.7	67.5	68.1	65.8	52.2	34.2	41.3
TransZero++(TPAMI'22)	72.6	64.6	82.7	72.5	78.3	67.5	73.6	70.4	67.6	48.6	37.8	42.5
ID2Former(NeurIPS'22)	76.4	66.8	76.8	71.5	45.4	35.3	57.6	43.8	–	–	–	–
DUET(AAAI'23)	69.9	63.7	84.7	72.7	72.3	62.9	72.8	67.5	64.4	45.7	45.8	45.8
ZSLViT(CVPR'24)	70.7	66.1	84.6	74.2	78.9	69.4	**78.2**	73.6	68.3	45.9	48.4	47.3
PSVMA+(TPAMI'25)	79.2	74.2	86.4	79.8	78.8	71.8	77.8	74.6	74.5	61.5	49.4	54.8
SAGE	**83.7**	**78.1**	**87.2**	**82.4**	82.1	**75.1**	77.5	**76.3**	**77.2**	63.7	**52.2**	**57.4**

Table 1: Comparison with state-of-the-art methods under the GZSL setting. The best result is indicated in boldface, and the second-best result is indicated with an underline.

Experiments

Datasets and Evaluation Metrics

We conduct experiments on three widely adopted GZSL benchmarks: AwA2, CUB, and SUN. Tab. 2 provides key statistics for these datasets, including the number of classes, the seen/unseen split, the number of attributes, and the total image count. We evaluate our model under the GZSL setting. The ZSL accuracy denoted as A_{zsl} is reported with validation solely on the unseen classes. For GZSL, we measure accuracy on both the seen and unseen classes and the harmonic mean accross all classes, denoted as A_s, A_u and H, respectively.

Dataset	Classes (S|U)	Attributes	Images
AwA2	50 (40|10)	85	37322
CUB	200 (150|50)	312	11788
SUN	717 (645|72)	102	14340

Table 2: Key statistics of the three benchmark datasets employed in our experiments. "S" denotes the number of seen classes, and "U" denotes the number of unseen classes.

Comparison with State-of-the-Art

Tab. 1 summarizes our results against several GZSL frameworks. Compared to generative methods (Han et al. 2021; Chen et al. 2021a,b; Lu et al. 2021; Kong et al. 2022; Cetin, Baran, and Cinbis 2022; Xu et al. 2022; Chen et al. 2023a; Hou et al. 2024), our approach consistently demonstrates stronger performance on Harmonic mean. When evaluated against embedding-based baselines (Chen et al. 2022b,c,a; Naeem et al. 2022; Chen et al. 2023b, 2024; Liu et al. 2025), we achieve more robust and better-harmonized results on both the GZSL and ZSL scenarios. These outcomes highlight SAGE's capacity to reconcile local and global features and to leverage attribute-level cues effectively, leading to reliable recognition even under significant domain generalization.

Ablation Studies

We conduct an ablation study to assess the impact of the key components in the SAGE framework, as shown in Tab. 3. All experiments are conducted under identical conditions, and we report the average results over three runs. The detailed analysis of hyperparameters is shown in the Appendix.

The removal of $\mathcal{L}_{\text{attr}}$ leads to the most significant performance drop, underscoring the necessity of explicit attribute supervision for cross-modal learning. Eliminating $\mathcal{L}_{\text{kl}}$ results in considerable degradation, validating its role in propagating multi-granularity knowledge and hard-sample guidance. While discarding $\mathcal{L}_{\text{cont}}$ causes milder declines, the consistent gains across datasets confirm the benefits of well-clustered semantic prototypes. Notably, omitting $\mathcal{L}_{\text{align}}$ minimally affects overall accuracy but disrupts seen-unseen class balance, revealing its specialized function in distribution alignment rather than pure performance boosting. The full model achieves the best performance on both the ZSL setting and GZSL setting across all benchmarks, demonstrating synergistic effects among components.

Method	AwA2				CUB				SUN			
	A_{zsl}	A_u	A_s	H	A_{zsl}	A_u	A_s	H	A_{zsl}	A_u	A_s	H
baseline	69.7	57.7	81.6	67.6	64.8	58.7	68.4	63.2	50.1	44.1	31.1	36.5
SAGE w/o $\mathcal{L}_{\text{attr}}$	72.3	66.4	72.2	69.2	67.5	61.3	69.4	65.1	51.6	46.2	31.8	37.7
SAGE w/o $\mathcal{L}_{\text{kl}}$	75.8	71.2	77.2	74.1	74.9	68.5	75.4	71.8	65.4	51.7	45.1	48.2
SAGE w/o $\mathcal{L}_{\text{cont}}$	77.6	73.9	78.4	76.1	77.1	70.6	76.6	73.5	72.8	62.1	45.3	52.4
SAGE w/o $\mathcal{L}_{\text{align}}$	82.9	73.2	85.3	78.8	78.2	70.8	**80.0**	75.1	76.6	**66.5**	48.5	56.1
SAGE	**83.7**	**78.1**	**87.2**	**82.4**	**82.1**	**75.1**	77.5	**76.3**	**77.2**	63.7	**52.2**	**57.4**

Table 3: Ablation study of SAGE on significant components. The best result is indicated in boldface, and the second-best result is indicated with an underline.

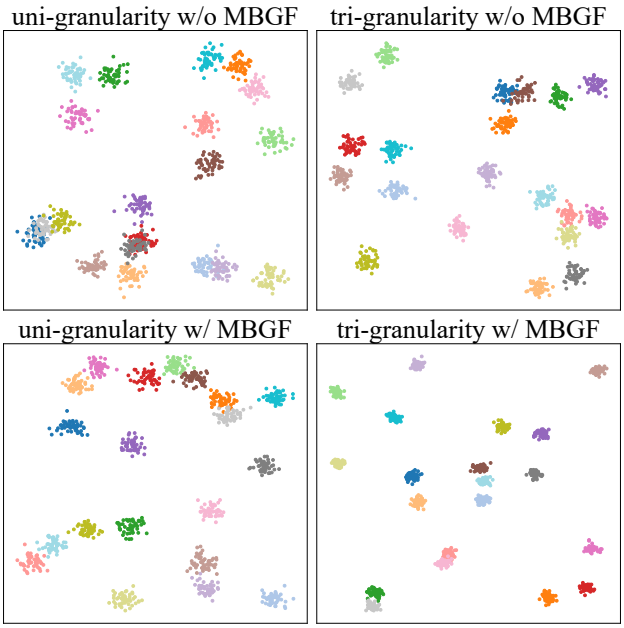
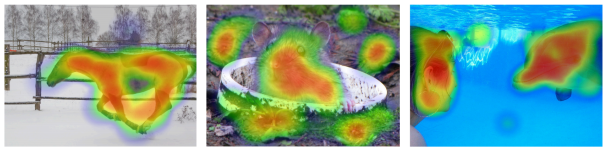


Figure 5: The t-SNE visualization of feature distributions.



(a) (b) (c)

Figure 6: Visualization of attention maps for abstract class-attribute pairs. (a) 'horse' - 'fast', (b) 'rat' - 'smelly', (c) 'dolphin' - 'smart'.

Cluster Behavior Visualization

To better understand how different components of SAGE influence the distribution of learned features, we conduct an analysis using t-SNE visualizations on the AwA2 dataset. We randomly sample 20 classes (10 from the seen and 10 from the unseen) and compare the clustering behavior of features across four configurations. The results are shown in Fig. 5. From the comparison, we observe that tri-granularity leads to better inter-class separability than uni-granularity, while the incorporation of MBGF consistently results in more compact intra-class clusters. These trends confirm that our proposed NAMED framework enhances semantic discrimination across classes and local consistency together, enabling SAGE to learn features exhibiting sharper boundaries" .

Attribute Attention Visualization

To further examine SAGE's ability to model abstract attributes, we visualize class-attribute attention maps for three representative pairs: *horse–fast*, *rat–smelly*, and *dolphin–smart*, as illustrated in Fig. 6. In each case, the attention is not limited to superficial features but extends to semantically relevant regions. For instance, *fast* activates the horse's body axis, *smelly* attends to both the rat and its environment, while *smart* highlights interactions between the dolphin and the human. These results demonstrate the model's ability to capture high-level semantic cues beyond visual appearance, enabling more nuanced alignment between semantic attributes and corresponding visual characteristics.

Conclusion & Future Work

We propose SAGE, a structured attribute-guided enhancement framework for GZSL. By integrating multi-granularity visual encoding, attribute-level semantic refinement, cross-modal attention alignment, and memory bank-guided fusion, SAGE effectively bridges the visual-semantic gap and improves recognition of both the seen and the unseen classes. Extensive experiments on three benchmarks demonstrate consistent improvements over state-of-the-art methods. Visualization analyses further confirm SAGE's capacity to model multi-grained attributes.

While SAGE demonstrates strong performance in GZSL, a key direction warrants further investigation. The current attribute grouping mechanism relies on human-defined configurations, exploring adaptive methods to automatically determine the optimal number of groups and their composition could further enhance the framework's generalization capability.

Acknowledgements

This work is supported by the National Science and Technology Major Project(2022ZD0116406).

References

Cetin, S.; Baran, O. B.; and Cinbis, R. G. 2022. Closed-Form Sample Probing for Learning Generative Models in Zero-Shot Learning. *International Conference on Learning Representations*.

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*.

Chen, S.; Hong, Z.; Xie, G.; Peng, Q.; You, X.; Ding, W.; and Shao, L. 2022b. 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. 2022c. MSDN: Mutually Semantic Distillation Network for Zero-Shot Learning. *IEEE Conference on Computer Vision and Pattern Recognition*.

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. *International Conference on Machine Learning*.

Chen, S.; Hou, W.; Khan, S.; and Khan, F. S. 2024. Progressive Semantic-Guided Vision Transformer for Zero-Shot Learning. *IEEE Conference on Computer Vision and Pattern Recognition*.

Chen, S.; Wang, W.; Xia, B.; Peng, Q.; You, X.; Zheng, F.; and Shao, L. 2021a. Free: Feature Refinement for Generalized Zero-Shot Learning. *International Conference on Computer Vision*.

Chen, S.; Xie, G.-S.; Liu, Y. Y.; Peng, Q.; Sun, B.; Li, H.; You, X.; and Shao, L. 2021b. HSVA: Hierarchical Semantic-Visual Adaptation for Zero-Shot Learning. *Advances in Neural Information Processing Systems*.

Chen, Z.; Huang, Y.; Chen, J.; Geng, Y.; Zhang, W.; Fang, Y.; Pan, J. Z.; Song, W.; and Chen, H. 2023b. DUET: Cross-Modal Semantic Grounding for Contrastive Zero-Shot Learning. *AAAI Conference on Artificial Intelligence*.

Chun, S.; Oh, S. J.; de Rezende, R. S.; Kalantidis, Y.; and Larlus, D. 2021. Probabilistic Embeddings for Cross-Modal Retrieval. *IEEE Conference on Computer Vision and Pattern Recognition*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. *Advances in Neural Information Processing Systems*.

Han, Z.; Fu, Z.; Chen, S.; and Yang, J. 2021. Contrastive Embedding for Generalized Zero-Shot Learning. *IEEE Conference on Computer Vision and Pattern Recognition*.

Hou, W.; Chen, S.; Chen, S.; Hong, Z.; Wang, Y.; Feng, X.; Khan, S.; Khan, F. S.; and You, X. 2024. Visual-Augmented Dynamic Semantic Prototype for Generative Zero-Shot Learning. *IEEE Conference on Computer Vision and Pattern Recognition*.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. *International Conference on Learning Representations*.

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. *IEEE Conference on Computer Vision and Pattern Recognition*.

Li, H.; Chen, Y.; Chen, Y.; Yu, R.; Yang, W.; Wang, L.; Ding, B.; and Han, Y. 2024a. Generalizable Whole Slide Image Classification with Fine-Grained Visual-Semantic Interaction. *IEEE Conference on Computer Vision and Pattern Recognition*.

Li, Y.; Luo, Y.; Wang, Z.; and Du, B. 2024b. Improving Generalized Zero-Shot Learning by Exploring the Diverse Semantics from External Class Names. *IEEE Conference on Computer Vision and Pattern Recognition*.

Lin, Z.; Yu, S.; Kuang, Z.; Pathak, D.; and Ramanan, D. 2023. Multimodality Helps Unimodality: Cross-Modal Few-Shot Learning with Multimodal Models. *IEEE Conference on Computer Vision and Pattern Recognition*.

Liu, M.; Bai, H.; Li, F.; Zhang, C.; Wei, Y.; Wang, M.; Chua, T.-S.; and Zhao, Y. 2025. PSVMA+: Exploring Multi-Granularity Semantic-Visual Adaption for Generalized Zero-Shot Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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*.

Mancini, M.; Naeem, M. F.; Xian, Y.; and Akata, Z. 2021. Open World Compositional Zero-Shot Learning. *IEEE Conference on Computer Vision and Pattern Recognition*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient Estimation of Word Representations in Vector Space. *International Conference on Learning Representations*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013b. Distributed Representations of Words and Phrases and Their Compositionality. *Advances in Neural Information Processing Systems*.

Naeem, M. F.; Xian, Y.; Gool, L. V.; and Tombari, F. 2022. I2DFormer: Learning Image to Document Attention for Zero-Shot Image Classification. *Advances in Neural Information Processing Systems*.

Pan, Z.; Wu, F.; and Zhang, B. 2023. Fine-Grained Image-Text Matching by Cross-Modal Hard Aligning Network. *IEEE Conference on Computer Vision and Pattern Recognition*.

Shi, B.; Ji, L.; Lu, P.; Niu, Z.; and Duan, N. 2019. Knowledge Aware Semantic Concept Expansion for Image-Text Matching. *International Joint Conference on Artificial Intelligence*.

Vyas, M. R.; Venkateswara, H.; and Panchanathan, S. 2020. Leveraging Seen and Unseen Semantic Relationships for Generative Zero-Shot Learning. *European Conference on Computer Vision*.

Wang, Q.; Liu, L.; Jing, C.; Chen, H.; Liang, G.; Wang, P.; and Shen, C. 2023. Learning Conditional Attributes for Compositional Zero-Shot Learning. *Computing Research Repository*.

Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2018. Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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. *IEEE Conference on Computer Vision and Pattern Recognition*.

Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; and Akata, Z. 2020. Attribute Prototype Network for Zero-Shot Learning. *Advances in Neural Information Processing Systems*.

Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; and Akata, Z. 2022. Attribute Prototype Network for Any-Shot Learning. *International Journal of Computer Vision*.