# ZeroMamba: Exploring Visual State Space Model for Zero-Shot Learning

**Wenjin Hou[1], Dingjie Fu[2], Kun Li[1], Shiming Chen[3], Hehe Fan[1]*, Yi Yang[1]**

[1]ReLER Lab, Zhejiang University, China
[2]Huazhong University of Science and Technology (HUST), China
[3]Mohamed bin Zayed University of AI
{wenjinhou, hehefan, yangyics}@zju.edu.cn

## Abstract

Zero-shot learning (ZSL) aims to recognize unseen classes by transferring semantic knowledge from seen classes to unseen ones, guided by semantic information. To this end, existing works have demonstrated remarkable performance by utilizing global visual features from Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs) for visual-semantic interactions. Due to the limited receptive fields of CNNs and the quadratic complexity of ViTs, however, these visual backbones achieve suboptimal visual-semantic interactions. In this paper, motivated by the visual state space model (*i.e.*, Vision Mamba), which is capable of capturing long-range dependencies and modeling complex visual dynamics, we propose a parameter-efficient ZSL framework called **ZeroMamba** to advance ZSL. Our ZeroMamba comprises three key components: Semantic-aware Local Projection (SLP), Global Representation Learning (GRL), and Semantic Fusion (SeF). Specifically, SLP integrates semantic embeddings to map visual features to local semantic-related representations, while GRL encourages the model to learn global semantic representations. SeF combines these two semantic representations to enhance the discriminability of semantic features. We incorporate these designs into Vision Mamba, forming an end-to-end ZSL framework. As a result, the learned semantic representations are better suited for classification. Through extensive experiments on four prominent ZSL benchmarks, ZeroMamba demonstrates superior performance, significantly outperforming the state-of-the-art (*i.e.*, CNN-based and ViT-based) methods under both conventional ZSL (CZSL) and generalized ZSL (GZSL) settings.

**Code** — https://github.com/Houwenjin/ZeroMamba

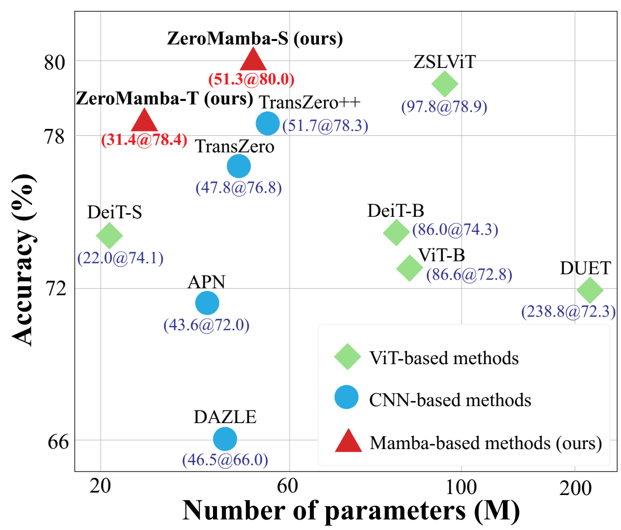


Figure 1: Compared with state-of-the-art CNN-based and ViT-based methods on CUB, our proposed ZeroMamba achieves the best trade-off between **Accuracy** and **Number of parameters**.

## Introduction

Zero-shot learning (ZSL) recognizes unseen classes with the support of shared semantic information (*e.g.*, category attributes (Lampert, Nickisch, and Harmeling 2009, 2013), semantic embeddings (Huynh and Elhamifar 2020; Radford et al. 2021)). Since images of unseen classes are unavailable during training, mainstream methods (Huynh and Elhamifar 2020; Wang et al. 2021; Xu et al. 2022; Chen et al. 2022c, 2023a, 2024c,a; Hou et al. 2024; Naeem et al. 2024) focus on establish visual-semantic interactions to align visual and semantic features, transferring semantic knowledge from seen classes to unseen ones. As a result, learning discriminative visual and semantic representations is crucial for ZSL.

Current methods typically leverage visual backbones with the powerful learning capabilities of CNNs (*e.g.*, ResNet-101 (He et al. 2016)) or ViTs (*e.g.*, ViT (Dosovitskiy et al. 2020)) to extract visual features. However, CNNs suffer from limited receptive fields in convolution operations (Luo et al. 2016), and ViTs face quadratic complexity issues in self-attention calculations (Habib, Saleem, and Lall 2023). Additionally, these visual backbones fail to learn desirable visual-semantic correspondences due to the lack of semantic guidance during pre-training on ImageNet (Russakovsky et al. 2015). Given these limitations, the performance and computational efficiency of CNN-based and ViT-based ZSL methods remain suboptimal, as illustrated in Fig. 1.

Recently, the visual state space model (*i.e.*, Vision Mamba) (Zhu et al. 2024; Huang et al. 2024; Li, Singh, and

*Corresponding author

Grover 2025), a novel and promising network architecture with advantages in linear complexity and global receptive fields, has demonstrated remarkable performance across various visual perception tasks, such as image (Liu et al. 2024; Weng et al. 2024), video (Li et al. 2024a; Park et al. 2024) and multi-modal (Fu et al. 2024) scenarios. Drawing inspiration from its ability to capture long-range dependencies and model complex visual dynamics, we *explore the visual state space model* to advance ZSL. To this end, the basic idea is straightforward: utilize Vision Mamba as the backbone to extract visual features and then directly map these features to the semantic space via a multi-layer perceptron (MLP) network for nearest-neighbor matching. However, significant challenges for ZSL lie in effectively bridging visual-semantic spaces and inserting discriminative semantic information. The basic solution tends to overfit to seen classes, severely limiting the transfer of intrinsic semantic knowledge. Thus, improving Visual Mamba to better suit ZSL and provide valuable insights are crucial issues.

In light of the above observations, our major innovations in this paper are **three** effective designs: *Firstly*, we introduce a Semantic-aware Local Projection (SLP) module conditioned on semantic embeddings to learn semantic-related local representations; *Secondly*, we develop a Global Representation Learning (GRL) module to bridge the visual and semantic spaces; *Thirdly*, we devise a Semantic Fusion (SeF) strategy to combine the two semantic representations, enhancing the discriminability of semantic features. By incorporating these three simple yet effective designs into Vision Mamba, we form an end-to-end ZSL framework called **ZeroMamba**. This unified framework allows us to simultaneously optimize visual and semantic representations during training, significantly boosting effective visual-semantic interactions. As shown in Fig. 1, our ZeroMamba is a parameter-efficient method that achieves the best trade-off between accuracy and the number of parameters. Overall, our contribution can be summarized as follows:

- We propose ZeroMamba, a parameter-efficient Mamba-based framework for ZSL that exhibits exceptional generalization of unseen classes and meanwhile has the advantages of global receptive fields. To the best of our knowledge, ZeroMamba is the first work attempting to explore Vision Mamba for the ZSL task.
- To adapt to ZSL, we design the Semantic-aware Local Projection (SLP) module, the Global Representation Learning (GRL) module, and the Semantic Fusion (SeF) strategy, integrating them into Vision Mamba to form an end-to-end ZSL framework.
- Extensive experiments with analyses on three prominent ZSL benchmarks and one large-scale dataset show that ZeroMamba sets a new state-of-the-art (SOTA) in both CZSL and GZSL settings, illustrating the effectiveness and superiority of our approach.

## Related Work

**Zero-Shot Learning.** Zero-shot learning (ZSL) transfers semantic knowledge from seen classes to unseen ones through visual-semantic interactions. According to the direction of visual and semantic mappings, mainstream ZSL methods can be categorized into generative and embedding-based approaches (Xian et al. 2019), implemented via semantic→visual and visual→semantic mappings, respectively. Typically, visual features are extracted using visual backbones like CNNs or ViTs. Semantic information is derived from category attributes. In this regard, enhancing the representation capability of these features has garnered widespread attention. In CNN feature-based methods, early works (Huynh and Elhamifar 2020; Chen et al. 2021b,a; Han et al. 2021; Cetin 2022; Feng et al. 2022) directly utilized global visual features, while recent studies (Wang et al. 2021; Xu et al. 2022; Chen et al. 2022b, 2024a; Li et al. 2024b) focus on local discriminative features. In these works, APN (Xu et al. 2020) and TransZero (Chen et al. 2022b) obtain semantic-related representations through local attention mechanisms; DPPN (Wang et al. 2021) simultaneously updates attributes and class prototypes; CE-GZSL (Han et al. 2021) and ICCE (Kong et al. 2022) map visual features to latent spaces; DSECN (Li et al. 2024b) explores diverse semantics from external class names.

Recently, ViTs with excellent self-attention mechanisms have also been applied to ZSL. For example, DUET (Chen et al. 2023b) introduces cross-modal masks; PSVMA (Liu et al. 2023) proposes a semantic-visual mutual adaption network; ZSLViT (Chen et al. 2024c) leverages semantic information to guide network learning; CVsC (Chen et al. 2024b) learns substantive visual-semantic correlations. In addition, large-scale vision-language models (*e.g.*, CLIP (Radford et al. 2021)) emerge with impressive ZSL transfer capabilities by aligning visual-semantic features in a common space. Analyzing these methods reveals that discriminative visual features and semantic representations are crucial for effective visual-semantic interactions.

**Vision Mamba.** State space models (SSMs), particularly Mamba (Gu and Dao 2023; Dao and Gu 2024), have recently demonstrated strong long-range modeling capabilities while maintaining linear complexity. In the field of Computer Vision, Vim (Zhu et al. 2024) is the first Mamba-based backbone network, showcasing superior performance and the ability to capture complex visual dynamics across a variety of vision tasks (*e.g.*, image classification, semantic segmentation and object detection). Subsequent Mamba-based models, such as VMamba (Liu et al. 2024) and Local-Mamba (Huang et al. 2024), primarily focus on image scanning strategies and Mamba blocks. VideoMamba (Li et al. 2024a; Park et al. 2024) extends Mamba to the video domain. HMNet (Xu et al. 2024) introduces Mamba to few-shot segmentation. However, Mamba-based ZSL methods have yet to be explored. Therefore, this work presents a simple Mamba-based framework specifically for ZSL and provides a comprehensive analysis of its performance.

## Proposed Method

ZeroMamba aims to incorporate the visual state space model (*i.e.*, Vision Mamba) into ZSL. Fig. 2 illustrates the framework of our ZeroMamba. In this section, we first present the

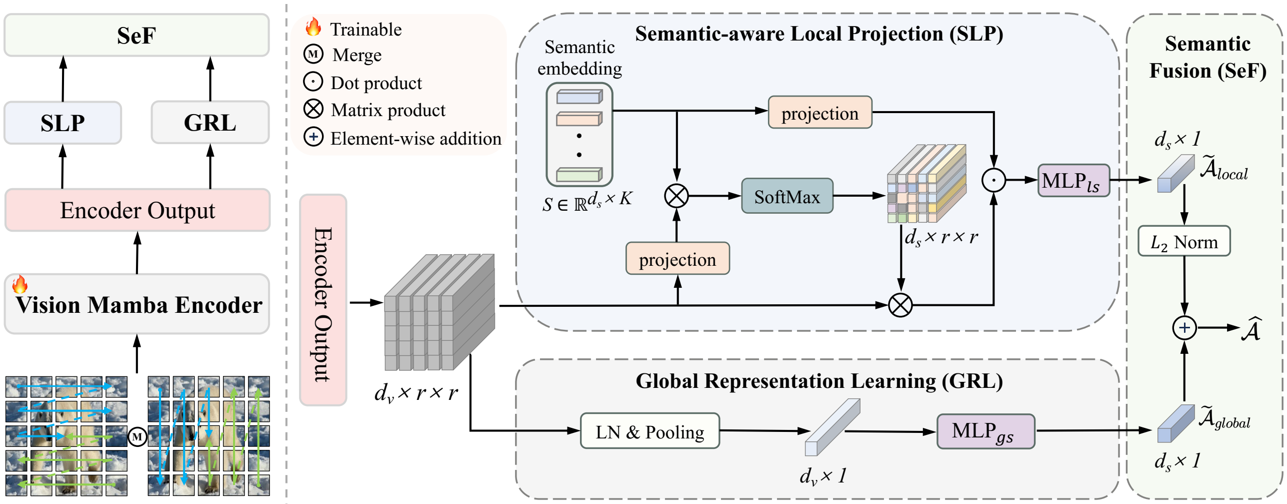


Figure 2: **Left:** The framework of **ZeroMamba**. Image patches are traversed along four-way scanning and fed into the Vision Mamba Encoder. The major innovation of our work lies in three simple yet effective designs: the SLP module, the GRL module, and the SeF strategy. By incorporating these designs into vanilla VMamba (Liu et al. 2024), we form an end-to-end framework capable of robust ZSL. **Right:** The structural details of the SLP module, the GRL module, and the SeF strategy.

definition of the ZSL task. Next, we introduce the preliminaries of the state space models and the selection mechanism. Then, we describe the details of the proposed method. Finally, we demonstrate the model optimization and zero-shot prediction process.

**Problem Definition.** Assume $\mathcal{D}^s = \{(x_i, y_i)|x_i \in \mathcal{X}^s, y_i \in \mathcal{Y}^s\}$ with $\mathcal{C}^s$ classes, where $x_i$ and $y_i$ represent the image of the *i*-th sample and its label, respectively. A disjoint set $\mathcal{D}^u = \{(x_i, y_i)|x_i \in \mathcal{X}^u,\ y_i \in \mathcal{Y}^u\}$ consists of $\mathcal{C}^u$ classes. Notably, $\mathcal{C}^s \cap \mathcal{C}^u = \emptyset$. In the conventional zero-shot learning (CZSL) setting, the training set $\mathcal{D}^{tr}$ is a subset of $\mathcal{D}^s$. The main goal is to classify a test image $x \in \mathcal{D}^u$ that belongs to an unseen class, given that there are no training images available for unseen classes. In the generalized zero-shot learning (GZSL) setting, the testing samples include both seen and unseen classes, *i.e.*, $c \in \mathcal{C}^s \cup \mathcal{C}^u = \mathcal{C}$. In this paper, we utilize class-level attribute vectors (*i.e.*, semantic prototypes) $\mathcal{A} \in \mathbb{R}^{K \times C}$ to bridge the gap between seen and unseen classes, where $K$ refers to the number of attributes and $C$ denotes the total number of categories. We also use the shared semantic embeddings $\mathcal{S} \in \mathbb{R}^{d_s \times K}$, extracted using GloVe, to boost visual-semantic interactions.

## Preliminaries

**State Space Models**. State space models (SSMs) have emerged as a novel alternative to CNNs and ViTs for modeling long-range dependency. Continuous-time SSMs map an input stimulation $x(t) \in \mathbb{R}$ to a response $y(t) \in \mathbb{R}$ via a hidden state $h(t) \in \mathbb{R}^N$. This progress can be mathematically formulated as follows:

$$\begin{aligned} h'(t) &= \boldsymbol{A}h(t) + \boldsymbol{B}x(t), \\ y(t) &= \boldsymbol{C}h(t), \end{aligned} \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{N \times N}$, $\boldsymbol{B} \in \mathbb{R}^{N \times 1}$ and $\boldsymbol{C} \in \mathbb{R}^{1 \times N}$ are the projection parameters. The term $h'(t)$ denotes the derivative of $h(t)$ with respect to time $t$.

To handle discrete input sequences, the commonly used zero-order hold assumption leverages a time step parameter $\Delta$ to convert the continuous parameters $\boldsymbol{A}$ and $\boldsymbol{B}$ into their discrete counterparts $\overline{\boldsymbol{A}}$ and $\overline{\boldsymbol{B}}$ as follows:

$$\begin{aligned} \overline{\boldsymbol{A}} &= \exp(\Delta \boldsymbol{A}), \\ \overline{\boldsymbol{B}} &= (\Delta \boldsymbol{A})^{-1}(\exp(\Delta \boldsymbol{A}) - \boldsymbol{I}) \cdot \Delta \mathbf{B}. \end{aligned} \tag{2}$$

After discretizing, the formulation of Eq. (1) is altered to:

$$\begin{aligned} h_t &= \overline{\boldsymbol{A}}h_{t-1} + \overline{\boldsymbol{B}}x_t, \\ y_t &= \boldsymbol{C}h_t. \end{aligned} \tag{3}$$

The model utilizes Eq. (3) to facilitate efficient inference, where the inputs are processed sequentially. For efficient and parallelizable training, Eq. (3) can be reformulated and computed as a global convolution:

$$\begin{aligned} \overline{\boldsymbol{K}} &= (\boldsymbol{C}\overline{\boldsymbol{B}}, \boldsymbol{C}\overline{\boldsymbol{A}}\overline{\boldsymbol{B}}, ..., \boldsymbol{C}\overline{\boldsymbol{A}}^{L-1}\overline{\boldsymbol{B}}), \\ y &= x * \overline{\boldsymbol{K}}, \end{aligned} \tag{4}$$

where $L$ represents the length of the input sequence, and $\overline{\boldsymbol{K}} \in \mathbb{R}^L$ is the convolutional kernel.

**Selection Mechanism**. Recent SSMs, such as Mamba models (Gu and Dao 2023; Dao and Gu 2024), leverage an input-dependent selection mechanism to address the limitations of fixed parameterization. By adopting such a mechanism, Mamba models exhibit linear scalability and demonstrate strong capabilities in long-range modeling.

In vision tasks, the Mamba model designed for 1-D sequence may not be optimal for handling 2-D image inputs. Therefore, Vision Mamba models have introduced a variety of selective scan mechanisms. Concretely, Vim (Zhu et al. 2024) combines Mamba with bidirectional SSM paths, VMamba (Liu et al. 2024) employs the 2D-Selective-Scan method, LocalMamba (Huang et al. 2024) segments the input image into local windows to perform SSM in different directions while preserving global SSM operations. In our work, we use SS2D to merge contextual information from four directions, following the design of VMamba.

**Overview.** Following the embedding-based ZSL methods, we can achieve ZSL through the following steps: First, use a pre-trained Vision Mamba model as the visual backbone to extract visual feature $\boldsymbol{v}_i$ of the image $x_i$. Then, apply a multi-layer perceptron (MLP) with a parameter matrix $\boldsymbol{W} \in \mathbb{R}^{d_s \times d_v}$ as a semantic predictor to obtain a global semantic feature $\boldsymbol{a}_i$ for classification:

$$\boldsymbol{a}_i = f(\boldsymbol{v}_i) = \boldsymbol{W}\boldsymbol{v}_i. \tag{5}$$

However, such a naive method inevitably leads to overfitting to seen classes, severely limiting the transfer of intrinsic semantic knowledge.

To tackle the above challenges, we proposed a Mamba-based ZSL framework, named ZeroMamba, which is shown in Fig. 2. ZeroMamba first acquires image patches through a four-way selective scan mechanism to aggregate contextual information. These patches are then fed into the Vision Mamba Encoder to extract visual features. More details can also be found in VMamba (Liu et al. 2024).). Next, three components – SLP module, GRL module, and Semantic Fusion strategy – are combined with the Vision Mamba Encoder. As a result, ZeroMamba forms an end-to-end framework that integrates semantic information into the network and simultaneously optimizes visual and semantic representations. The following sections detail our proposed designs.

## Semantic-aware Local Projection (SLP)

To enhance ZeroMamba with rich semantic information, we propose a Semantic-aware Local Projection (SLP) module. As shown in Fig. 2, the SLP first extracts semantic embeddings: $\mathcal{S} = \{\boldsymbol{s}_i\}_{i=1}^{K}$, where $\boldsymbol{s}_i \in \mathbb{R}^{d_s}$ denotes the semantic embedding of the $i$-th attribute. Subsequently, we flatten the encoder's output feature map (*i.e.*, $\mathcal{F} \in \mathbb{R}^{d_v \times r \times r} \rightarrow \boldsymbol{F}_{r'} \in \mathbb{R}^{d_v \times r'}$), where $r'$ corresponds to $r \times r$ image regions. For these features, we apply two projection parameters $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ to calculate the attention map $\mathcal{M}$ and obtain the local semantic representation $f_s \in \mathbb{R}^{d_s \times 1}$. This process is formalized as follows:

$$\mathcal{M} = \text{SoftMax}(\mathcal{S}\boldsymbol{W}_1\boldsymbol{F}_{r'}),\ f_s = \mathcal{S}\boldsymbol{W}_2 \odot (\mathcal{M}\boldsymbol{F}_{r'})^T, \tag{6}$$

where $\odot$ denotes dot product operation, and $\mathcal{M}\boldsymbol{F}_{r'}$ represents the semantic-related attention feature.

Upon obtaining $f_s$, we employ an MLP (*i.e.*, $\text{MLP}_{ls}$) that maps local semantic representations to semantic space, denoted as $\widetilde{\mathcal{A}}_{local}$:

$$\widetilde{\mathcal{A}}_{local} = \text{MLP}_{ls}(f_s). \tag{7}$$

Through this process, the SLP effectively captures local semantic-related representations, facilitating visual-semantic interactions.

## Global Representation Learning (GRL)

We devise a Global Representation Learning (GRL) module to discover global semantic information and bridge visual and semantic spaces. As shown in Fig. 2, the GRL employs LayerNorm to stabilize feature activations and subsequently uses average pooling to obtain an informative global visual feature $f_v \in \mathbb{R}^{d_v \times 1}$. At last, GRL leverages a MLP (*i.e.*, $\text{MLP}_{gs}$) to project the visual embedding to global semantic representation, denoted as $\widetilde{\mathcal{A}}_{global}$:

$$\widetilde{\mathcal{A}}_{global} = \text{MLP}_{gs}(f_v). \tag{8}$$

## Semantic Fusion Strategy (SeF)

By applying SLP and GRL, we obtain global and local semantic representations. We argue that fusing these representations enhances the exploration of complementary semantic information for classification. To this end, we propose a simple Semantic Fusion (SeF) strategy, formulated as follows:

$$\widehat{\mathcal{A}} = \widetilde{\mathcal{A}}_{global} + \text{Norm}(\widetilde{\mathcal{A}}_{local}), \tag{9}$$

where Norm denotes the $L_2$ normalization, which suppresses the significance of $\widetilde{\mathcal{A}}_{local}$. To ensure alignment with the semantic prototypes $\mathcal{A}$, we introduce a semantic constraint loss $\mathcal{L}_{sc}$ to guide the optimization of the semantic representation $\widehat{\mathcal{A}}$, defined as:

$$\mathcal{L}_{sc} = \mathbb{E}[\|\widehat{\mathcal{A}} - \mathcal{A}\|_1]. \tag{10}$$

Based on the above fusion operations, we optimize both global and local semantic representations, boosting effective visual-semantic interactions.

## Model Optimization and Zero-Shot Prediction

**Optimization**. In addition to the semantic constraint loss mentioned above, we optimize the cross-entropy loss between the prediction and ground-truth label:

$$\mathcal{L}_{ce} = -\log \frac{\exp(p_i^c)}{\sum_{c' \in \mathcal{C}^s} \exp(p_i^{c'})}, \tag{11}$$

where $p_i^c$ is the probability that image $i$ belongs to class $c$, formulated as:

$$p_i^c = \cos(\widehat{\mathcal{A}_i^c},\ \mathcal{A}^c). \tag{12}$$

Here, $\cos(\cdot)$ denotes the cosine function, which measures the similarity between predicted semantic feature $\widehat{\mathcal{A}}$ and class semantic prototypes $\mathcal{A}$.

Finally, the overall loss function for our model is:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_{sc}\mathcal{L}_{sc}, \tag{13}$$

where $\lambda_{sc}$ is a hyper-parameter to control the magnitude of $\mathcal{L}_{sc}$, with its value varying across different datasets.

**Prediction**. After training, we input test sample $x_i$ into ZeroMamba, obtain its semantic representation $\widehat{\mathcal{A}_i}$, and calculate the probabilistic vector $p_i$. To this end, we predict the class label $y^*$ using the following formulation:

$$y^* = \underset{y \in \mathcal{Y}^u / \mathcal{Y}^s \cup \mathcal{Y}^u}{\arg\max} (p_i + \lambda_{col}\mathbb{I}_{[y \in \mathcal{Y}^u]}). \tag{14}$$

$\mathbb{I}_{[y \in \mathcal{Y}^u]}$ is an indicator function, which is 1 when $y \in \mathcal{Y}^u$ and 0 otherwise. $\lambda_{col}$ is a calibration coefficient that explicitly calibrates the sensitivity of the predictor to unseen classes, and $\mathcal{Y}^u / \mathcal{Y}^s \cup \mathcal{Y}^u$ corresponds to the CZSL/GZSL setting.

| Method | Backbone | Venue | CUB | | | | SUN | | | | AWA2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | CZSL | GZSL | | | CZSL | GZSL | | | CZSL | GZSL | | |
| | | | Acc | U | S | H | Acc | U | S | H | Acc | U | S | H |
| **CNN-based methods** | | | | | | | | | | | | | | |
| DAZLE (Huynh and Elhamifar 2020) | ResNet-101 | 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 |
| APN (Xu et al. 2020) | ResNet-101 | 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 |
| HSVA (Chen et al. 2021b) | ResNet-101 | NeurIPS'21 | 62.8 | 52.7 | 58.3 | 55.3 | 63.8 | 48.6 | 39.0 | 43.3 | – | 59.3 | 76.6 | 66.8 |
| SE-GZSL (Kim, Shim, and Shim 2022) | ResNet-101 | AAAI'22 | – | 53.1 | 60.3 | 56.4 | – | 45.8 | 40.7 | 43.1 | – | 59.9 | 80.7 | 68.8 |
| TransZero++ (Chen et al. 2022a) | ResNet-101 | 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 |
| FREE + ESZSL (Cetin 2022) | ResNet-101 | ICLR'22 | – | 51.6 | 60.4 | 55.7 | – | 48.2 | 36.5 | 41.5 | – | 51.3 | 78.0 | 61.8 |
| f-CLSWGAN + DSP (Chen et al. 2023a) | ResNet-101 | ICML'23 | – | 51.4 | 63.8 | 56.9 | – | 48.3 | 43.0 | 45.5 | – | 60.0 | 86.0 | 70.7 |
| CDL + OSCO (2023) | ResNet-101 | TPAMI'23 | – | 29.0 | 69.0 | 40.6 | – | 32.0 | **65.0** | 42.9 | – | 48.0 | 71.0 | 57.1 |
| ICIS (Christensen et al. 2023) | ResNet-101 | ICCV'23 | 60.0 | 45.8 | 73.7 | 56.5 | 51.8 | 45.2 | 25.6 | 32.7 | 64.6 | 35.6 | **93.3** | 51.6 |
| HAS (Chen et al. 2023c) | ResNet-101 | ACM MM'23 | 76.5 | 69.6 | 74.1 | 71.8 | 63.2 | 42.8 | 38.9 | 40.8 | 71.4 | 63.1 | 87.3 | 73.3 |
| TransZero + ALR (Chen et al. 2024a) | ResNet-101 | SCIS'24 | 78.8 | **70.4** | 69.0 | 69.7 | 66.2 | **52.7** | 34.0 | 41.3 | 71.2 | 61.6 | 82.3 | 70.5 |
| DSECN (Li et al. 2024b) | ResNet-101 | CVPR'24 | 40.9 | – | – | 45.3 | 49.1 | – | – | 38.5 | 40.0 | – | – | 53.7 |
| **ViT-based methods** | | | | | | | | | | | | | | |
| ViT-ZSL (Alamri and Dutta 2021) | ViT-Large | IMVIP'21 | – | 67.3 | 75.2 | 71.0 | – | 44.5 | **55.3** | **49.3** | – | 51.9 | 90.0 | 68.5 |
| CLIP (Radford et al. 2021) | ViT-Base | ICML'21 | – | 55.2 | 54.8 | 55.0 | – | – | – | – | – | – | – | – |
| DUET (Chen et al. 2023b) | ViT-Base | AAAI'23 | 72.3 | 62.9 | 72.8 | 67.5 | 64.4 | 45.7 | 45.8 | 45.8 | 69.9 | 63.7 | 84.7 | 72.7 |
| TFVAEGAN + SHIP (Wang et al. 2023) | ViT-Base | ICCV'23 | – | 21.1 | **84.4** | 34.0 | – | – | – | – | – | 43.7 | **96.3** | 60.1 |
| I2MVFormer-Wiki (Naeem et al. 2023) | ViT-Base | CVPR'23 | 42.1 | 32.4 | 63.1 | 42.8 | – | – | – | – | **73.6** | 66.6 | 82.9 | 73.8 |
| I2DFormer+ (Naeem et al. 2024) | ViT-Base | IJCV'24 | 45.9 | 38.3 | 55.2 | 45.3 | – | – | – | – | **77.3** | **69.8** | 83.2 | **75.9** |
| ZSLViT (Chen et al. 2024c) | ViT-Base | CVPR'24 | **78.9** | 69.4 | **78.2** | **73.6** | **68.3** | 45.9 | 48.4 | 47.3 | 70.7 | 66.1 | 84.6 | 74.2 |
| **Mamba-based method** | | | | | | | | | | | | | | |
| **ZeroMamba (Ours)** | VMamba-Small | – | **80.0** | **72.1** | 76.4 | **74.2** | **72.4** | **56.5** | 41.4 | **47.7** | 71.9 | **67.9** | 87.6 | **76.5** |

Table 1: Comparing our ZeroMamba with CNN-based and ViT-based methods on CUB, SUN, and AWA2 benchmark datasets in the CZSL and GZSL settings. Our ZeroMamba significantly sets a new state-of-the-art (SOTA) for zero-shot learning. The best and second-best results are highlighted in **Red** and **Blue**, respectively. Symbol "–" denotes no results are reported.

| Method | SEKG (2018) | CADA-VAE (2019) | Transzero++ (2022a) | I2DFormer (2022) | I2DFormer+ (2024) | **ZeroMamba (Ours)** |
|---|---|---|---|---|---|---|
| *Acc* (%) | 10.8 | 9.8 | **23.9** | 15.5 | 17.6 | **24.5** |

Table 2: Comparison results on ImageNet. The best and second-best results are highlighted in **Red** and **Blue**, respectively.

## Experiments

**Benchmark Datasets.** To evaluate the effectiveness of our proposed framework, we conduct extensive experiments on three prominent ZSL datasets: Caltech-USCD Birds-200-2011 (**CUB**) (Welinder et al. 2010), SUN Attribute (**SUN**) (Patterson and Hays 2012) and Animals with Attributes 2 (**AWA2**) (Xian et al. 2019). We use the Proposed Split (PS) (Xian et al. 2019) division, as detailed in Tab. 3. Additionally, we verify the generalization of ZeroMamba on large-scale ImageNet (Russakovsky et al. 2015) benchmark.

| Datasets | Type | Images | $N_S$ | Classes ($s$ \| $u$) |
|---|---|---|---|---|
| **CUB** | bird | 11788 | 312 | 200 (150 \| 50) |
| **SUN** | scene | 14340 | 102 | 717 (645 \| 72) |
| **AWA2** | animal | 37322 | 85 | 50 (40 \| 10) |

Table 3: Details of the ZSL datasets. $N_S$ denotes the dimensions of the semantic vector. $s$ and $u$ are the number of seen and unseen classes.

**Evaluation Metrics.** Following previous work (Xian et al. 2019; Naeem et al. 2024; Chen et al. 2024c), we measure the average Top-1 accuracy per unseen class for the CZSL setting, denoted as $Acc$. In the GZSL setting, we calculate the harmonic mean among seen and unseen classes: $H = (2 \times S \times U) / (S + U)$. $S$ and $U$ indicate the Top-1 accuracy of seen and unseen classes.

**Implementation Details.** We extend the first Vision Mamba-based model for ZSL, which differs from previous CNN-based and ViT-based models. Specifically, we leverage the VMamba-Small model pre-trained on ImageNet-1K to initialize our Vision Mamba Encoder. We implement our experiments in PyTorch and utilize the SGD optimizer (momentum $= 0.9$, weight decay $= 0.001$) with learning rate of $5 \times 10^{-4}$ on a single NVIDIA A100 GPU. All models are trained with a batch size of 16. In our method, we empirically set $\{\lambda_{sc}, \lambda_{col}\}$ to $\{1.0,0.3\}$, $\{0.2,0.35\}$, and $\{0.0,0.98\}$ for CUB, SUN, and AWA2, respectively.

### Comparison with State-of-the-Art Methods

We compare the proposed ZeroMamba with SOTA CNN-based and ViT-based methods, with the results shown in Tab. 1. In the CZSL task, ZeroMamba achieves the best accuracy of 80.0% and 72.4% on CUB and SUN, respectively. For the coarse-grained dataset (*i.e.*, AWA2), ZeroMamba remains competitive ($Acc = 71.9\%$). For the challenging GZSL task, ZeroMamba surpasses the previous best methods (*e.g.*, ZSLViT (Chen et al. 2024c) on CUB and I2DFormer+ (Naeem et al. 2024) on AWA2) that use ViT-Base backbone with more parameters. Additionally, compared to large-scale vision-language models (*e.g.*, CLIP (Radford et al. 2021)), our method improves $H$ by 19.2% on CUB. It is worth noting that ZeroMamba significantly improves the accuracy of unseen classes while maintaining superior performance on seen classes. Furthermore, the results in Fig. 1 strongly demonstrate that ZeroMamba attains more efficient computation than CNN-based and ViT-based methods, *i.e.*, superior performance with fewer parameters.

| Model | Params (M) | CUB | | | | SUN | | | | AWA2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CZSL | GZSL | | | CZSL | GZSL | | | CZSL | GZSL | | |
| | | Acc | U | S | H | Acc | U | S | H | Acc | U | S | H |
| LocalVim-Tiny (2024) | 8.3 | 66.5 | 55.8 | 60.4 | 58.0 | 62.6 | 42.3 | 33.3 | 37.3 | 64.3 | 58.9 | 76.6 | 66.6 |
| Vim-Small (2024) | 25.9 | 64.7 | 58.2 | 52.1 | 55.0 | 64.4 | 58.5 | 21.2 | 31.2 | 62.2 | 58.2 | 79.1 | 67.1 |
| LocalVim-Small (2024) | 27.9 | 68.1 | 56.6 | 63.7 | 59.9 | 65.6 | 46.2 | 36.4 | 40.7 | 66.7 | 53.2 | 84.4 | 65.2 |
| LocalVMamba-Small (2024) | 50.1 | 74.0 | 64.6 | 64.3 | 64.4 | 69.4 | 51.5 | 37.4 | 43.3 | 64.2 | 55.9 | 84.4 | 67.3 |
| **ZeroMamba-Tiny (Ours)** | 31.4 | 78.4 | 71.6 | 74.3 | 72.9 | 69.9 | 52.0 | 41.1 | 45.9 | 67.7 | 61.2 | 87.2 | 71.9 |
| **ZeroMamba-Small (Ours)** | 51.3 | **80.0** | 72.1 | 76.4 | **74.2** | **72.4** | 56.5 | 41.4 | **47.7** | **71.9** | 67.9 | 87.6 | **76.5** |

Table 4: Comparing our ZeroMamba with different Vision Mamba models on CUB, SUN, and AWA2 benchmark datasets in the CZSL and GZSL settings. The best result is highlighted in **boldface** and underline.

| baseline | $\mathcal{L}_{sc}$ | SLP | GRL | CUB | | | | SUN | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | CZSL | GZSL | | | CZSL | GZSL | | |
| | | | | Acc | U | S | H | Acc | U | S | H |
| ✓ | ✓ | | | 75.7 | 73.2 | 62.9 | 67.6 | 70.6 | 49.5 | 41.3 | 45.1 |
| ✓ | ✓ | ✓ | | 76.2 | 73.0 | 65.4 | 69.0 | 71.3 | 46.0 | 45.6 | 45.8 |
| ✓ | ✓ | | ✓ | 76.9 | 73.7 | 65.8 | 69.6 | 71.8 | 55.4 | 40.0 | 46.5 |
| ✓ | | ✓ | ✓ | 77.6 | 59.0 | 80.6 | 68.1 | 71.2 | 50.3 | 43.4 | 46.6 |
| **ZeroMamba (full)** | ✓ | ✓ | ✓ | **80.0** | 72.1 | 76.4 | **74.2** | **72.4** | 56.5 | 41.4 | **47.7** |

Table 5: Ablation studies for different components of ZeroMamba on CUB and SUN datasets. We directly remove the classification head of VMamba and then insert an MLP with $\mathcal{L}_{sc}$ to map visual features to the semantic space as our baseline. SLP and GRL represent the Semantic-aware Local Projection and the Global Representation Learning modules, respectively. The best result is highlighted in **boldface** and underline.

## Experiments on Large-scale Dataset

To further evaluate the effectiveness and generalization capabilities of ZeroMamba, we conduct experiments on a challenging large-scale benchmark (*i.e.*, ImageNet). Due to the difficulty of this task, most recent work has excluded it. In this work, we follow TransZero++ (Chen et al. 2022a) to obtain semantic embeddings of class names using word2vec (Mikolov et al. 2013) with 300-dimensions. We randomly split the training/test set into 800/200. We only use 1/10 of the data for every class for training. The experimental results are in Tab. 2 for the CZSL task. We observe that ZeroMamba outperforms all methods, setting a new SOTA on ImageNet. Also, ZeroMamba exhibits an impressive zero-shot accuracy of 24.5% compared to I2DFormer+ ($Acc = 17.6\%$) (Naeem et al. 2024) which directly learns a class embedding from the document text with global and fine-grained alignment.

## Ablation Study and Analysis

In this section, we perform ablation studies and analyses to comprehensively demonstrate the effectiveness of our proposed ZeroMamba, including the effectiveness of different Vision Mamba models and components.

**Effectiveness of Different Vision Mamba Models.** To further assess the effectiveness of our proposed ZeroMamba, we conduct experiments on different Vision Mamba models. The comparison results are shown in Tab. 4. We map these models' visual features (*i.e.*, LocalVim, Vim and LocalVMamba) to the semantic space using an MLP for classification. Evidently, our ZeroMamba (using VMamba (Liu et al. 2024) as the visual encoder) achieves the best results on all datasets with only slightly increased parameters. Moreover, our method incorporates two semantic feature learning modules and fuses semantic information, demonstrating the rationality and effectiveness of our design.

**Effectiveness of Each Component.** To verify the effectiveness of our proposed components, we report the improvements of exerting components on our baseline model in Tab. 5. The results indicate that merely altering the classification head to a semantic-mapping MLP does not yield satisfactory performance (*i.e.*, baseline). When ZeroMamba lacks $\mathcal{L}_{sc}$ to exert semantic constraint, the $Acc/H$ significantly drops by 4.3%/6.1% and 1.8%/1.1% on CUB and SUN, respectively. This is because fine-grained datasets can't ensure semantic consistency, leading to severe overfitting of the classification results to seen classes. Moreover, both SLP and GRL bring notable performance gains, verifying that enhancing global features and utilizing local features benefit discriminative semantic learning. As a result, the model achieves $Acc/H$ improvements of 4.3%/6.6% and 1.8%/2.6% on CUB and SUN, respectively. In general, these results indicate the effectiveness of each component of ZeroMamba.

## Qualitative Results

**Visualization of Activation Maps.** We visualize the activation maps of our ZeroMamba alongside pioneering visual backbones, as shown in Fig. 3. CNN captures global areas, and ViT focuses on local object parts, while our ZeroMamba detects the semantic-related regions that are beneficial for classification. The part enclosed in a red rectangle displays a hard case where all three models perform suboptimally, yet ZeroMamba can still localize more semantic-related information. This shows the capability of ZeroMamba to handle complex visual-semantic relationships.

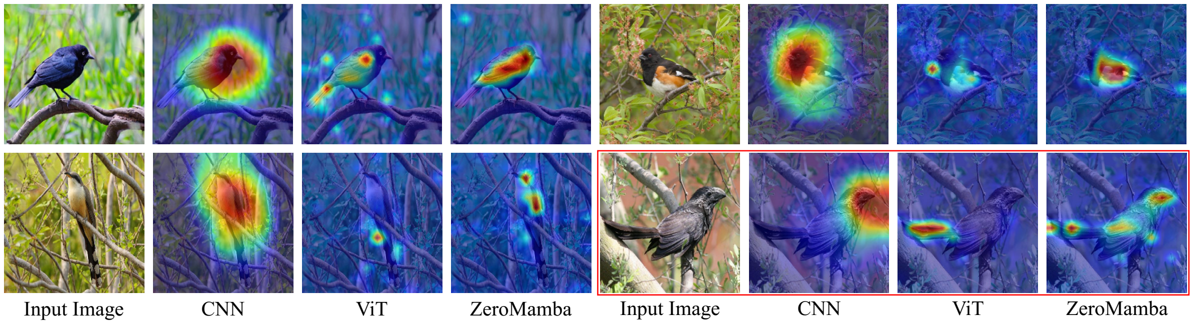


Figure 3: Visualization of the activation maps of different visual backbones, including CNN (*e.g.*, ResNet-101 (He et al. 2016)), ViT (*e.g.*, ViT-Base (Dosovitskiy et al. 2020)), and ZeroMamba. Our ZeroMamba can accurately capture the semantic-related information. We use CUB as an example, with the red box indicating challenging cases.

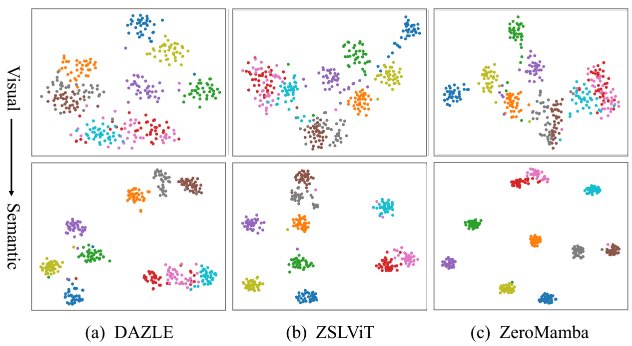


Figure 4: (Please Zoom in for details.) Visualizations with t-SNE embeddings of different methods produced by (a) DAZLE (2020), (b) ZSLViT (2024c), and (c) ZeroMamba (ours) in both visual and semantic spaces. The 10 colors denote 10 different classes randomly selected from CUB.

**Visualization of t-SNE Embeddings.** We also present t-SNE embeddings (2008) to compare our ZeroMamba with classic CNN-based and ViT-based methods (*i.e.*, DAZLE (2020) and ZSLViT (2024c)). We visualize both visual and semantic features. As shown in Fig. 4, it demonstrates that visual features are more scattered and mixed among classes, while semantic features are relatively cohesive and distinguishable. The comparison indicates that ZeroMamba optimizes visual and semantic representations, forming more distinct class boundaries in both visual and semantic spaces.
**Visualization of Effective Receptive Fields.** The Effective Receptive Field (ERF) (2016) plays a crucial role in visual tasks, as the output must respond to sufficiently large areas to capture semantic-related information from the entire image. In this work, we intuitively visualize the central pixel's ERF across different visual backbones. As shown in Fig. 5, we can observe that CNN (*e.g.*, ResNet-101 (2016)) exhibits local ERF, while ViT (*e.g.*, DeiT (2021)) and ZeroMamba demonstrate global ERFs. Furthermore, ZeroMamba shows higher intensity for the central pixel compared to ViT, indicating its superior capability in representing visual data.

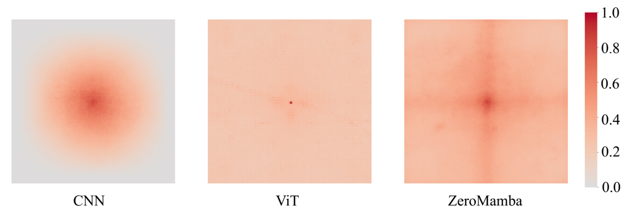


Figure 5: Comparison of effective receptive fields (ERF) (2016) between CNN (*e.g.*, ResNet-101 (2016)), ViT (*e.g.*, DeiT (2021)) and ZeroMamba. Pixels with higher intensity (darker color) indicate a stronger response to the central pixel. It is evident that ZeroMamba and ViT exhibit a global receptive field, while CNN only has a local one on CUB.

## Concluding Remarks

In this paper, we have advocated Vision Mamba for ZSL, a parameter-efficient framework called ZeroMamba. The method is straightforward: incorporating semantic information into network learning and discovering local and global discriminative semantic-related representations to form an end-to-end ZSL framework. Our main contribution shows that despite its simplicity, ZeroMamba consistently outperforms existing CNN-based and ViT-based methods that have been developed specifically for ZSL and are more complex. Thus, this paper provides strong evidence that ZeroMamba can serve as an alternative baseline for ZSL. Furthermore, our investigation highlights the strengths of Vision Mamba in handling complex visual-semantic interactions, offering valuable insights for future research and improvements in the promising ZSL field.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (62472381), the Fundamental Research Funds for the Zhejiang Provincial Universities (226-2024-00208) and the Earth System Big Data Platform of the School of Earth Sciences, Zhejiang University.

## References


Alamri, F.; and Dutta, A. 2021. Multi-head self-attention via vision transformer for zero-shot learning. *arXiv preprint arXiv:2108.00045*.

Cavazza, J.; Murino, V.; and Del Bue, A. 2023. No Adversaries to Zero-Shot Learning: Distilling an Ensemble of Gaussian Feature Generators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cetin, S. 2022. *Closed-form sample probing for training generative models in zero-shot learning*. Master's thesis, Middle East Technical University.

Chen, S.; Chen, S.; Xie, G.-S.; Shu, X.; You, X.; and Li, X. 2024a. Rethinking attribute localization for zero-shot learning. *SCIENCE CHINA Information Sciences*, 67(7).

Chen, S.; Fu, D.; Chen, S.; Hou, W.; You, X.; et al. 2024b. Causal Visual-semantic Correlation for Zero-shot Learning. In *ACM Multimedia 2024*.

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. *TPAMI*.

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*, volume 2, 3.

Chen, S.; Hong, Z.; Xie, G.-S.; Yang, W.; Peng, Q.; Wang, K.; Zhao, J.; and You, X. 2022c. MSDN: Mutually Semantic Distillation Network for Zero-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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. *arXiv preprint arXiv:2306.06931*.

Chen, S.; Hou, W.; Khan, S.; and Khan, F. S. 2024c. Progressive Semantic-Guided Vision Transformer for Zero-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23964–23974.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, 122–131.

Chen, S.; Xie, G.; Liu, 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*, 34: 16622–16634.

Chen, Z.; Huang, Y.; Chen, J.; Geng, Y.; Zhang, W.; Fang, Y.; Pan, J. Z.; and Chen, H. 2023b. Duet: Cross-modal semantic grounding for contrastive zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 405–413.

Chen, Z.; Zhang, P.; Li, J.; Wang, S.; and Huang, Z. 2023c. Zero-shot learning by harnessing adversarial samples. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4138–4146.

Christensen, A.; Mancini, M.; Koepke, A.; Winther, O.; and Akata, Z. 2023. Image-free classifier injection for zero-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19072–19081.

Dao, T.; and Gu, A. 2024. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In *International Conference on Machine Learning (ICML)*.

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

Feng, Y.; Huang, X.; Yang, P.; Yu, J.; and Sang, J. 2022. Non-generative generalized zero-shot learning via task-correlated disentanglement and controllable samples synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9346–9355.

Fu, C.; Wang, Y.; Zhang, J.; Jiang, Z.; Mao, X.; Wu, J.; Cao, W.; Wang, C.; Ge, Y.; and Liu, Y. 2024. MambaGesture: Enhancing Co-Speech Gesture Generation with Mamba and Disentangled Multi-Modality Fusion. *arXiv preprint arXiv:2407.19976*.

Gu, A.; and Dao, T. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2312.00752*.

Habib, G.; Saleem, T. J.; and Lall, B. 2023. Knowledge distillation in vision transformers: A critical review. *arXiv preprint arXiv:2302.02108*.

Han, Z.; Fu, Z.; Chen, S.; and Yang, J. 2021. Contrastive embedding for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2371–2381.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

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. In *CVPR*, 23627–23637.

Huang, T.; Pei, X.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2024. LocalMamba: Visual State Space Model with Windowed Selective Scan. *arXiv preprint arXiv:2403.09338*.

Huynh, D.; and Elhamifar, E. 2020. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4483–4493.

Kim, J.; Shim, K.; and Shim, B. 2022. Semantic feature extraction for generalized zero-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 1166–1173.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9306–9315.

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

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2013. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3): 453–465.

Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; and Qiao, Y. 2024a. Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2403.06977*.

Li, S.; Singh, H.; and Grover, A. 2025. Mamba-nd: Selective state space modeling for multi-dimensional data. In *European Conference on Computer Vision*, 75–92. Springer.

Li, Y.; Luo, Y.; Wang, Z.; and Du, B. 2024b. Improving Generalized Zero-Shot Learning by Exploring the Diverse Semantics from External Class Names. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23344–23353.

Liu, M.; Li, F.; Zhang, C.; Wei, Y.; Bai, H.; and Zhao, Y. 2023. Progressive Semantic-Visual Mutual Adaption for Generalized Zero-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15337–15346.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; and Liu, Y. 2024. VMamba: Visual State Space Model. *arXiv preprint arXiv:2401.10166*.

Luo, W.; Li, Y.; Urtasun, R.; and Zemel, R. 2016. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15169–15179.

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*, 35: 12283–12294.

Naeem, M. F.; Xian, Y.; Gool, L. V.; and Tombari, F. 2024. I2DFormer+: Learning Image to Document Summary Attention for Zero-Shot Image Classification. *IJCV*, 1–17.

Park, J.; Kim, H.-S.; Ko, K.; Kim, M.; and Kim, C. 2024. VideoMamba: Spatio-Temporal Selective State Space Model. *arXiv preprint arXiv:2407.08476*.

Patterson, G.; and Hays, J. 2012. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2751–2758. IEEE.

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

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *IJCV*, 115: 211–252.

Schonfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; and Akata, Z. 2019. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8247–8255.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, C.; Min, S.; Chen, X.; Sun, X.; and Li, H. 2021. Dual Progressive Prototype Network for Generalized Zero-Shot Learning. *Advances in Neural Information Processing Systems*, 34: 2936–2948.

Wang, X.; Ye, Y.; and Gupta, A. 2018. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6857–6866.

Wang, Z.; Liang, J.; He, R.; Xu, N.; Wang, Z.; and Tan, T. 2023. Improving zero-shot generalization for clip with synthesized prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 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,*.

Weng, J.; Yan, Z.; Tai, Y.; Qian, J.; Yang, J.; and Li, J. 2024. MambaLLIE: Implicit Retinex-Aware Low Light Enhancement with Global-then-Local State Space. *arXiv preprint arXiv:2405.16105*.

Xian, Y.; Sharma, S.; Schiele, B.; and Akata, Z. 2019. f-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10275–10284.

Xu, Q.; Liu, X.; Zhu, L.; Lin, G.; Long, C.; Li, Z.; and Zhao, R. 2024. Hybrid mamba for few-shot segmentation. *arXiv preprint arXiv:2409.19613*.

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*, 33: 21969–21980.

Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; and Akata, Z. 2022. Attribute prototype network for any-shot learning. *International Journal of Computer Vision*, 130(7): 1735–1753.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*.