Progressive Semantic-Guided Vision Transformer for Zero-Shot Learning

Shiming Chen[1], Wenjin Hou[2], Salman Khan[1,3], Fahad Shahbaz Khan[1,4]
[1]Mohamed bin Zayed University of AI [2]Huazhong University of Science and Technology
[3]Australian National University [4]Linköping University
{shimingchen, houwj17}@gmail.com {salman.khan, fahad.khan}@mbzuai.ac.ae

Abstract

Zero-shot learning (ZSL) recognizes the unseen classes by conducting visual-semantic interactions to transfer semantic knowledge from seen classes to unseen ones, supported by semantic information (e.g., attributes). However, existing ZSL methods simply extract visual features using a pre-trained network backbone (i.e., CNN or ViT), which fail to learn matched visual-semantic correspondences for representing semantic-related visual features as lacking of the guidance of semantic information, resulting in undesirable visual-semantic interactions. To tackle this issue, we propose a progressive semantic-guided vision transformer for zero-shot learning (dubbed ZSLViT). ZSLViT mainly considers two properties in the whole network: i) discover the semantic-related visual representations explicitly, and ii) discard the semantic-unrelated visual information. Specifically, we first introduce semantic-embedded token learning to improve the visual-semantic correspondences via semantic enhancement and discover the semantic-related visual tokens explicitly with semantic-guided token attention. Then, we fuse low semantic-visual correspondence visual tokens to discard the semantic-unrelated visual information for visual enhancement. These two operations are integrated into various encoders to progressively learn semantic-related visual representations for accurate visual-semantic interactions in ZSL. The extensive experiments show that our ZSLViT achieves significant performance gains on three popular benchmark datasets, i.e., CUB, SUN, and AWA2.

1. Introduction

Zero-shot learning (ZSL), aiming to recognize unseen classes by exploiting the intrinsic semantic relatedness between seen and unseen categories during training [26, 28, 38, 56], has achieved significant progress. Inspired by the way humans learn unknown concepts, semantic information (e.g., attributes [27]) shared by seen and unseen classes is employed to support knowledge transfer from seen classes to unseen ones. Targeting this goal, ZSL conducts effective visual-semantic interactions between visual and semantic spaces to align them. For example, discovering the semantic representations in visual spaces and matching them with the semantic information. As such, exploring the shared semantic knowledge between the visual and semantic spaces is essential.

Existing ZSL methods [2, 3, 9, 11–13, 25, 31, 35, 46, 49, 51, 52] typically take a network backbone (convolutional neural network (CNN) or vision Transformer (ViT)) pre-trained on ImageNet [42] to extract visual features. However, the network backbone fails to learn matched visual-semantic correspondences for representing semantic-related visual features, because they lack sufficient guidance of semantic information, as shown in Fig. 1(a). As shown in Fig. 1(c1), the CNN backbone learns the representations focused on the meaningless background information or the whole object. Although some methods adopt the attention mechanism to enhance the CNN visual features via attribute localization [8, 9, 33, 37, 52, 53, 58], they only obtain the sub-optimal visual representations as the visual spaces are almost fixed after the CNN backbone learning.

Thanks to the strong capability of modeling long-range association of whole image, some methods simply take the pre-trained ViT to extract visual features for ZSL tasks [2, 3, 13, 31, 35] and achieve better performance than CNN features-based ZSL methods. Unfortunately, they localize the semantic attribute incorrectly without explicit guidance of semantic information, which also fails to represent the correspondences between visual-semantic features, as shown in Fig. 1(c2). Therefore, the visual features learned by CNN or ViT backbone cannot be well related to their corresponding semantic attributes (*e.g.*, the '*neck color yellow*' of Yellow_Headed_Blackbird), resulting in undesirable visual-semantic interactions. Consequently, the semantic knowledge transferring in ZSL is limited, thus leading to inferior ZSL performance. As such, properly constructing matched visual-semantic correspondences for learning semantic-related visual features in the feature extraction network for advancing ZSL is highly necessary.

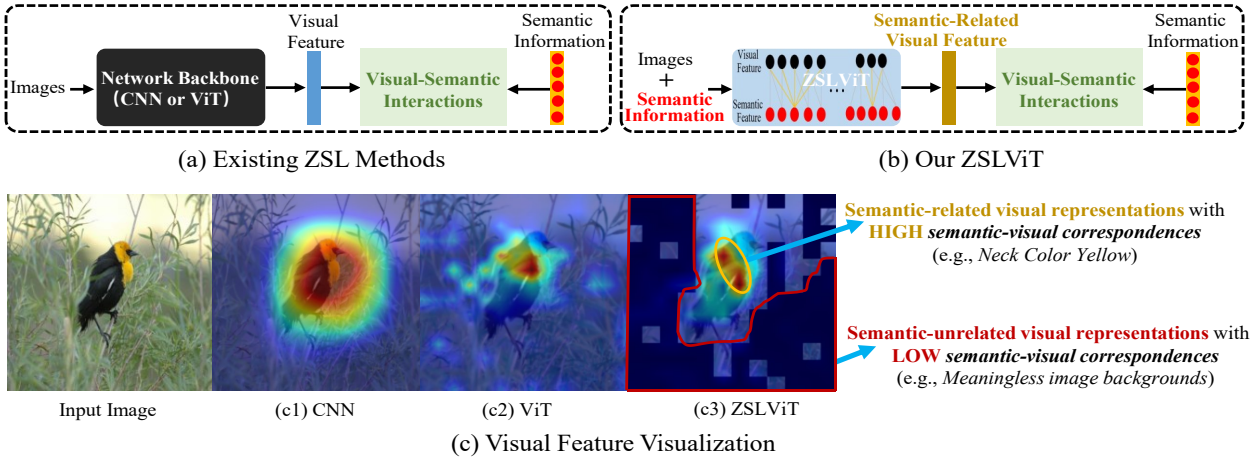


Figure 1. Motivation Illustration. (a) Existing ZSL methods simply take the pre-trained network backbone (*i.e.*, CNN or ViT) to extract visual features. (b) Our ZSLViT progressively learns semantic-visual correspondences to represent semantic-related visual features in the *whole network* for advancing ZSL. (c) The visual feature visualization. (*c1*) The heat map of visual features learned by CNN backbone (*e.g.*, ResNet101 [20]) includes the whole object and background, which fail to capture the semantic attributes. (*c2*) The attention map of visual features learned by the standard ViT [16], which localizes the semantic attributes incorrectly. (*c3*) The attention map learned by our ZSLViT, which discovers the semantic-related visual representations and discards the semantic-unrelated visual information according to semantic-visual correspondences.

To learn semantic-related visual features for desirable visual-semantic interactions, we propose a *progressive semantic-guided vision transformer* specifically for ZSL, dubbed ZSLViT. As shown in Fig. 1(b) and (c3), ZSLViT takes two considerations in the whole network: i) how to discover the semantic-related visual representations explicitly, and ii) how to discard the semantic-unrelated visual information (*e.g.*, meaningless image backgrounds). We first introduce a semantic-embedded token learning (SET) mechanism consisting of a semantic enhancement and a semantic-guided token attention. The semantic enhancement improves semantic-visual correspondences for visual tokens via visual-semantic consistency learning and semantic embedding. Accordingly, the semantic-guided token attention explicitly discovers the semantic-related visual tokens, which have high visual-semantic correspondences and are preserved into the next layer. Then, we introduce visual enhancement (ViE) to fuse the visual tokens with low visual-semantic correspondences into one new token for purifying the semantic-unrelated information. Thus, the semantic-unrelated visual information is discarded for enhancing visual features. These two operations are integrated into various encoders to progressively learn semantic-related visual representations, enabling desirable visual-semantic interactions for ZSL task. The quantitative and qualitative results demonstrate the superiority and great potential of ZSLViT.

Our main contributions can be summarized:

- We propose a progressive semantic-guided visual transformer, dubbed ZSLViT, which learns matched visual-semantic correspondences for representing semantic-related visual representations, enabling effective visual-semantic interactions for ZSL.
- We introduce semantic-embedded token learning and visual enhancement to discover the semantic-related visual representations explicitly and discard the semantic-unrelated visual information, respectively.
- We conduct extensive experiments on three challenging benchmark datasets (*i.e.*, CUB [48], SUN [39], and AWA2 [50]) under both conventional and generalized ZSL settings. Results show that our ZSLViT achieves significant improvements and new state-of-the-art results.

2. Related Works

Zero-Shot Learning. ZSL typically transfers semantic knowledge from seen classes to unseen ones by conducting visual-semantic interactions, and thus the unseen classes can be recognized [1, 6, 11, 18, 19, 49, 51]. There two methods are typically adopted, *i.e.*, embedding-based methods [7–9, 13, 23, 31, 32] and generative methods [10, 12, 17, 21, 24, 36, 51]. Embedding-based methods map visual features into semantic space and match them with their corresponding semantic prototypes by nearest-neighbor matching. The generative ZSL methods learn a generator conditioned by the semantic prototypes to synthesize the visual features for unseen classes, which are utilized to train a supervised clas-

sifier (*e.g.*, softmax). Different from the generic image classification task that classifies classes based on the semantic-unrelated labels, ZSL aims to classify the unseen class samples according to the semantic prototypes that are represented by the specific semantic attributes. Thus, discovering semantic-related visual representations and discarding the semantic-unrelated visual information to conduct effective semantic knowledge transferring from seen classes to unseen ones for ZSL is very necessary. These methods take pre-trained CNN backbone (*e.g.*, ResNet101) to extract the global visual features, which cannot accurately capture the semantic information of visual appearances (*e.g.*, the '*neck color yellow*' of Yellow_Headed_Blackbird), resulting in undesirable visual-semantic interactions for semantic knowledge transferring. Thus their results are essentially limited.

Although some methods take attention mechanism [9, 33, 52, 53, 58] to refine the extracted visual features from CNN backbone, they obtain the sub-optimal visual representations as the visual spaces are almost fixed after the CNN backbone learning. Considering that vision Transformer (ViT) [16, 30, 41, 55] has the advantages of learning implicitly semantic-context visual information using self-attention mechanisms, in this work, we devise a novel ViT backbone to progressively learn the semantic-related visual features under the guidance of semantic information in the whole network. This encourages the model to conduct effective visual-semantic interactions in ZSL.

Vision Transformer. Transformers [44] have achieved significant progress in computer vision recently due to their strong capability of modeling long-range relation, *e.g.*, image classification [22], object detection [4], and semantic segmentation [14]. Vision Transformer (ViT) [16] is the first pure Transformer backbone introduced for image classification, and it is further employed for other vision tasks [41]. Some methods simply take ViT to extract the global visual features for ZSL tasks [2, 3, 13, 31]. Unfortunately, they fail to construct matched visual-semantic correspondences explicitly with semantic information and cannot well explore the potential of ViT for ZSL. In this work, we aim to design a ViT backbone specifically for advancing ZSL considering two properties: i) discover the semantic-related visual representations explicitly, and ii) discard the semantic-unrelated visual information.

3. Semantic-Guided Vision Transformer

The task of ZSL is formulated as follows. Let we have C^s seen classes data $\mathcal{D}^s = \{(x_i^s, y_i^s)\}$, where $x_i^s \in \mathcal{X}$ denotes the i-th sample, and $y_i^s \in \mathcal{Y}^s$ is its class label. The $\mathcal{D}^s$ is split into a training set $\mathcal{D}_{tr}^s$ and a testing set $\mathcal{D}_{te}^s$ following [50]. Meanwhile, we have C^u unseen classes data $\mathcal{D}_{te}^u = \{(x_i^u, y_i^u)\}$, where $x_i^u \in \mathcal{X}$ is the sample of unseen classes, and $y_i^u \in \mathcal{Y}^u$ is its class label. Thus, the total class number in one dataset is $c \in \mathcal{C}^s \cup \mathcal{C}^u$. The semantic prototypes are represented by vectors. Each vector corresponds to one class. Each semantic vector $z^c = [z^c(1), \ldots, z^c(A)]^\top \in \mathbb{R}^{|A|}$ is with the $|A|$ dimension, where each dimension is a semantic attribute value annotated by human. In the conventional ZSL setting (CZSL) setting, we learn a classifier only for unseen classes (*i.e.*, $f_{\mathrm{CZSL}} : \mathcal{X} \rightarrow \mathcal{Y}^U$). Differently in generalized ZSL (GZSL), we learn a classifier for both seen and unseen classes (*i.e.*, $f_{\mathrm{GZSL}} : \mathcal{X} \rightarrow \mathcal{Y}^U \cup \mathcal{Y}^S$).

In the following, we introduce our ZSLViT specifically. As shown in Fig. 2, the novel operations of ZSLViT include a semantic-embedding token learning (SET) and a visual enhancement (ViE). These two operations are integrated into various encoders between the multi-head self-attention and feed-forward network layers to progressively learn semantic-related visual representations, enabling accurate visual-semantic interactions for ZSL. At the end of this section, we demonstrate how we perform zero-shot prediction using the semantic-related visual representations learned by ZSLViT.

3.1. Semantic-Embedded Token Learning

Semantic-embedded token learning (SET) is employed to discover semantic-related visual features. SET consists of a semantic enhancement module and a semantic-guided token attention module. The semantic enhancement explicitly improves the visual-semantic correspondences with visual-semantic consistency learning and semantic embedding. The semantic-guided token attention discovers the semantic-related visual representations based on the semantic-enhanced tokens.

Semantic Enhancement. We first conduct visual-semantic consistency learning based on the visual features and semantic vectors. Here, we take the $[cls]$ token (*i.e.*, $token[cls]$) as the visual features due to it pays more attention (*i.e.*, having a larger attention value) on class-specific tokens to represent one image for classification [5, 44]. Specifically, we take two multi-layer perceptrons (MLP), *i.e.*, MLP_{V2S} and MLP_{S2V}, to map the features from visual space to semantic space (*i.e.*, $Visual \rightarrow Semantic$) and from semantic space to visual space (*i.e.*, $Semantic \rightarrow Visual$), respectively. As such, the MLP_{V2S} and MLP_{S2V} can effectively improve their consistency.

$$Visual \rightarrow Semantic : \tilde{z} = MLP_{V2S}(Token[cls]), \quad (1)$$

$$Semantic \rightarrow Visual : \widetilde{Token}[cls] = MLP_{S2V}(z), \quad (2)$$

where $\tilde{z}$ is the reconstructed semantic vector from visual space, and $\widetilde{Token}[cls]$ is the reconstructed visual feature from semantic space. To enable visual-semantic consistency learning, we take a semantic reconstruction loss $\mathcal{L}_{SR}$ and a visual reconstruction loss $\mathcal{L}_{VR}$ to guide the optimiza-

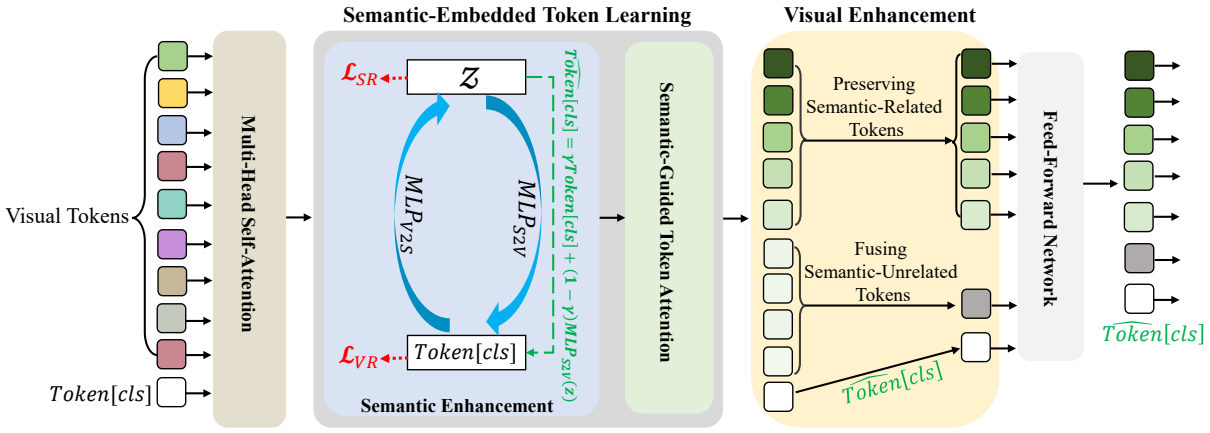


Figure 2. A single ZSLViT encoder. ZSLViT encoder includes a semantic-embedded token learning (SET) and a visual enhancement (ViE) between the multi-head self-attention and feed-forward network layers. SET improves the visual-semantic correspondences via semantic enhancement and discovers the semantic-related visual tokens explicitly with semantic-guided token attention. ViE fuses the visual tokens of low visual-semantic correspondences to discard the semantic-unrelated visual information for visual tokens enhancement. The ZSLViT encoder are integrated into various layers to progressively learn semantic-related visual representations, enabling effective visual-semantic interactions for ZSL.

tion of MLP_{V2S} and MLP_{S2V}, formulated as:

$$\mathcal{L}_{SR} = \|z - \tilde{z}\|_1, \tag{3}$$

$$\mathcal{L}_{VR} = \|Token[cls] - \widetilde{Token}[cls]\|_1. \tag{4}$$

Notably, we set a larger weight on $\mathcal{L}_{VR}$ than $\mathcal{L}_{SR}$ as we mainly aim to enhance semantic information into visual representations for subsequent learning. This will also facilitate stable optimization for MLP_{V2S} and MLP_{S2V}.

Considering the semantic vectors are informative attribute representations, we explicitly boost the semantic information into visual features for semantic enhancement via semantic embedding. Specifically, we concatenate the reconstructed visual features from semantic space with the real visual features $Token[cls]$:

$$\widehat{Token}[cls] = \gamma Token[cls] + (1-\gamma)MLP_{S2V}(z), \tag{5}$$

where γ is a combination coefficient, which is set to a relatively large value for progressive enhancement, enabling stable learning for ZSLViT. $\widehat{Token}[cls]$ is the semantically enhanced token, which is served as a new $Token[cls]$ (*i.e.*, $Token[cls] = \widehat{Token}[cls]$) to update the original $Token[cls]$ for subsequent learning. We should note that semantic embedding is only used in the training stage but not in the inference stage.

Semantic-Guided Token Attention. After semantic enhancement, we take semantic-guided token attention to identify the semantic-related and semantic-unrelated visual tokens based on $\widehat{Token}[cls]$. Specifically, we perform the interaction between the $\widehat{Token}[cls]$ and other visual tokens, where the packed outputs of the multi-head self-attention layer are used as keys (K) and values (V), and $\widehat{Token}[cls]$ is the query vector. It is defined as:

$$f(x) = \text{Softmax}\left(\frac{\widehat{Token}[cls] \cdot K^\top}{\sqrt{d}}\right) V = a \cdot V, \tag{6}$$

where d is a scaling factor. $a = \{a_1, a_2, \cdots, a_n\}$ (n is the number of input visual tokens in a ZSLViT encoder) is *attention scores being visual-semantic correspondences* from $[cls]$ token to other visual tokens. Accordingly, $f(x)$ is a linear combination of the value vectors $V = \{v_1, v_2, \cdots, v_n\}$. Since v_i comes from the i-th visual token, the attention score a_i determines how much information of the i-th visual token is embedded into the output of $[cls]$ token. It is natural to assume that the visual-semantic correspondence a_i indicates the importance of the i-th token corresponding to the semantic information for visual representations. To this end, ZSLViT can effectively localize the image regions most relevant to semantic attributes for discovering the semantic-related visual features, as shown in Fig. 1(c3).

3.2. Visual Enhancement

Visual enhancement (ViE) is devised to discard the semantic-unrelated visual features for enhancing visual features further. According to the visual-semantic correspondences a in Eq. 6, ZSLViT can easily determine the semantic-related visual tokens (*i.e.*, with the Top-k largest a, and the indices set denoted as $\mathcal{P}$), and the semantic-unrelated visual tokens(*i.e.*, with the $n-k$ smallest a, and the indices set denoted as $\mathcal{N}$). We take a hyper-parameter $\kappa = k/n$ to determine $\mathcal{P}$ and $\mathcal{N}$. Since visual-semantic

interactions in ZSL rely on semantic-related visual information, we can preserve the semantic-related visual tokens and discard the semantic-unrelated visual tokens to alleviate the negative effects of meaningless visual information (*e.g.*, the background of image). Thus the visual features are enhanced to enable effective visual-semantic interactions in ZSL. Considering ZSLViT cannot completely learn the accurate semantic-related visual representation in an encoder at one time, we fuse the semantic-unrelated visual tokens at the current stage to supplement semantic-related ones:

$$T(x) = \{f(x)_i\}_{i=1}^{\mathcal{P}} \cup \sum_{j \in \mathcal{N}} a_j f(x)_j, \tag{7}$$

$T(x)$ is the semantic-related visual features in the current encoder and is used for subsequent learning in the feed-forward network layer and next encoder. Thus, ZSLViT purifies the visual tokens in various encoders to discard the meaningless visual information progressively. Meanwhile, ZSLViT can be effectively lightened to reduce computational costs, enabling model acceleration.

3.3. Model Optimization

We now introduce the optimization objectives of our ZSLViT. First, ZSLViT conducts semantic-embedded token learning in various layers/encoders (indexed by S), where include $\mathcal{L}_{SR}$ (Eq. 3) and $\mathcal{L}_{SR}$ (Eq. 4). Assuming we are dealing with a minibatch of B samples $x_i \in \mathcal{D}_{tr}^s$, it can be formulated as:

$$\mathcal{L}_{SET} = \frac{1}{B}\frac{1}{S}\sum_{i=1}^{B}\sum_{s=1}^{S}(\lambda_{SR}\mathcal{L}_{SR}^s(x_i) + \lambda_{VR}\mathcal{L}_{VR}^s(x_i)). \tag{8}$$

Further, ZSLViT also trains the prediction module at the last layer such that it can produce favorable predictions and fine-tune the backbone to make it adapt to semantic-related visual feature learning. Since our ZSLViT is an embedding-based model, we should map the visual features (*i.e.*, the $token[cls]$ in the last layer) into their corresponding semantic vectors, *i.e.*, $\phi(x_i) = Token[cls]^\top W_{V2S}$, where W_{V2S} is a learnable mapping matrix. Following [8], we also employ the attribute-based cross-entropy loss for optimization:

$$\mathcal{L}_{pre} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(\phi(x_i) \times z^c\right)}{\sum_{\hat{c}\in\mathcal{C}^s}\exp\left(\phi(x_i) \times z^{\hat{c}}\right)}. \tag{9}$$

Different from existing embedding-based ZSL methods [2, 3, 7, 8, 53] that take an additional self-calibration loss to tackle the seen-unseen bias problem [11] during training, our ZSLViT can automatically avoid this issue as the discovered semantic-related visual features have good generalization from seen classes to unseen ones.

To this end, the overall loss function of ZSLViT is defined as:

$$\mathcal{L}_{ZSLViT} = \mathcal{L}_{SET} + \mathcal{L}_{pre}. \tag{10}$$

As the $\mathcal{L}_{pre}$ is the base loss in embedding-based ZSL, we set its weight to one.

3.4. Zero-Shot Prediction

We conduct zero-shot prediction in the inference stage. We first obtain the embedding features $\phi(x_i)$ of a test instance x_i from testing set (*i.e.*, $x_i \in \mathcal{D}_{te}^s \cup \mathcal{D}_{te}^u$) in the semantic space. Then, we take an explicit calibration to predict the test label of x_i, which is formulated as:

$$c^* = \arg\max_{c\in\mathcal{C}^u/\mathcal{C}} softmax(\phi(x_i) \times z^c) + \mathbb{I}_{[c\in\mathcal{C}^u]}. \tag{11}$$

$\mathbb{I}_{[c\in\mathcal{C}^u]}$ is an indicator function (*i.e.*, it is τ when $c \in \mathcal{C}^u$, otherwise zero, where τ is a hyper-parameter to control the calibration degree). We empirically set τ to 0.4 for all datasets. $\mathcal{C}^u/\mathcal{C}$ corresponds to the CZSL/GZSL setting, respectively. Since we do not use the unlabeled samples of unseen classes during training, our ZSLViT is an inductive method.

4. Experiments

Benchmark Datasets. We conduct extensive experiments on three popular ZSL benchmarks, including two fine-grained datasets (e.g., CUB [48] and SUN [39]) and a coarse-grained dataset (e.g., AWA2 [50]), to verify our ZSLViT. In specific, CUB includes 11,788 images of 200 bird classes (seen/unseen classes = 150/50) captured with 312 attributes. SUN consists of 14,340 images from 717 scene classes (seen/unseen classes = 645/72) captured with 102 attributes. AWA2 has 37,322 images from 50 animal classes (seen/unseen classes = 40/10) captured with 85 attributes.

Evaluation Protocols. Following [50], we measure the top-1 accuracy both in the CZSL and GZSL settings. In the CZSL setting, we only measure the accuracy of the test samples from the unseen classes, *i.e.*, $\boldsymbol{acc}$. In the GZSL setting, we compute the accuracy of the test samples from both the seen classes (denoted as $\boldsymbol{S}$) and unseen classes (denoted as $\boldsymbol{U}$). To generally evaluate the performance, the harmonic mean between seen and unseen classes is a main evaluation protocol in the GZSL setting, defined as $\boldsymbol{H = (2 \times S \times U)/(S + U)}$.

Implementation Details. We use the training splits proposed in [49]. For a fair comparison, we take the ViT-base model [43] pre-trained on ImageNet-1k as a baseline and for initialization. The MLP_{S2V} and MLP_{V2S} has multiple hidden layer with ReLU activation. We incorporate our semantic-embedded token learning and visual enhancement operations into the 4-th, 7-th and 10-th encoder by default for CUB and AWA2 (4-th and 7-th layers for SUN) to progressively learn the semantic-related visual features. We

Table 1. State-of-the-art comparisons on CUB, SUN, and AWA2 under the CZSL and GZSL settings. CNN features-based ZSL methods are categorized as ♯, and ViT features-based ZSL methods are categorized as ♭. The symbol "*" denotes attention-based ZSL methods using CNN features. The symbol † denotes ZSL methods based on large-scale vision-language model. The best and second-best results are marked in **Red** and **Blue**, respectively.

	Methods	Venue	CUB				SUN				AWA2			
			CZSL	GZSL			CZSL	GZSL			CZSL	GZSL		
			acc	U	S	H	acc	U	S	H	acc	U	S	H
♯	AREN* [52]	CVPR'19	71.8	38.9	**78.7**	52.1	60.6	19.0	38.8	25.5	67.9	15.6	**92.9**	26.7
♯	f-VAEGAN [51]	CVPR'19	61.0	48.4	60.1	53.6	64.7	45.1	38.0	41.3	71.1	57.6	70.6	63.5
♯	TF-VAEGAN [36]	ECCV'20	64.9	52.8	64.7	58.1	**66.0**	45.6	40.7	43.0	72.2	59.8	75.1	66.6
♯	LsrGAN [45]	ECCV'20	–	48.1	59.1	53.0	–	44.8	37.7	40.9	–	54.6	74.6	63.0
♯	DAZLE* [23]	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* [53]	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
♯	Composer* [24]	NeurIPS'20	69.4	56.4	63.8	59.9	62.6	**55.1**	22.0	31.4	71.5	62.1	77.3	68.8
♯	FREE [11]	ICCV'21	–	55.7	59.9	57.7	–	47.4	37.2	41.7	–	60.4	75.4	67.1
♯	GCM-CF [54]	CVPR'21	–	61.0	59.7	60.3	–	47.9	37.8	42.2	–	60.4	75.1	67.0
♯	HSVA [12]	NeurIPS'21	62.8	52.7	58.3	55.3	63.8	48.6	39.0	43.3	–	59.3	76.6	66.8
♯	MSDN* [8]	CVPR'22	76.1	68.7	67.5	68.1	65.8	52.2	34.2	41.3	70.1	62.0	74.5	67.7
♯	GEM-ZSL* [33]	CVPR'22	**77.8**	64.8	77.1	**70.4**	62.8	38.1	35.7	36.9	67.3	64.8	77.5	70.6
♯	SE-GZSL [25]	AAAI'22	–	53.1	60.3	56.4	–	45.8	40.7	43.1	–	59.9	80.7	68.8
♯	TransZero* [7]	AAAI'22	76.8	**69.3**	68.3	68.8	65.6	52.6	33.4	40.8	70.1	61.3	82.3	70.2
♯	VS-Boost [29]	IJCAI'23	–	68.0	68.7	68.4	–	49.2	37.4	42.5	–	–	–	–
♯	ICIS [15]	ICCV'23	60.6	45.8	73.7	56.5	51.8	45.2	25.6	32.7	64.6	35.6	**93.3**	51.6
♭	CLIP† [40]	ICML'21	–	55.2	54.8	55.0	–	–	–	–	–	–	–	–
♭	CoOp† [57]	IJCV'22	–	49.2	63.8	55.6	–	–	–	–	–	–	–	–
♭	I2DFormer-Wiki [35]	NeurIPS'22	45.4	35.3	57.6	43.8	–	–	–	–	**76.4**	**66.8**	76.8	71.5
♭	CoOp+SHIP† [47]	ICCV'23	–	55.3	58.9	57.1	–	–	–	–	–	–	–	–
♭	I2MVFormer-Wiki [34]	CVPR'23	42.1	32.4	63.1	42.8	–	–	–	–	**73.6**	**66.6**	82.9	**73.8**
♭	DUET [13]	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
♭	**ZSLViT (Ours)**	–	**78.9**	**69.4**	**78.2**	**73.6**	**68.3**	**45.9**	**48.4**	**47.3**	70.7	66.1	84.6	**74.2**

use the Adam optimizer with hyper-parameters (β_1 = 0.9, β_2 = 0.999) to optimize our model. The batch size is set to 32. We empirically set γ and κ to 0.9 for all datasets. We perform experiments on a single NVIDIA Tesla V100 graphic card with 32GB memory. We use PyTorch[1] for the implementation of all experiments.

4.1. Experimental Results

Results of Conventional Zero-Shot Learning. Table 1 shows the results of various methods on various datasets in the CZSL setting. Results show that the ViT features-based ZSL methods obtain better performances on all datasets than the CNN features-based ZSL methods. This is because ViT has obvious advantages of learning implicitly semantic-context visual information using a self-attention mechanism, which is well beneficial for ZSL task. Compared to all CNN features-based and ViT features-based methods, our ZSLViT consistently achieves the best performance of 78.9% and 68.3% on CUB and SUN datasets, respectively. Our ZSLViT obtains significant performance gains than the other ViT features-based methods [2, 13, 34, 35] on fine-grained datasets, *i.e.*, by 6.6% and 3.9% on CUB and SUN respectively. Our significant performance improvements demonstrate that ZSLViT effectively learns the semantic-related visual features for desirable semantic knowledge transferring from seen classes to unseen ones. Since the coarse-grained dataset (*e.g.*, AWA2) has a number of visual-unrelated semantic attributes for class descriptions (*e.g.*, "*eat fish*"), ZSLViT cannot achieve similar improvements on the AWA2.

Results of Generalized Zero-Shot Learning. Table 1 presents the GZSL performances of various methods, including CNN features-based and ViT features-based methods. Similar to the CZSL setting, ViT features-based ZSL methods perform better than CNN features-based methods generally on all datasets in the GZSL setting. Although some CNN features-based methods [7, 8, 23, 24, 33, 52, 53, 58] take attention mechanism to localize semantic attribute for enhancing CNN features further, the visual space is relatively fixed after CNN backbone training, resulting in sub-optimal performance. This indicates the advantages of ViT in ZSL task further. Compared to all various methods, our ZSLViT obtains the best results on all datasets, *e.g.*, 73.6%, 47.3% and 74.2% on CUB, SUN and AWA2, respectively. Especially, our ZSLViT outperforms the latest ViT features-based ZSL method (*i.e.*, I2MVFormer [34]) by a large margin, resulting in harmonic mean improvements by 30.8%. Notably, our ZSLViT also significantly outperforms the large-scale vision-language based ZSL methods (*e.g.*, CLIP [40] and CoOP [57]). These results demon-

[1] https://pytorch.org/

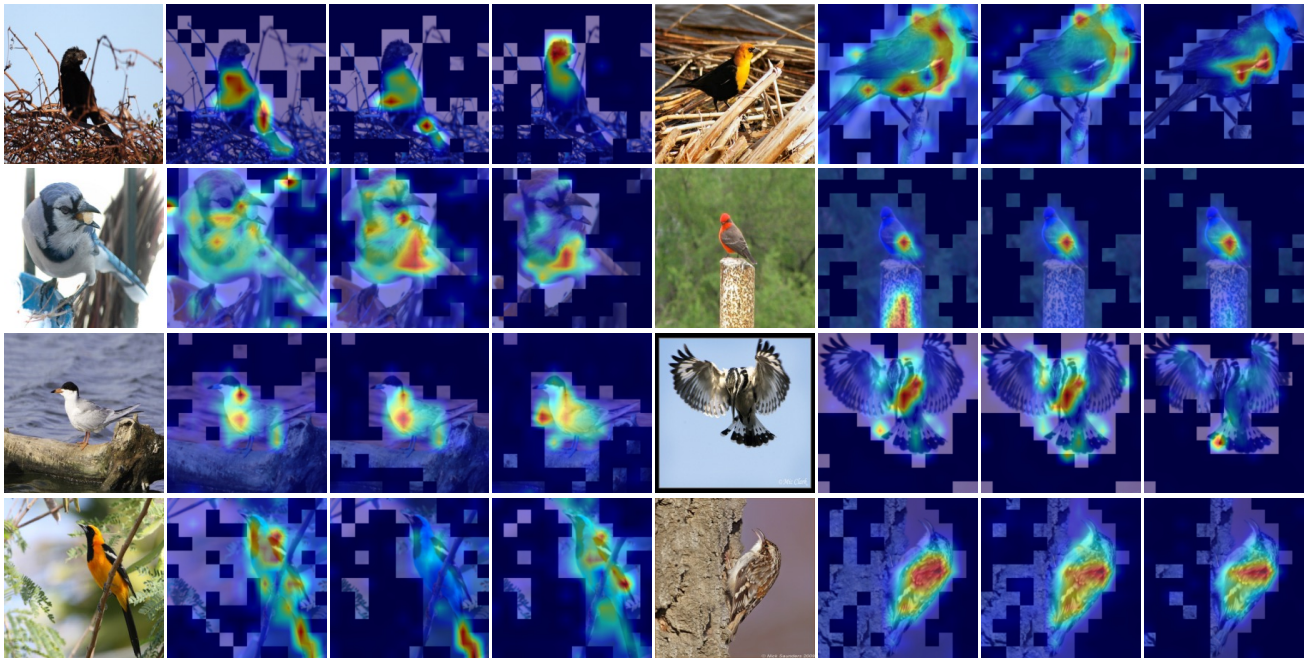


Figure 3. Visualizations of attention mask and map of our ZSLViT in various layers. The masked regions represent the semantic-unrelated visual tokens with low visual-semantic correspondences, which are fused into a new token for subsequent learning. The highlighted attention maps are the semantic-related visual tokens with high visual-semantic correspondences, which are preserved to next layer. Results show that ZSLViT can accurately identify the semantic-relateds/unrelated visual tokens in images for visual enhancement.

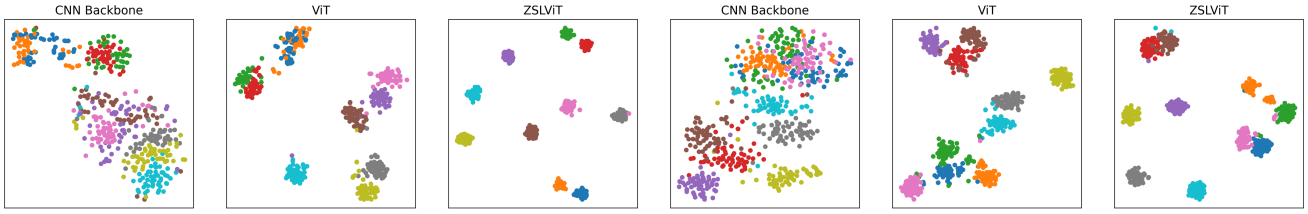
Table 2. Ablation studies for different components of ZSLViT on the CUB and AWA2 datasets.

Method	CUB		AWA2	
	acc	H	acc	H
ZSLViT w/o SET ($\mathcal{L}_{VR}$)	76.4	63.7	61.4	66.3
ZSLViT w/o SET ($\mathcal{L}_{SR}$)	78.0	72.8	70.2	73.6
ZSLViT w/o SET (Eq. 5)	76.2	69.7	69.9	71.2
ZSLViT w/o ViE	77.9	72.4	70.5	73.9
ZSLViT (full)	**78.9**	**73.6**	**70.7**	**74.2**

strate that our ZSLViT can sufficiently explore the potential of ViT to learn semantic-related visual features for ZSL.

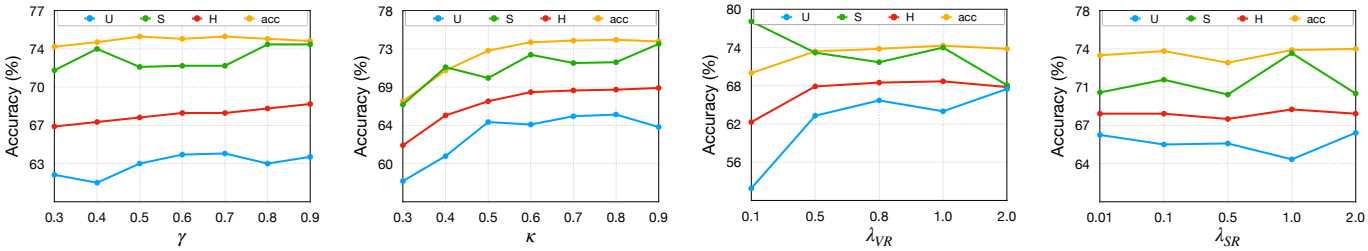
Ablation Study. We conduct ablation studies to evaluate the effectiveness of our ZSLViT in terms of the visual regression loss in SET (denoted as SET ($\mathcal{L}_{VR}$)), semantic regression loss in SET (denoted as SET ($\mathcal{L}_{SR}$)), semantic embedding in SET (denoted as SET (Eq. 5)), and ViE. Results are shown in Table 2. ZSLViT performs poorer slightly than its full model when no semantic reconstruction loss is used in SET, *i.e.*, the acc/harmonic mean drop by 0.9%/0.8% on CUB and 0.5%/0.6% on AWA2. If visual reconstruction loss is not used in SET, ZSLViT achieves very poor results compared to its full model, *i.e.*, the acc/harmonic mean drops by more than 2.5%/9.9% on CUB and 9.3%/7.9% on AWA2. These results show that visual reconstruction loss $\mathcal{L}_{VR}$ is essential for semantic-embedded token learning, because $\mathcal{L}_{VR}$ encourages SET to incorporate semantic information into visual tokens. This is typically neglected by existing ViT features-based ZSL methods [2,13,34,35]. Moreover, ZSLViT cooperates with semantic embedding to improve its performance of acc/harmonic mean by 2.7%/3.9% on CUB and 0.8%/3.0% on AWA2, respectively. ViE further improves the performance of ZSLViT by discarding the semantic-unrelated visual tokens. Generally, these results indicate the effects of various components of ZSLViT.

Attention Mask and Map Visualization. To intuitively show the effectiveness of our ZSLViT at learning semantic-related visual features for advancing ZSL, we visualize the attention mask (the tokens with low visual-semantic correspondences) and attention maps learned by ZSLViT on CUB. Results are shown in Fig. 3. Thanks to the semantic-embedded token learning that identifies the semantic-related/unrelated tokens according to the semantic-guided token attention, ZSLViT effectively i) preserves the semantic-related visual tokens of high visual-semantic correspondences (*e.g.*, the class-related attributes), and ii) discards the semantic-unrelated visual tokens of low visual-semantic correspondences (*e.g.*, the meaningless image background) by fusing them into a new token for subsequent learning. Accordingly, ZSLViT progressively learns



(a) Seen Classes (b) Unseen Classes

Figure 4. t-SNE visualizations of visual features for (a) seen classes and (b) unseen classes, learned by the CNN backbone (*e.g.*, ResNet101 [20]), standard ViT [43] and our ZSLViT. The 10 colors denote 10 different seen/unseen classes randomly selected from CUB. (Best viewed in color)



(a) Varying γ effect (b) Varying κ effect (c) Varying λ_{VR} effect (d) Varying λ_{SR} effect

Figure 5. The effects of (a) embedding coefficient γ, (b) fusing rate κ, (c) loss weights λ_{VR}, and (d) loss weights λ_{SR}. We take CUB as an example.

semantic-related visual representations for desirable visual-semantic interactions in ZSL. As such, ZSLViT achieves significant performances both in seen and unseen classes.

t-SNE Visualization of Visual Features. As shown in Fig. 4, we present the t-SNE visualization of visual features for (a) seen classes and (b) unseen classes on CUB, learned by the CNN backbone (*e.g.*, ResNet101 [20]), standard ViT [43] and our ZSLViT. Results show the visual features extracted from the CNN Backbone are confused between different classes, while the visual features learned by ViT are high-quality. This intuitively shows ViT is more suitable for ZSL task than CNN backbone. Furthermore, the visual features learned by our ZSLViT are desirable intra-class compactness and inter-class separability, because our ZSLViT discovers the semantic-related visual tokens and discards the semantic-unrelated visual tokens in the whole network according to the visual-semantic correspondences. As such, our ZSLViT significantly improves the ZSL performances over the CNN and ViT backbones.

Hyper-Parameter Analysis. We perform experiments to investigate the effects of various hyper-parameters on CUB, *i.e.*, embedding coefficient γ in Eq. 5, fusing rate κ in ViE, and loss weights λ_{VR} and λ_{SR} in Eq. 8. From the results in Fig. 5(a), ZSLViT achieves better performances with a relatively large value of γ (*i.e.*, 0.9). The reason is that the semantic information should be progressively embedded into visual features for stable optimization. Fig. 5(b) indicates that the performance of ZSLViT drops when the fusing rate κ in the visual enhancement module is set to small (*e.g.*, $\kappa < 0.6$), because some informative visual tokens are discarded. The results shown in Fig. 5(c) and Fig. 5(d) show the better effects of our ZSLViT can be achieved when λ_{VR} is larger than λ_{SR}. Because we mainly aim to enhance semantic information into visual representations for subsequent learning. Overall, ZSLViT is robust to all hyper-parameters and easy to optimize.

5. Conclusion

In this work, we devise a novel zero-shot learning framework, *i.e.*, progressive semantic-guided visual Transformer (ZSLViT), to learn semantic-related visual features for effective visual-semantic interactions. By conducting semantic-embedded token learning and visual enhancement at various stages, our ZSLViT effectively discovers the semantic-related visual features and discards semantic-unrelated visual features according to visual-semantic correspondences. We quantitatively and qualitatively demonstrate that ZSLViT achieves consistent improvements over the current state-of-the-art methods on three ZSL benchmarks, e.g., CUB, SUN, and AWA2.

References

[1] Zeynep Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:1425–1438, 2016. 2

[2] Faisal Alamri and Anjan Dutt. Implicit and explicit attention for zero-shot learning. In *GCPR*, 2021. 1, 3, 5, 6, 7

[3] Faisal Alamri and Anjan Dutta. Multi-head self-attention via vision transformer for zero-shot learning. In *IMVIP*, 2021. 1, 3, 5

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9630–9640, 2021. 3

[6] Shiming Chen, Shihuang Chen, Wen Qing Hou, Weiping Ding, and Xinge You. Egans: Evolutionary generative adversarial network search for zero-shot learning. *IEEE Transactions on Evolutionary Computation*, 2023. 2

[7] Shiming Chen, Ziming Hong, Yang Liu, Guo-Sen Xie, Baigui Sun, Hao Li, Qinmu Peng, Ke Lu, and Xinge You. Transzero: Attribute-guided transformer for zero-shot learning. In *AAAI*, 2022. 2, 5, 6

[8] Shiming Chen, Ziming Hong, Guo-Sen Xie, Wenhan Yang, Qinmu Peng, Kai Wang, Jian Zhao, and Xinge You. Msdn: Mutually semantic distillation network for zero-shot learning. In *CVPR*, 2022. 1, 2, 5, 6

[9] Shiming Chen, Zi-Quan Hong, Guosen Xie, Jian Zhao, Hao Li, Xinge You, Shuicheng Yan, and Ling Shao. Transzero++: Cross attribute-guided transformer for zero-shot learning. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 1, 2, 3

[10] Shiming Chen, Wen Qing Hou, Ziming Hong, Xiaohan Ding, Yibing Song, Xinge You, Tongliang Liu, and Kun Zhang. Evolving semantic prototype improves generative zero-shot learning. In *ICML*, 2023. 2

[11] Shiming Chen, Wenjie Wang, Beihao Xia, Qinmu Peng, Xinge You, Feng Zheng, and Ling Shao. Free: Feature refinement for generalized zero-shot learning. In *ICCV*, 2021. 1, 2, 5, 6

[12] Shiming Chen, Guo-Sen Xie, Yang Yang Liu, Qinmu Peng, Baigui Sun, Hao Li, Xinge You, and Ling Shao. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. In *NeurIPS*, 2021. 1, 2, 6

[13] Zhuo Chen, Yufen Huang, Jiaoyan Chen, Yuxia Geng, Wen Zhang, Yin Fang, Jeff Z. Pan, Wenting Song, and Huajun Chen. Duet: Cross-modal semantic grounding for contrastive zero-shot learning. In *AAAI*, 2023. 1, 2, 3, 6, 7

[14] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 3

[15] Anders Christensen, Massimiliano Mancini, A. Sophia Koepke, Ole Winther, and Zeynep Akata. Image-free classifier injection for zero-shot classification. In *ICCV*, 2023. 6

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 3

[17] Akshita Gupta, Sanath Narayan, Salman Hameed Khan, Fahad Shahbaz Khan, Ling Shao, and Joost van de Weijer. Generative multi-label zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:14611–14624, 2021. 2

[18] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. Contrastive embedding for generalized zero-shot learning. In *CVPR*, 2021. 2

[19] Jameel Hassan, Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *NeurIPS*, 2023. 2

[20] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 8

[21] Ziming Hong, Shiming Chen, Guosen Xie, Wenhan Yang, Jian Zhao, Yuanjie Shao, Qinmu Peng, and Xinge You. Semantic compression embedding for generative zero-shot learning. In *IJCAI*, 2022. 2

[22] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *ICCV*, pages 3463–3472, 2019. 3

[23] D. Huynh and E. Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *CVPR*, pages 4482–4492, 2020. 2, 6

[24] Dat T. Huynh and E. Elhamifar. Compositional zero-shot learning via fine-grained dense feature composition. In *NeurIPS*, 2020. 2, 6

[25] Junhan Kim, Kyuhong Shim, and Byonghyo Shim. Semantic feature extraction for generalized zero-shot learning. In *AAAI*, 2022. 1, 6

[26] Christoph H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1

[27] Christoph H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:453–465, 2014. 1

[28] H. Larochelle, D. Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, 2008. 1

[29] Xiaofan Li, Yachao Zhang, Shiran Bian, Yanyun Qu, Yuan Xie, Zhongchao Shi, and Jianping Fan. Vs-boost: Boosting visual-semantic association for generalized zero-shot learning. In *IJCAI*, 2023. 6

[30] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *ICLR*, 2022. 3

[31] Man Liu, Feng Li, Chunjie Zhang, Yunchao Wei, Huihui Bai, and Yao Zhao. Progressive semantic-visual mutual adaption for generalized zero-shot learning. In *CVPR*, 2023. 1, 2, 3

[32] Shichen Liu, Mingsheng Long, J. Wang, and Michael I. Jordan. Generalized zero-shot learning with deep calibration network. In *NeurIPS*, 2018. 2

[33] Yang Liu, Lei Zhou, Xiao Bai, Yifei Huang, Lin Gu, Jun Zhou, and T. Harada. Goal-oriented gaze estimation for zero-shot learning. In *CVPR*, 2021. 1, 3, 6

[34] Muhammad Ferjad Naeem, Muhammad Gul Zain Ali Khan, Yongqin Xian, Muhammad Zeshan Afzal, Didier Stricker, Luc Van Gool, and Federico Tombari. I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification. In *CVPR*, 2023. 6, 7

[35] Muhammad Ferjad Naeem, Yongqin Xian, Luc Van Gool, and Federico Tombari. I2dformer: Learning image to document attention for zero-shot image classification. In *NeurIPS*, 2022. 1, 6, 7

[36] Sanath Narayan, A. Gupta, F. Khan, Cees G. M. Snoek, and L. Shao. Latent embedding feedback and discriminative features for zero-shot classification. In *ECCV*, 2020. 2, 6

[37] Sanath Narayan, Akshita Gupta, Salman Hameed Khan, Fahad Shahbaz Khan, Ling Shao, and Mubarak Shah. Discriminative region-based multi-label zero-shot learning. In *ICCV*, pages 8711–8720, 2021. 1

[38] Mark Palatucci, D. Pomerleau, Geoffrey E. Hinton, and Tom Michael Mitchell. Zero-shot learning with semantic output codes. In *NeurIPS*, 2009. 1

[39] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, pages 2751–2758, 2012. 2, 5

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 6

[41] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *NeurIPS*, 2021. 3

[42] Olga Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Zhiheng Huang, A. Karpathy, A. Khosla, Michael S. Bernstein, A. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015. 1

[43] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herv'e J'egou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 5, 8

[44] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[45] M. R. Vyas, Hemanth Venkateswara, and S. Panchanathan. Leveraging seen and unseen semantic relationships for generative zero-shot learning. In *ECCV*, 2020. 6

[46] Ziyu Wan, Dongdong Chen, Yan Li, Xingguang Yan, Junge Zhang, Yizhou Yu, and Jing Liao. Transductive zero-shot learning with visual structure constraint. In *NeurIPS*, 2019. 1

[47] Z. Wang, Jian Liang, Ran He, Nana Xu, Zilei Wang, and Tien-Ping Tan. Improving zero-shot generalization for clip with synthesized prompts. In *ICCV*, 2023. 6

[48] P. Welinder, S. Branson, T. Mita, C. Wah, Florian Schroff, Serge J. Belongie, and P. Perona. Caltech-ucsd birds 200. *Technical Report CNS-TR-2010-001, Caltech,*, 2010. 2, 5

[49] Yongqin Xian, T. Lorenz, B. Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018. 1, 2, 5

[50] Yongqin Xian, B. Schiele, and Zeynep Akata. Zero-shot learning — the good, the bad and the ugly. *CVPR*, pages 3077–3086, 2017. 2, 3, 5

[51] Yongqin Xian, Saurabh Sharma, B. Schiele, and Zeynep Akata. F-vaegan-d2: A feature generating

framework for any-shot learning. In *CVPR*, pages 10267–10276, 2019. 1, 2, 6

[52] Guo-Sen Xie, L. Liu, Xiaobo Jin, F. Zhu, Zheng Zhang, J. Qin, Yazhou Yao, and L. Shao. Attentive region embedding network for zero-shot learning. In *CVPR*, pages 9376–9385, 2019. 1, 3, 6

[53] Wenjia Xu, Yongqin Xian, Jiuniu Wang, B. Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *NeurIPS*, 2020. 1, 3, 5, 6

[54] Zhongqi Yue, Tan Wang, Hanwang Zhang, Qianru Sun, and Xiansheng Hua. Counterfactual zero-shot and open-set visual recognition. In *CVPR*, 2021. 6

[55] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Ouyang Wanli, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *CVPR*, 2022. 3

[56] Shengxiang Zhang, Muzammal Naseer, Guangyi Chen, Zhiqiang Shen, Salman A. Khan, Kun Zhang, and Fahad Shahbaz Khan. Towards realistic zero-shot classification via self structural semantic alignment. In *AAAI*, 2023. 1

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

[58] Yizhe Zhu, Jianwen Xie, Z. Tang, Xi Peng, and A. Elgammal. Semantic-guided multi-attention localization for zero-shot learning. In *NeurIPS*, 2019. 1, 3, 6