TransZero++: Cross Attribute-Guided Transformer for Zero-Shot Learning

Shiming Chen, Ziming Hong, Guo-Sen Xie, Jian Zhao, Xinge You, *Senior Member, IEEE*, Shuicheng Yan, *Fellow, IEEE*, and Ling Shao, *Fellow, IEEE*

Abstract—Zero-shot learning (ZSL) tackles the novel class recognition problem by transferring semantic knowledge from seen classes to unseen ones. Semantic knowledge is typically represented by attribute descriptions shared between different classes, which act as strong priors for localizing object attributes that represent discriminative region features, enabling significant visual-semantic interaction of ZSL. Existing attention-based models have attempted to learn inferior region features in a single image using unidirectional attention, ignoring the transferability and discriminative attribute localization of visual features. In this paper, we propose a cross attribute-guided Transformer network, termed TransZero++, to refine visual features and learn accurate attribute localization for semantic-augmented visual embedding representations in ZSL. TransZero++ consists of an attribute→visual Transformer sub-net (AVT) and a visual→attribute Transformer sub-net (VAT). Specifically, AVT first takes a feature augmentation encoder to alleviate the cross-dataset bias between ImageNet and ZSL benchmarks, and improves the transferability of visual features by reducing the entangled relative geometry relationships among region features. Then, an attribute→visual decoder is employed to localize the image regions most relevant to each attribute in a given image for attribute-based visual feature representations. Analogously, VAT uses the similar feature augmentation encoder to refine the visual features, which are further applied in visual→attribute decoder to learn visual-based attribute features. By further introducing feature-level and prediction-level semantical collaborative losses, the two attribute-guided transformers teach each other to learn semantic-augmented visual embeddings via semantical collaborative learning. Finally, the semantic-augmented visual embeddings learned by AVT and VAT are fused to conduct desirable visual-semantic interaction cooperated with semantic vectors for ZSL classification. Extensive experiments show that TransZero++ achieves the new state-of-the-art performances on three popular challenging ZSL benchmarks. The codes are available at: https://github.com/shiming-chen/TransZero_pp.

Index Terms—Zero-Shot Learning; Transformer; Attribute Localization; Semantic-Augmented Visual Embedding; Semantical Collaborative Learning.

conference papers position the abstract like regular

1 Introduction and Motivation

HUMAN beings are capable to learn novel concepts based on prior experience without seeing them in advance. For example, given the clues that zebras appear like horses yet with black-and-white stripes of tigers, one can quickly recognize a zebra if he/she has seen horses and tigers before. Nevertheless, unlike humans, supervised machine learning can only classify samples belonging to the classes that have already appeared during the training phase, and they are not capable of handling samples from previously unseen categories. Motivated by this challenge, zero-shot learning (ZSL) was proposed to recognize new classes by exploiting the intrinsic semantic relatedness during learning [1], [2], [3], [4], [5], [6]. Since ZSL is a foundational method of artificial intelligence, it is commonly used in tasks with wide real-world applications, *e.g.*, image classification [7], [8], image retrieval [9], [10], semantic segmentation [11] and object detection [12]. Particularly, the core idea of ZSL is to learn discriminative visual features for conducting effective visual-semantic interactions based on the semantic information (*e.g.*, attribute vectors [4], sentence embeddings [13], and DNA [14]), which are shared between the seen and unseen classes employed to support the knowledge transfer. According to the different ranges of the label space during testing, ZSL methods can be categorized into conventional ZSL (CZSL), which aims to predict unseen classes, and generalized ZSL (GZSL), which can predict both seen and unseen classes [15]. Moreover, ZSL can also be classified as inductive ZSL [16], [17], which only utilizes the labeled seen data, and transductive ZSL [18], [19], assuming that unlabeled unseen data are available [15]. Inductive ZSL is more reasonable and challenging, we are thus interested in the inductive ZSL setting in this paper.

To enable visual-semantic interactions for transferring knowledge from seen to unseen classes, early embedding-based ZSL methods are trying to learn the embedding between seen classes and their class semantic vectors, and then classify unseen classes by nearest neighbor search in the embedding space. However, these embedding-based methods inevitably overfit to seen classes under the GZSL setting (known as the bias problem), since the embedding is only learned by seen class samples. To mitigate this bias problem, many generative ZSL methods have been proposed to synthesize feature samples for unseen classes by leveraging generative models (*e.g.*, variational autoencoders (VAEs) [21], [22], [23], generative adversarial nets (GANs) [8], [16], [24], and generative flows [25]) for data augmentation. Thus the generative ZSL methods can compensate for the lack of training samples of unseen classes and convert the ZSL task into a supervised classification task.

S. Chen, Z. Hong and X. You are with the School of Electronic Information and Communication, Huazhong University of Science and Technology, Wuhan 430074, China. (Corresponding author: Xinge You. e-mail: youxg@hust.edu.cn)
J. Zhao is with the Institute of North Electronic Equipment, Beijing, China.
G. Xie is with the Mohamed bin Zayed University of AI (MBZUAI), Abu Dhabi, UAE.
S. Yan is with Sea AI Lab (SAIL), Singapore.
L. Shao is with the National Center for Artificial Intelligence (NCAI), Saudi Data and Artificial Intelligence Authority (SDAIA), Riyadh, Saudi Arabia.

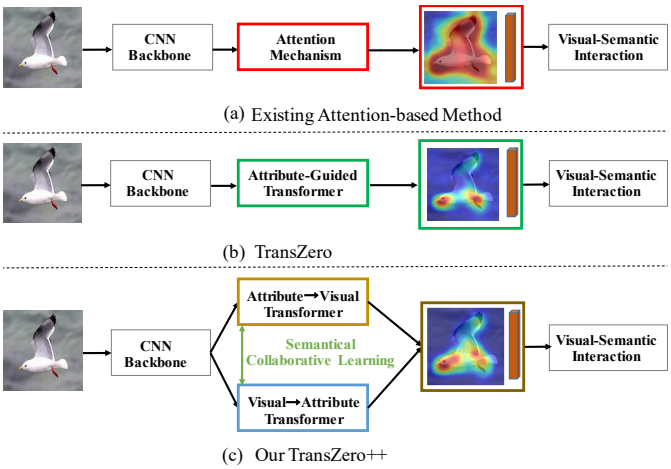


Fig. 1: Motivation illustration. (a) Existing attention-based ZSL methods simply learn inferior region embeddings (e.g., the whole bird body) using unidirectional attention, ignoring the transferability and discriminative attribute localization (e.g., the distinctive bird body parts) of visual features; (b) TransZero [20] (Conference version) employs an attribute-augmented Transformer reduces the entangled relationships among region features to improve their transferability, localizes the object attributes to represent discriminative region features, enabling significant visual-semantic interaction. (c) Our TransZero++ takes two cross attribute-augmented Transformer (*i.e.*, attribute→visual Transformer and visual→attribute Transformer) to furter discover more intrinsic semantic knowledge for semantic-augmented visual embedding representations via semantical collaborative learning, encouraging desirable visual-semantic interaction.

Although these methods have achieved significant improvement, they rely on global (holistic) visual features which are insufficient for representing the fine-grained information (*e.g.*, "head pattern plain" of *Acadian Flycatcher*) for distinguishing fine-grained classes [30], [31]. Because the discriminative information is contained in a few regions corresponding to a few attributes. Thus, the visual feature representations are inferior, resulting in undesirable visual-semantic interactions. More recently, few attention-based models [30], [32], [33], [34], [35], [36] have attempted to explore more discriminative region features under the guidance of the semantic information, as shown in Fig. 1 (a). However, these methods are limited in: i) They directly take the entangled region (grid) features for ZSL classification, which hinders the transferability of visual features from seen to unseen classes; ii) They simply learn region embeddings (*e.g.*, the whole bird body) using unidirectional attention, neglecting the importance of discriminative attribute localization (*e.g.*, the distinctive bird body parts) for semantic-augmented visual embedding representations. Furthermore, most of aforementioned methods directly use feature extraction models (*e.g.*, ResNet [26]) pre-trained on ImageNet [27] alone, ignoring the cross-dataset bias between ImageNet and GZSL benchmarks (*e.g.*, CUB [28]). Such a bias inevitably results in poor-quality visual features in which not all the dimensions are semantically related to the pre-defined attributes for ZSL tasks, which potentially limits the recognition performance on both seen and unseen classes [24], [29]. Thus, properly improving the transferability and localizing the object attributes of visual features for enabling significant visual-semantic interaction in ZSL has become very necessary.

To tackle the above challenges, in this paper, we propose a cross attribute-guided Transformer, termed TransZero++, which reduces the entangled relationships among region features to improve their transferability and localizes the object attributes to represent semantic-augmented region features in ZSL, as shown in Fig. 1 (c). To the best of our knowledge, TransZero++ is the first work extending the Transformer to the ZSL task. Specifically, TransZero++ consists of two attribute-guided Transformer sub-nets (*i.e.*, attribute→visual Transformer (AVT) and visual→attribute Transformer (VAT)) that learn attribute-based visual features and visual-based attribute features respectively, which are further mapped into the semantic embedding space using two mapping functions $\mathcal{M}_1$ and $\mathcal{M}_2$ to conduct desirable visual-semantic interaction. In AVT and VAT, we first take a feature augmentation encoder to i) alleviate the cross-dataset bias between ImageNet and ZSL benchmarks, and ii) reduce the entangled relative geometry relationships between different regions for improving the transferability from seen to unseen classes. They are ignored by existing ZSL methods. To learn locality-augmented visual features, we employ an attribute→visual decoder in AVT to localize the image regions most relevant to each attribute in a given image (denoted as attribute-based visual features), under the guidance of semantic attribute information. We also take a visual→attribute decoder to learn visual-based attribute features in VAT. By introducing feature-level and prediction-level semantical collaborative losses further, the two attribute-guided transformers teach each other to learn semantic-augmented visual embeddings via semantical collaborative learning. Finally, the two semantic-augmented visual embeddings cooperated with the semantic vectors to conduct desirable visual-semantic interaction for ZSL classification. Extensive experiments show that TransZero++ achieves the new state-of-the-art on three ZSL benchmarks. The qualitative results also demonstrate that TransZero++ refines visual features and accurately localizes attribute regions for semantic-augmented feature representations..

A preliminary version of this work was presented as a conference paper (termed TransZero [20]). As shown in Fig. 1 (b), although TransZero can localize some important attributes for discriminative region feature representations with low confident scores, some other valuable attributes are failed (*e.g.*, "white wing color" of *Red Legged Kittiwake*). In this version, we strengthen the work from four aspects: i) we propose VAT to learn visual-based semantic attribute representations that are complementary to the attribute-based visual features learned by AVT, improving the confidence scores for attribute localization. ii) we introduce feature-level and prediction-level semantic collaborative losses to encourage AVT and VAT to calibrate each other to discover more intrinsic semantic knowledge between visual and attribute features for semantic-augmented visual embedding representations, under the guidance of semantic collaborative learning. iii) since the learned attribute-based visual features and visual-based attribute features are complementary to each other, we combine the two semantic-augmented visual embeddings learned by AVT and VAT to conduct desirable visual-semantic interaction for ZSL classification. iv) We conduct substantially more experiments to demonstrate the effectiveness of the proposed framework and verify the contribution of each component. Thus, TransZero [20] is extended to be TransZero++.

The main contributions of this paper are summarized as follows:

- We introduce a novel ZSL method, termed TransZero++, which simultaneously refines the visual features, localizes the object attributes for discriminative region feature representations and learns semantic-augmented visual embeddings via semantical collaborative learning. TransZero++ consists of an attribute→visual Transformer sub-net (AVT) and a visual→attribute Transformer sub-net (VAT) that learns attribute-based visual features and visual-based attribute features, respectively, which are complementary to each other.
- We propose a feature augmentation encoder to i) alleviate the cross-dataset bias between ImageNet and ZSL benchmarks, and ii) reduce the entangled relative geometry relationships between different regions to improve the transferability of visual features. They are ignored by existing ZSL methods. This feature augmentation encoder is incorporated into AVT and VAT.
- We introduce feature-level and prediction-level semantic collaborative losses to enable semantical collaborative learning between the AVT and VAT, encouraging TransZero++ to learn semantic-augmented visual embeddings by discovering more intrinsic semantic knowledge between visual and attribute features.
- Extensive experiments demonstrate that TransZero++ achieves the new state-of-the-art on three popular challenging ZSL benchmarks. Compared with the latest attention-based method (*i.e.*, APN [34]), it leads to significant improvements of 6.3%/3.2%, 6.0%/4.9% and 4.2%/8.6% in $\boldsymbol{acc}/\boldsymbol{H}$ on CUB [28], SUN [37] and AWA2 [15], respectively.

The rest of this paper is organized as follows. Sec. 2 discusses related works. The proposed TransZero++ is illustrated in Sec. 3. Experimental results and discussions are provided in Sec. 4, respectively. Finally, we present a summary in Sec. 5.

2 Related Work

In this section, we mainly review three streams of related works: zero-shot learning, Transformer, and collaborative learning.

2.1 Zero-Shot Learning

Early embedding-based ZSL methods [8], [16], [18], [24], [38], [39], [40], [41], [41], [42] aim to learn a mapping from the visual domain to semantic domains to transfer semantic knowledge from seen to unseen classes. They usually extract global visual features from pre-trained or end-to-end trainable networks, *e.g.*, ResetNet [26]. Note that end-to-end models achieve better performance than pre-trained ones because they fine-tune the visual features, thus the cross-dataset bias between ImageNet and ZSL benchmarks is alleviated [8], [24].

However, these methods inevitably overfit to seen classes in GZSL since they only learn the model on seen classes [15], [43], [44]. As such, the generative ZSL methods are introduced to tackle this challenge using various generative models (*e.g.*, VAEs [21], [22], [23], [29], GAN [21], [22], [23], and generative flows [25]) to synthesize a number of images or visual features for unseen classes based on the class semantic vector (attribute values manually annotated by humans). Thus, the ZSL task is converted to supervised classification. Arora *et al.* [21] uses a conditional VAE model (SE-ZSL) to synthesize images for unseen classes. Since synthesizing the high dimensional image is not feasible, Xian *et al.* [8], [16] propose f-CLSWGAN and f-VAEGAN to synthesize visual features based on GANs. Different from these generative methods that learn semantic-to-visual mapping as a generator, the common space learning-based ZSL methods are also a special generative model that maps visual and semantic features into a common space simultaneously using VAEs [22], [29], [45].

Although the aforementioned ZSL methods have achieved significant improvements, they still yield relatively undesirable results. This is because they compress holistic visual features to perform global embedding cannot efficiently capture the subtle differences among various fine-grained classes [31]. Furthermore, the holistic visual features are limited to poor transferable from one domain to another domain (*e.g.*, from seen to unseen classes) [46], [47]. More relevant to this work are the recent attention-based ZSL methods [30], [32], [33], [34], [48] that utilize attribute descriptions as guidance to discover the more discriminative region (or part) features. Unfortunately, They simply learn region embeddings (*e.g.*, the whole bird body) neglecting the importance of discriminative attribute localization (*e.g.*, the distinctive bird body parts). Furthermore, the end-to-end attention models are also time-consuming when it comes to fine-tuning the CNN backbone. In contrast, we propose an attribute-guided Transformer to learn the attribute localization for discriminative region feature representations under non end-to-end ZSL model.

2.2 Transformer Model

Transformer models [49], [50], [51], [52] have recently achieved excellent performance on a wide range of language and computer vision tasks, *e.g.*, machine translation [53], image recognition [54], video understanding [55], visual question answering [56], etc. Generally, the success of Transformer can be attributed to its self-supervision and self-attention [51]. The self-supervision enables complex models to be trained without the high cost of manual annotation, which in turn allows generalizable representations that encode useful relationships between the entities presented in a given dataset to be learned. The self-attention layers consider the broad context of a given sequence by learning the relationships between the elements in the token set (*e.g.*, words in the language, or patches in an image). Some methods [55], [57], [58], [59] have also shown that the transformer can better capture the relationship between various modals (*e.g.*, visual features and language) in parallel during training. Motivated by these, we design an attribute-guided Transformer that reduces the relationships among different regions to improve the transferability of visual features and learns the attribute localization for representing discriminative region features. In contrast to most of the vision Transformers that learn feature representations on image patches, our TransZero++ learn semantic-augmented visual embeddings on visual features learned by CNN backbone (*e.g.*, ResNet).

2.3 Collaborative Learning

Recently, Cooperative Learning [60] has been introduced to learn multiple models jointly for the same task. Teacher-student models to create consistent training supervisions for labeled/unlabeled data using collaborative learning, enabling two-way knowledge transfer from each other. Thus the intrinsic knowledge between different models is distilled for feature representations [61], [62]. Some methods adopt a pool of student models instead of the teacher models by training them with supervision from each other [63],

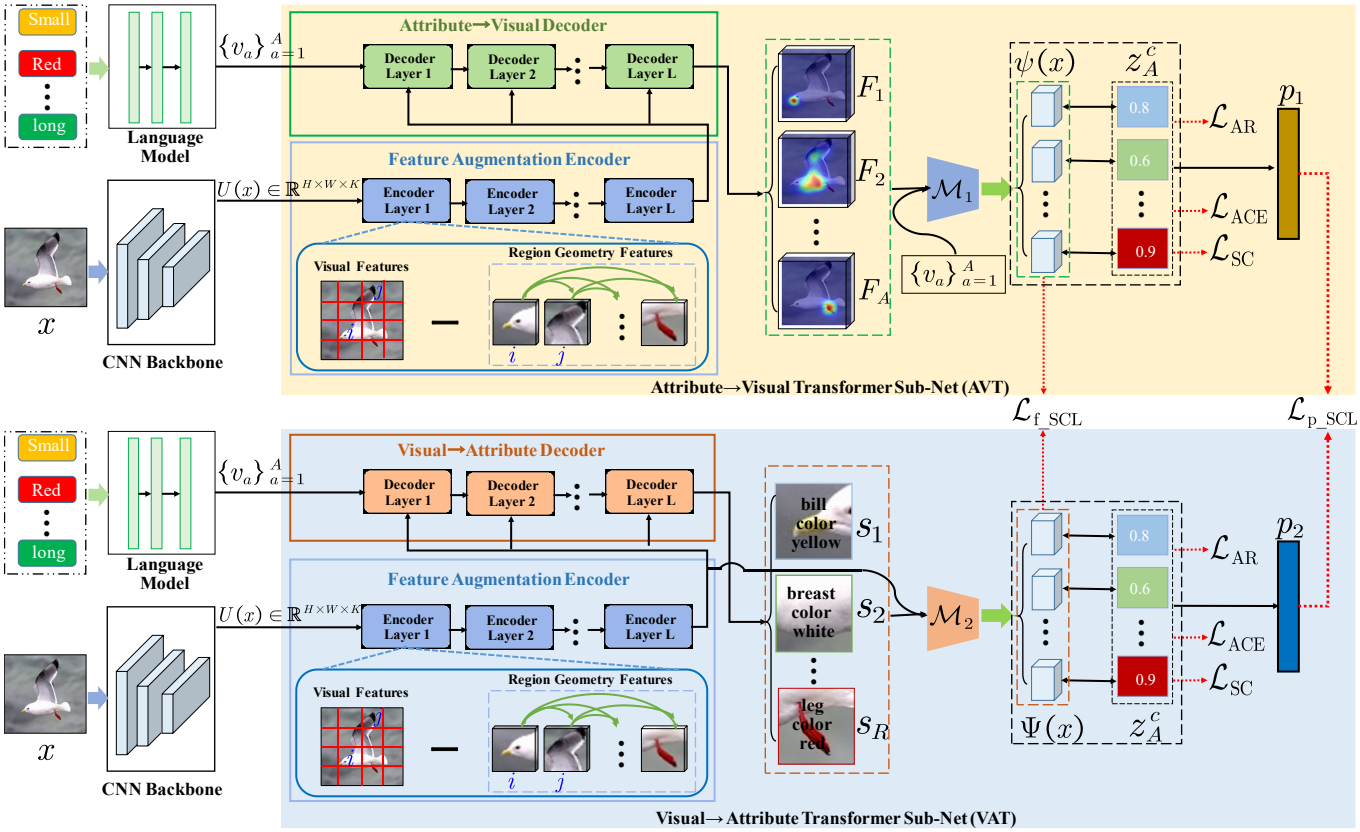


Fig. 2: The architecture of the proposed TransZero++ model. TransZero++ consists of an attribute→visual Transformer sub-net (AVT) and a visual→attribute Transformer sub-net (VAT). AVT includes a feature augmentation encoder that alleviates the cross-dataset bias between ImageNet and ZSL benchmarks and reduces the entangled geometry relationships between different regions for improving the transferability of visual features, and an attribute→visual decoder that localizes object attributes for attribute-based visual feature representations based on the semantic attribute information. Analogously, VAT learns visual-based attribute features using the similar feature augmentation encoder and a visual→attribute decoder. Finally, two mapping functions $\mathcal{M}_1$ and $\mathcal{M}_2$ map the learned attribute-based visual features and visual-based attribute features into semantic embedding space respectively under the guidance of semantical collaborative learning, enabling desirable visual-semantic interaction for ZSL classification.

[64]. These motivate us to design semantical collaborative learning to discover more intrinsic semantic knowledge (*e.g.*, attributes) for semantic-augmented visual embedding representations on the two attribute-guided Transformers. Different from existing collaborative methods that employ multiple similar networks for implicit knowledge distillation, our semantical collaborative learning is based on two attribute-guided Transformers that learn attribute-based visual features and visual-based attribute features respectively for explicit knowledge distillation.

3 Proposed Method

First, we introduce some notations and the problem definition of ZSL. We denote $\mathcal{D}^s = \{(x_i^s, y_i^s)\}$ as training data with C^s seen classes, where $x_i^s \in \mathcal{X}$ refers to the image i, and $y_i^s \in \mathcal{Y}^s$ is its corresponding class label. The unseen classes C^u have unlabeled samples $\mathcal{D}^u = \{(x_i^u, y_i^u)\}$, where $x_i^u \in \mathcal{X}$ are the unseen class images, and $y_i^u \in \mathcal{Y}^u$ are the corresponding labels. A set of class semantic vectors of the class $c \in \mathcal{C}^s \cup \mathcal{C}^u = \mathcal{C}$ with A attributes $z^c = [z_1^c, \dots, z_A^c]^\top = \phi(y)$ (the attribute values annotated by humans) helps knowledge transfer from seen to unseen classes. According to each word in attribute names, we also take a language model (*i.e.*, GloVe [65]) to learn the semantic attribute features $v_A = \{v_1, \cdots, v_a\}_{a=1}^A$ of all attributes as auxiliary information. ZSL aims to predict the class labels $y^u \in \mathcal{Y}^u$ in the CZSL settings and $y \in \mathcal{Y} = \mathcal{Y}^s \cup \mathcal{Y}^u$ in the GZSL setting, where $\mathcal{Y}^s \cap \mathcal{Y}^u = \emptyset$.

In this paper, we propose a cross attribute-guided Transformer network (termed TransZero++) to refine the visual features, localize the object attributes for discriminative region feature representations, and learn semantic-augmented visual embeddings via semantical collaborative learning under a non end-to-end model. This facilitates desirable visual-semantic interaction in ZSL. As illustrated in Fig. 2, our TransZero++ includes an attribute→visual Transformer sub-net (AVT) and visual→attribute Transformer sub-net (VAT). AVT refines the visual feature using a feature augmentation encoder, and employs an attribute→visual decoder to learns attribute-based visual features, which is further mapped as the semantic-augmented visual embedding $\psi(x)$ in the semantic embedding space using a mapping function $\mathcal{M}_1$. Analogously, VAT uses a similar feature augmentation encoder and a visual→attribute decoder to learn visual-based attribute features, which are further mapped as the semantic-augmented visual embeddings $\Psi(x)$ in the semantic embedding space using another mapping function $\mathcal{M}_2$. Finally, the two semantic-augmented visual embeddings are combined to conduct desirable visual-semantic interaction for ZSL classification based on the class semantic vectors. To encourage TransZero++ to learn semantic-augmented

visual embeddings, we introduce feature-level and prediction-level semantical collaborative losses to encourage the two cross AVT and VAT to learn collaboratively and teach each other throughout the training process.

3.1 Attribute→Visual Transformer

3.1.1 Feature Augmentation Encoder.

Since the cross-dataset bias between ImageNet and ZSL benchmarks potentially limits the quality of visual feature extraction [24], [29], we first propose a feature augmentation encoder to refine the visual features of ZSL benchmarks. In addition, previous ZSL methods simply flatten the grid features $U(x) \in \mathbb{R}^{H\times W\times K}$ (extracted by a CNN backbone) of a single image into a feature vector, which is further applied to generative models or embedding learning. Unfortunately, such a feature vector implicitly entangles the visual feature representations among various regions in an image, which hinders their transferability from one domain to other domains (e.g., from seen to unseen classes) [34], [56]. Atzmon *et al.* [46] and Chen [47] show that the local visual features are more transferable than the holistic ones. As such, we propose a feature-augmented scaled dot-product attention to further enhance the encoder layer by reducing the relative geometry relationships among the grid features.

Following [66], we first calculate the relative center coordinates $(v_i^{\text{cen}}, t_i^{\text{cen}})$ based on the pair of 2-D relative positions of the i-th grid $\left\{\left(v_i^{\min}, t_i^{\min}\right), \left(v_i^{\max}, t_i^{\max}\right)\right\}$ during learning the relative geometry features, formulated as:

$$(v_i^{\text{cen}}, t_i^{\text{cen}}) = \left(\frac{v_i^{\min} + v_i^{\max}}{2}, \frac{t_i^{\min} + t_i^{\max}}{2}\right), \tag{1}$$

$$w_i = \left(v_i^{\max} - v_i^{\min}\right) + 1, \tag{2}$$

$$h_i = \left(t_i^{\max} - t_i^{\min}\right) + 1, \tag{3}$$

where $(v_i^{\min}, t_i^{\min})$ and $(v_i^{\max}, t_i^{\max})$ are the relative position coordinates of the top left corner and bottom right corner of the grid i, respectively. Different from [66] that uses 4-D feature vectors for geometry feature representations, we only need to calculate the 2-D geometry feature representations as our grid features are irrelevant to the edges (*i.e.*, width and lendth) of grid.

Then, we construct region geometry features G_{ij} between grid i and grid j:

$$G_{ij} = \text{ReLU}\left(w_g^T g_{ij}\right), \tag{4}$$

where

$$g_{ij} = FC\left(r_{ij}\right), \qquad r_{ij} = \begin{pmatrix} \log\left(\frac{\left|v_i^{\text{cen}} - v_j^{\text{cen}}\right|}{w_i}\right) \\ \log\left(\frac{\left|t_i^{\text{cen}} - t_j^{\text{cen}}\right|}{h_i}\right) \end{pmatrix}, \tag{5}$$

where r_{ij} is the relative geometry relationship between grids i and j, FC is a fully connected layer followed by a $ReLU$ activation, and w_g^T is a set of learnable weight parameters.

Finally, we substract the region geometry features from the visual features in the feature-augmented scaled dot-product attention to provide a more accurate attention map, formally defined as:

$$Q^e = U(x)W_q^e, K^e = U(x)W_k^e, V^e = U(x)W_v^e, \tag{6}$$

$$Z_{aug} = \text{softmax}\left(\frac{Q^e {K^e}^\top}{\sqrt{d^e}} - G\right)V^e, \tag{7}$$

$$U_{aug}(x) \leftarrow U(x) + Z_{aug}, \tag{8}$$

where Q, K, V are the query, key, and value matrices, W_q^e, W_k^e, W_v^e are the learnable matrices of weights, d^e is a scaling factor, and Z_{aug} is the augmented features. $U(x) \in \mathbb{R}^{HW\times C}$ are the packed visual features, which are learned from the flattened features embedded by a fully connected layer followed by a ReLU and a Dropout layer. $U_{aug}(x)$ is the augmented visual features from the feature augmentation encoder.

3.1.2 Attribute→Visual Decoder.

To learn attribute-based visual features, we design attribute→visual decoder to localize the most relevant image regions to the corresponding attributes to extract attribute-based visual features from a given image for each attribute. We can attend to image regions with respect to each attribute, and compare each attribute to the corresponding attended visual region features to determine the importance of each attribute. Specifically, following the standard Transformer [49], our attribute→visual decoder employs a multi-head self-attention layer and feed-forward network (FFN) to build the decoder layer. In the decoding process, the attribute→visual decoder continuously localizes the local visual information under the guidance of semantic attribute features v_A. Thus, our attribute→visual decoder can effectively localize the image regions most relevant to each attribute in a given image. The multi-head self-attention layer uses the outputs of the encoder U_{aug} as keys $(K_t^{a\to v})$ and values $(V_t^{a\to v})$ and a set of learnable semantic embeddings v_A as queries $(Q_t^{a\to v})$. It is defined as:

$$Q_t^{a\to v} = v_A W_{qt}^{a\to v}, \tag{9}$$

$$K_t^{a\to v} = U_{aug}(x)W_{kt}^{a\to v}, \tag{10}$$

$$V_t^{a\to v} = U_{aug}(x)W_{vt}^{a\to v}, \tag{11}$$

$$\text{head}_t = \text{softmax}\left(\frac{Q_t^d {K_t^{a\to v}}^\top}{\sqrt{\tau}}\right)V_t^{a\to v}, \tag{12}$$

$$\hat{F} = \|_{t=1}^{T}(\text{head}_t)W_o^{a\to v}, \tag{13}$$

where $W_{qt}^{a\to v}$, $W_{kt}^{a\to v}$, $W_{vt}^{a\to v}$, and $W_o^{a\to v}$ are the learnable weights, τ is a scaling factor, and $\|$ is a concatenation operation. Thus, we get a set of attribute-based visual features $\hat{F} = \{\hat{F}_1, \cdots, \hat{F}_A\}$, which captures the visual evidence used to localize the corresponding semantic attributes in the image. Specifically, our AVT will assign a high positive score to the a-th attribute if an image has an obvious attribute v_a. Otherwise, AVT will assign a negative score to the a-th attribute. Then, an FFN with two linear transformations followed a ReLU activation in between is applied to the attended features $\hat{F}$:

$$F = ReLu\left(\hat{F}W_1^{a\to v} + b_1^{a\to v}\right)W_2^{a\to v} + b_2^{a\to v}, \tag{14}$$

where $W_1^{a\to v}$, $W_2^{a\to v}$, $b_1^{a\to v}$ and $b_2^{a\to v}$ are the weights and biases of the linear layers respectively, and $F = \{F_1, \cdots, F_A\}$ are the final attribute-based visual features that will be fed into VSEN for desirable visual-semantic interaction.

3.1.3 Visual-Semantic Embedding Mapping

After learning attribute-based visual features that are locality-augmented, we further map them into the semantic embedding space. Based on a mapping function $(\mathcal{M}_1)$, we take the semantic attribute vectors $v_A = \{v_1, \cdots, v_a\}_{a=1}^{A}$ as support to encourage the mapping to be more accurate. Specifically, $\mathcal{M}_1$ matches the attribute-based visual features F with the semantic attribute information v_a, formulated as:

$$\psi(x_i) = \mathcal{M}_1(F) = v_A^\top W_3^{a\to v} F, \tag{15}$$

where $W_3^{a\to v}$ is an embedding matrix that embeds F into the semantic attribute space. Similar to the class semantic vector z^c, $\psi_a(x_i)$ is an attribute score that represents the confidence of having the a-th attribute in the image x_i. Given a set of semantic attribute vectors $v_A = \{v_1, \cdots, v_a\}_{a=1}^A$, TransZero++ obtains a mapped semantic embedding $\psi(x_i)$ of a single image x_i.

3.2 Visual→Attribute Transformer

Likewise, we introduce visual→attribute Transformer (VAT) to attend to attributes with respect to each image region, and thus the visual-based attribute features are learned. They are complementary to the attribute-based visual features, enabling them to calibrate each other to discover more intrinsic semantic knowledge between visual and attribute features. VAT first applies the similar feature augmentation encoder to refine the visual features as $U_{aug}(x)$, which are further used in visual→attribute decoder of VAT.

3.2.1 *Visual→Attribute Decoder.*

After improving the visual features, we design a visual→attribute decoder to learns visual-based attribute features. Formally, it is formulated as:

$$Q_t^{v\to a} = U_{aug}(x)W_{qt}^{v\to a}, \tag{16}$$

$$K_t^{v\to a} = v_A W_{kt}^{v\to a}, \tag{17}$$

$$V_t^{v\to a} = v_A W_{vt}^{v\to a}, \tag{18}$$

$$\text{head}_t = \text{softmax}\left(\frac{Q_t^d {K_t^{v\to a}}^\top}{\sqrt{\tau}}\right) V_t^{v\to a}, \tag{19}$$

$$\hat{S} = \|_{t=1}^T (\text{head}_t) W_o^{v\to a}, \tag{20}$$

where $W_{qt}^{v\to a}$, $W_{kt}^{v\to a}$, $W_{vt}^{v\to a}$, and $W_o^{v\to a}$ are the learnable weights, and $\|$ is a concatenation operation. As such, we get a set of visual-based attribute features $\hat{S} = \{\hat{S}_1, \cdots, \hat{S}_K\}$. Intrinsically, $\hat{S}$ is the visual semantic representations corresponding to the K visual regions in a single image. Specifically, our VAT will assign a high positive score to the k-th visual region with respect to the corresponding attribute, otherwise, VAT will assign a negative score. Then, an FFN with two linear transformations followed by a ReLU activation in between is applied to the attended features $\hat{S}$:

$$S = ReLu\left(\hat{S}W_1^{v\to a} + b_1^{v\to a}\right) W_2^{v\to a} + b_2^{v\to a}, \tag{21}$$

where $W_1^{v\to a}$, $W_2^{v\to a}$, $b_1^{v\to a}$ and $b_2^{v\to a}$ are the weights and biases of the linear layers respectively, and $S = \{S_1, \cdots, S_K\}$ are the final visual-based attribute features, which will be fed into VSEN for significant visual-semantic interaction.

3.2.2 *Visual-Semantic Embedding Mapping*

Once visual-based attribute features are learned, we map them into the semantic embedding space based on a mapping function ($\mathcal{M}_2$). To conduct an effective map, we take the augmented visual features $U_{aug}(x)$ learned by feature augmentation encoder as support. Thus, $\mathcal{M}_2$ first maps the visual-based attribute features S into K region scores $\bar{S}$, formulated as:

$$\bar{S} = \mathcal{M}_2(S) = U_{aug}(x)^\top W_3^{v\to a} S, \tag{22}$$

where $W_3^{v\to a}$ is a learnable mapping matrix. Here, $\bar{S}$ is K-D, which is not match with the dimension of class semantic vector A-D. Thus, $\mathcal{M}_2$ further embeds $\bar{S}$ into the semantic attribute space with dimension of A based on an attention score $Att = v_A^\top W_{att} U(x)$, written as:

$$\Psi(x_i) = \mathcal{M}_2(S) = \bar{S} \times Att, \tag{23}$$

Similar to the $\psi(x_i)$, $\Psi_a(x_i)$ is an attribute score that represents the confidence of having the a-th attribute in the image x_i. As such, TransZero++ obtains a mapped semantic embedding $\Psi(x_i)$ of a single image x_i in VAT.

3.3 Model Optimization

To achieve effective optimization for TransZero++, each attribute-guided Transformer sub-net is trained with three supervised losses that have been used in our conference version [20], *i.e.*, the attribute regression loss, attribute-based cross-entropy loss, and self-calibration loss. To enable semantic collaborative learning between the two attribute-guided Transformer sub-nets, *i.e.*, AVT and VAT, we introduce a feature-level semantic collaborative loss and prediction-level semantic collaborative loss, which align each other's visual embeddings and class posterior probabilities respectively.

Attribute Regression Loss. To encourage $\mathcal{M}_1$ and $\mathcal{M}_2$ to accurately map visual/attribute features into their corresponding semantic embeddings, we introduce an attribute regression loss to constrain TransZero++. Here, we regard visual-semantic mapping as a regression problem and minimize the mean square error between the embedded attribute score $f(x_i)$ and the corresponding ground truth attribute score z^c of a batch of n_b images $\{x_i^s\}_{i=1}^{n_b}$:

$$\mathcal{L}_{\text{AR}} = \frac{1}{n_b}\sum_{i=1}^{n_b} \|f(x_i^s) - z^c\|_2^2. \tag{24}$$

where $f(x_i^s) = \psi(x_i^s)$ for AVT and $f(x_i^s) = \Psi(x_i^s)$ for VAT.

Attribute-Based Cross-Entropy Loss. Since the associated visual/attribute embedding is projected near its class semantic vector z^c when an attribute is visually present in an image, we take the attribute-based cross-entropy loss $\mathcal{L}_{\text{ACE}}$ to optimize the parameters of the TransZero++, i.e., the dot product between the visual/attribute embedding and each class semantic vector is calculated to produce class logits. This encourages the image/attribute to have the highest compatibility score with its corresponding class semantic vector. Given a batch of n_b training images $\{x_i^s\}_{i=1}^{n_b}$ with their corresponding class semantic vectors z^c, $\mathcal{L}_{\text{ACE}}$ is defined as:

$$\mathcal{L}_{\text{ACE}} = -\frac{1}{n_b}\sum_{i=1}^{n_b} \log \frac{\exp\left(f(x_i^s) \times z^c\right)}{\sum_{\hat{c}\in\mathcal{C}} \exp\left(f(x_i^s) \times z^{\hat{c}}\right)}. \tag{25}$$

Self-Calibration Loss. Since $\mathcal{L}_{\text{AR}}$ and $\mathcal{L}_{\text{ACE}}$ optimize the model on only seen classes, TransZero++ inevitably overfits to these classes [31], [33], [34]. To tackle this challenge, we further employ a self-calibration loss $\mathcal{L}_{\text{SC}}$ to explicitly shift some of the prediction probabilities from seen to unseen classes. $\mathcal{L}_{\text{SC}}$ is thus formulated as:

$$\mathcal{L}_{\text{SC}} = -\frac{1}{n_b}\sum_{i=1}^{n_b}\sum_{c'=1}^{\mathcal{C}^u} \log \frac{\exp\left(f(x_i^s) \times z^{c'} + \mathbb{I}_{[c'\in\mathcal{C}^u]}\right)}{\sum_{\hat{c}\in\mathcal{C}} \exp\left(f(x_i^s) \times z^{\hat{c}} + \mathbb{I}_{[\hat{c}\in\mathcal{C}^u]}\right)}, \tag{26}$$

where $\mathbb{I}_{[c\in\mathcal{C}^u]}$ is an indicator function (*i.e.*, it is 1 when $c \in \mathcal{C}^u$, otherwise -1). Intuitively, $\mathcal{L}_{\text{ACE}}$ encourages non-zero probabilities to be assigned to the unseen classes during training, which allows TransZero++ to produce a large/non-zero probability for the true unseen class when given test samples from unseen classes.

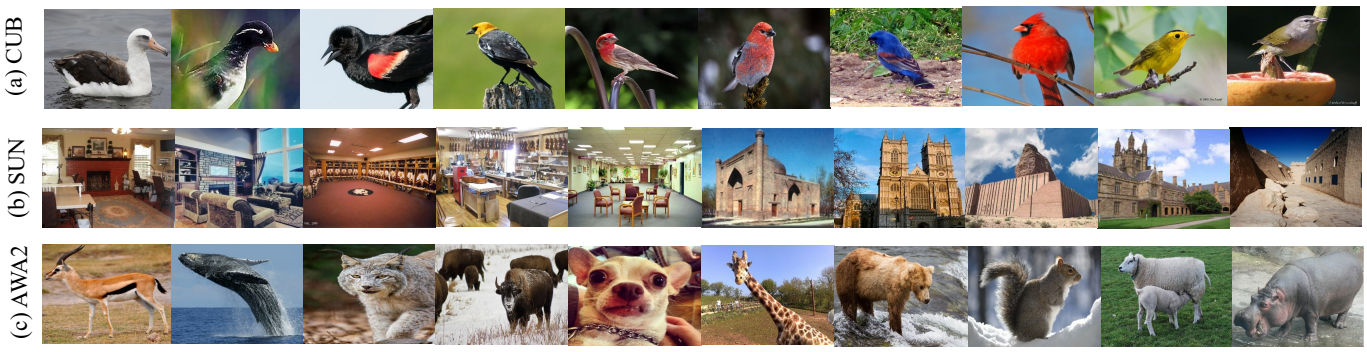


Fig. 3: Some samples on various datasets, including two fine-grained datasets (*i.e.*, CUB and SUN), and one coarse-grained dataset (*i.e.*, AWA2). Each sample is extract from various classes. (Best viewed in color.)

Algorithm 1 The algorithm of TransZero++.

Input: The training set $\mathcal{D}^s = \{(x_i^s, y_i^s)\}_{i=1}^{N^s}$, the test set $\mathcal{D}^u = \{(x_i^u, y_i^u)\}_{i=1}^{N^u}$, the pretrained CNN backbone ResNet101, the maximum iteration epoch $\max_{iter}$, loss weights (*i.e.*, $\mathcal{L}_{\text{AR}}$, $\mathcal{L}_{\text{SC}}$, λ_{VAT}, $\lambda_{\text{f\_SCL}}$ and $\lambda_{\text{p\_SCL}}$), the combination coefficients α, and hyperparameters (momentum = 0.9, weight_decay = 0.0001) of the Adam optimizer.

Output: The predicted label c^* for the test samples.

while $iter \leq \max_{iter}$ **do** ▷ *Optimization*
  Take CNN backbone (*e.g.*, ResNet101 [26]) to extract the visual features $U(x)$ for all image samples.
  Take language model (*i.e.*, GloVe [65]) to learn the semantic attribute features $v_A = \{v_1, \cdots, v_a\}_{a=1}^{A}$ for each attribute.
  Optimize TransZero++ with Eq. 31.
end while
Predict the label c^* of the test samples using Eq. 32. ▷ *Prediction*

Semantical Collaborative Loss. To enable the two attribute-augmented Transformer sub-nets to learn collaboratively and teach each other throughout the training process, we further introduce a feature-level semantical collaborative loss $\mathcal{L}_{\text{f\_SCL}}$ and a prediction-level semantical collaborative loss $\mathcal{L}_{\text{p\_SCL}}$ for optimization. These two losses are based on ℓ_2 distance. Note that the ℓ_2 distance can be replaced with other metrics, *e.g.*, the Kullback Leibler (KL) Divergence or Jensen-Shannon Divergence (JSD).

Specifically, $\mathcal{L}_{\text{f\_SCL}}$ uses an ℓ_2 distance between the visual semantic embeddings of AVT and VAT (*i.e.*, $\psi(x_i)$ and $\Psi(x_i)$) for test sample x_i, formulated as:

$$\mathcal{L}_{\text{f\_SCL}} = -\frac{1}{n_b}\sum_{i=1}^{n_b} \|\psi(x_i) - \Psi(x_i)\|_2^2. \tag{27}$$

Likewise, $\mathcal{L}_{\text{p\_SCL}}$ calculates the ℓ_2 distance between the predictions of the two attribute-augmented Transformer sub-nets (*i.e.*, p_1 and p_2), formulated as:

$$\mathcal{L}_{\text{p\_SCL}} = -\frac{1}{n_b}\sum_{i=1}^{n_b} \|p_1(x_i) - p_2(x_i)\|_2^2, \tag{28}$$

Similar to the TransZero, the AVT and VAT are optimized with the three supervised losses, *i.e.*, $\mathcal{L}_{\text{ACE}}$, $\mathcal{L}_{\text{AR}}$ and $\mathcal{L}_{\text{SC}}$, formulated as:

$$\mathcal{L}_{\text{AVT}} = \mathcal{L}_{\text{ACE}}^{\text{AVT}} + \lambda_{\text{AR}}\mathcal{L}_{\text{AR}}^{\text{AVT}} + \lambda_{\text{SC}}\mathcal{L}_{\text{SC}}^{\text{AVT}}, \tag{29}$$

$$\mathcal{L}_{\text{VAT}} = \mathcal{L}_{\text{ACE}}^{\text{VAT}} + \lambda_{\text{AR}}\mathcal{L}_{\text{AR}}^{\text{VAT}} + \lambda_{\text{SC}}\mathcal{L}_{\text{SC}}^{\text{VAT}}, \tag{30}$$

where λ_{AR} and λ_{SC} are the loss weights to control the loss $\mathcal{L}_{\text{AR}}$ and $\mathcal{L}_{\text{SC}}$, respectively, in the AVT and VAT. Finally, we formulate the overall loss function of TransZero++:

$$\begin{aligned}\mathcal{L}_{total} = \mathcal{L}_{\text{AVT}} + \lambda_{\text{VAT}}\mathcal{L}_{\text{VAT}} \\ + \lambda_{\text{f\_SCL}}\mathcal{L}_{\text{f\_SCL}} + \lambda_{\text{p\_SCL}}\mathcal{L}_{\text{p\_SCL}},\end{aligned} \tag{31}$$

where λ_{VAT}, $\lambda_{\text{f\_SCL}}$ and $\lambda_{\text{p\_SCL}}$ are the weights to control their corresponding loss terms. To enable the training process of TransZero++ more stable, we set the loss weight to one for $\mathcal{L}_{\text{AVT}}$.

3.4 Zero-Shot Prediction

After training TransZero++, We first obtain the visual embeddings of a test instance x_i in the semantic space w.r.t. AVT and VAT, *i.e.*, $\psi(x)$ and $\Psi(x)$. Considering the semantic-augmented visual embeddings learned by AVT and VAT are complementary to each other, we fuse their predictions using a combination coefficients α to predict the test label of x_i with an explicit calibration, formulated as:

$$c^* = \arg\max_{c \in \mathcal{C}^u/\mathcal{C}} (\alpha\psi(x_i) + (1-\alpha)\Psi(x_i))^\top \times z^c + \mathbb{I}_{[c\in\mathcal{C}^u]}. \tag{32}$$

Here, $\mathcal{C}^u/\mathcal{C}$ corresponds to the CZSL/GZSL setting, respectively. The complete procedures (including model training and prediction) for TransZero++ are illustrated by the pseudocode in Algorithm 1.

4 Experiments

Dataset. We conduct extensive experiments on three popular challenging ZSL benchmarks, including two fine-grained datasets (*i.e.*, CUB [28] and SUN [37]) and a coarse-grained dataset (*i.e.*, AWA2 [15]). CUB has 11,788 images of 200 bird classes (seen/unseen classes = 150/50) depicted with 312 attributes. SUN includes 14,340 images from 717 scene classes (seen/unseen classes = 645/72) depicted with 102 attributes. AWA2 consists of 37,322 images from 50 animal classes (seen/unseen classes = 40/10) depicted with 85 attributes. Here, we show some samples on these datasets, as shown in Fig. 3.

TABLE 1: Results (%) of the state-of-the-art CZSL and GZSL modes on CUB, SUN and AWA2, including end-to-end and non end-to-end methods (generative and non-generative methods). The best and second-best results are marked in **Red** and **Blue**, respectively. The Symbol "–" indicates no results. The Symbol "*" denotes attention-based methods.

Methods	CUB				SUN				AWA2			
	CZSL	GZSL			CZSL	GZSL			CZSL	GZSL		
	acc	U	S	H	acc	U	S	H	acc	U	S	H
End-to-End												
QFSL [18]	58.8	33.3	48.1	39.4	56.2	30.9	18.5	23.1	63.5	52.1	72.8	60.7
LDF [38]	67.5	26.4	**81.6**	39.9	–	–	–	–	65.5	9.8	87.4	17.6
SGMA* [33]	71.0	36.7	71.3	48.5	–	–	–	–	68.8	37.6	87.1	52.5
AREN* [32]	71.8	38.9	78.7	52.1	60.6	19.0	38.8	25.5	67.9	15.6	**92.9**	26.7
LFGAA* [36]	67.6	36.2	**80.9**	50.0	61.5	18.5	**40.0**	25.3	68.1	27.0	**93.4**	41.9
APN* [34]	72.0	65.3	69.3	67.2	61.6	41.9	34.0	37.6	68.4	57.1	72.4	63.9
Non End-to-End												
Generative Methods												
SE-ZSL [21]	59.6	41.5	53.3	46.7	63.4	40.9	30.5	34.9	69.2	58.3	68.1	62.8
f-CLSWGAN [16]	57.3	43.7	57.7	49.7	60.8	42.6	36.6	39.4	68.2	57.9	61.4	59.6
f-VAEGAN [8]	61.0	48.4	60.1	53.6	64.7	45.1	38.0	41.3	71.1	57.6	70.6	63.5
E-PGN [39]	72.4	52.0	61.1	56.2	–	–	–	–	**73.4**	52.6	83.5	64.6
Composer [67]	69.4	56.4	63.8	59.9	62.6	**55.1**	22.0	31.4	71.5	62.1	77.3	68.8
GCM-CF [68]	–	61.0	59.7	60.3	–	47.9	37.8	42.2	–	60.4	75.1	67.0
SDGZSL [29]	75.5	59.9	66.4	63.0	–	–	–	–	72.1	64.6	73.6	68.8
FREE [24]	–	55.7	59.9	57.7	–	47.4	37.2	41.7	–	60.4	75.4	67.1
HSVA [23]	62.8	52.7	58.3	55.3	63.8	48.6	**39.0**	**43.3**	–	59.3	76.6	66.8
Non-Generative Methods												
SP-AEN [69]	55.4	34.7	70.6	46.6	59.2	24.9	38.6	30.3	58.5	23.3	90.9	37.1
PQZSL [70]	–	43.2	51.4	46.9	–	35.1	35.3	35.2	–	31.7	70.9	43.8
IIR [71]	63.8	30.4	65.8	41.2	63.5	22.0	34.1	26.7	67.9	17.6	87.0	28.9
TCN [72]	59.5	52.6	52.0	52.3	61.5	31.2	37.3	34.0	71.2	61.2	65.8	63.4
DVBE [40]	–	53.2	60.2	56.5	–	45.0	37.2	40.7	–	**63.6**	70.8	67.0
DAZLE* [31]	66.0	56.7	59.6	58.1	59.4	52.3	24.3	33.2	67.9	60.3	75.7	67.1
TransZero [20](Conference Version)	**76.8**	**69.3**	68.3	**68.8**	**65.6**	**52.6**	33.4	40.8	70.1	61.3	82.3	**70.2**
TransZero++ (Ours)	**78.3**	**67.5**	73.6	**70.4**	**67.6**	48.6	37.8	**42.5**	**72.6**	**64.6**	82.7	**72.5**

TABLE 2: Ablation studies for different components of TransZero++ on the CUB and SUN datasets. "FAE" is the feature augmentation encoder, "FA" means feature augmentation, and "DEC" denotes the decoders in AVT and VAT.

Method	CUB				SUN			
	acc	U	S	H	acc	U	S	H
TransZero++ w/o AVT	49.0	36.4	48.5	41.6	57.7	36.9	28.7	32.3
TransZero++ w/o VAT	75.6	62.8	72.3	67.2	63.1	44.1	35.5	39.3
TransZero++ w/o FAE	70.7	63.4	57.5	60.3	64.0	52.0	33.5	40.8
TransZero++ w/o FA	76.4	66.6	70.3	68.4	65.3	46.7	36.9	41.2
TransZero++ w/o DEC	62.7	49.2	63.4	55.4	64.1	48.4	34.0	39.9
TransZero++ (full)	**78.3**	67.5	73.6	**70.4**	**67.6**	48.6	37.8	**42.5**

Evaluation Protocols. Following [15], we measure the top-1 accuracy both in the CZSL and GZSL settings. In the CZSL setting, we predict the unseen classes to compute the accuracy of test samples, i.e., $\boldsymbol{acc}$. In the GZSL setting, we calculate the accuracy of the test samples from both the seen classes (denoted as $\boldsymbol{S}$) and unseen classes (denoted as $\boldsymbol{U}$). Meanwhile, their harmonic mean (defined as $\boldsymbol{H} = (\mathbf{2} \times \boldsymbol{S} \times \boldsymbol{U})/(\boldsymbol{S} + \boldsymbol{U})$) is also used for evaluation in the GZSL setting.

Implementation Details. We use the training splits proposed by [16]. We take a ResNet101 pre-trained on ImageNet as the CNN backbone to extract the visual feature map $U(x) \in \mathbb{R}^{H \times W \times K}$ (H and W are the height and width of the feature maps, K is the number of channels) without fine-tuning. We use the Adam optimizer with hyperparameters (momentum = 0.9, weight decay = 0.0001) to optimize our model. The learning rate and batch size are set to 0.0001 and 50 for all datasets, respectively. Since the training data for the ZSL model is a medium scale which leads to over-fitting with more complex Transformer architectures, the encoder and decoder layers are set to 1 with one attention head both in AVT and VAT. All experiments are performed on a single NVIDIA RTX A6000 graphic card with 48GB memory. We use PyTorch[1] for the implementation of all experiments. More hyperparameter settings are shown in Sec. 4.4.

4.1 Comparison with State of the Art

Our TransZero++ is a non end-to-end and non-generative manner. We compare it with other state-of-the-art methods both in CZSL and GZSL settings, including end-to-end methods (*e.g.*, QFSL [18], SGMA [33], AREN [32], LFGAA [36], APN [34]), generative methods (*e.g.*, SE-ZSL [21], f-VAEGAN [8], Composer [67], E-PGN [39], SDGZSL [29], FREE [24], HSVA [23]) and non-generative methods (*e.g.*, SP-AEN [69], PQZSL [70], IIR [71],

1. https://pytorch.org/

TABLE 3: Ablation studies for different losses of TransZero++ on the CUB and SUN datasets. Note that $\mathcal{L}_{\text{SCL}} = \mathcal{L}_{\text{f\_SCL}} + \mathcal{L}_{\text{p\_SCL}}$.

Method	CUB				SUN			
	acc	U	S	H	acc	U	S	H
TransZero++(VAT) w/o $\mathcal{L}_{\text{SCL}}$	49.0	36.4	48.5	41.6	57.7	36.9	28.7	32.3
TransZero++(AVT) w/o $\mathcal{L}_{\text{SCL}}$	75.6	62.8	72.3	67.2	63.1	44.1	35.5	39.3
TransZero++(VAT) w/ $\mathcal{L}_{\text{SCL}}$	49.2	37.7	51.9	43.7	63.3	48.0	31.5	38.0
TransZero++(AVT) w/ $\mathcal{L}_{\text{SCL}}$	77.6	67.2	73.4	70.2	63.8	45.3	34.7	39.3
TransZero++ w/o $\mathcal{L}_{\text{SC}}$	77.0	46.6	76.4	58.9	65.1	41.5	36.4	38.7
TransZero++ w/o $\mathcal{L}_{\text{AR}}$	77.3	67.1	73.4	70.1	64.7	45.2	35.4	39.7
TransZero++(AVT and VAT) w/o $\mathcal{L}_{\text{f\_SCL}}$	78.1	67.9	72.1	69.9	65.6	47.4	37.5	41.9
TransZero++(AVT and VAT) w/o $\mathcal{L}_{\text{d\_SCL}}$	75.4	64.4	71.0	67.5	65.2	46.0	37.6	41.4
TransZero (full)	**78.3**	67.5	73.6	**70.4**	**67.6**	48.6	37.8	**42.5**

TCN [72], DVBE [40], DAZLE [31]), to demonstrate its effectiveness and advantages.

4.1.1 *Conventional Zero-Shot Learning.*

Here, we first compare our TransZero with the state-of-the-art methods in the CZSL setting. As shown in Table 1, our TransZero++ achieves the best accuracies of 78.3% and 67.6% on CUB and SUN, respectively. This shows that TransZero++ effectively learns the locality-augmented visual feature representations for distinguishing various fine-grained classes. On the coarse-grained dataset (*i.e.*, AWA2), TransZero++ still performs competitive performance, with the second-best top-1 accuracy of 72.6%. Compared with other attention-based methods (*e.g.*, SGMA [33], AREN [32], APN [34]), TransZero++ gets significant gains of over 6.3% and 5.0% on CUB and SUN, respectively. This demonstrates that the attribute localization representations learned by our TransZero++ are more discriminative than the region embeddings learned by the existing attention-based methods on fine-grained datasets. Benefiting from the semantic collaborative learning and the two complementary semantic-augmented embeddings learned by AVT and VAT, TransZero++ further improves the performance by 1.5%, 2.0% and 2.5% over its conference version (TransZero [20]) on CUB, SUN and AWA2, respectively.

4.1.2 *Generalized Zero-Shot Learning.*

Table 1 shows the results of different methods in the GZSL setting. The results show that the unseen accuracy ($\boldsymbol{U}$) of all methods is usually lower than the seen accuracy ($\boldsymbol{S}$) on the CUB and AWA2 datasets, *i.e.*, $\boldsymbol{U} < \boldsymbol{S}$. Meanwhile, $\boldsymbol{U} > \boldsymbol{S}$ on the SUN dataset since the number of seen classes is much larger than the number of unseen classes. Since per class only contains 16 training images on SUN, which heavily limits the ZSL models, the data augmentation is very effective for improving the performance on SUN. As such, most of the generative methods perform better than the non-generative methods on SUN, *e.g.*, GCM-CF [68], FREE [24] and HSVA [23].

We can see that most state-of-the-art methods achieve good results on seen classes but fail on unseen classes, while our TransZero++ generalizes better to unseen classes with high unseen and seen accuracies. For example, TransZero++ obtains the best performance with a harmonic mean of 70.4% and 72.5% on CUB and AWA2, respectively, and second-best performance with 42.5% on SUN. We argue these desirable results benefit from the fact that i) the feature augmentation encoders in AVT and VAT effectively refine the visual features that are more discriminative and transferable than the ones directly extracted from the CNN backbone, ii) the VAT and AVT localize the local attributes for locality-augmented feature representations. Compared to the latest attention-based method (*e.g.*, APN [34]), our TransZero++ achieves significant improvements of 3.2%, 4.9% and 9.6% in harmonic mean on CUB, SUN and AWA2, respectively. This demonstrates the superiority and great potential of our cross attribute-guided Transformer for the ZSL task. Since the semantical collaborative learning encourages AVT and VAT to learn semantic-augmented embedding for desirable visual-semantic interaction, TransZero++ continuously improve the performance of its conference version (TransZero [20]), *e.g.*, the improvements of 1.6%, 1.7% and 2.3% in harmonic mean on CUB, SUN and AWA2, respectively.

4.2 Ablation Study

To provide further insight into TransZero++, we conduct ablation studies to evaluate the effect of different model components, loss functions, and distance metrics for semantical collaborative losses.

Analysis of Model Components. As shown in Table 2, we conduct ablation studies to evaluate the effects of different model components , *i.e.*, AVT, VAT, feature augmentation encoder (denoted as FAE), feature augmentation in FAE (denoted as FA), and visual-semantic decoder (denoted as DEC). TransZero++ performs significantly worse than if no AVT is used, *i.e.*, the acc/harmonic mean drops by 29.3%/28.8% on CUB and 9.9%/10.2% on SUN. Meanwhile, Transzero++ also achieves inferior performance than the full model. These results show that it is necessary to simultaneously learn the semantic-augmented embeddings with AVT and VAT for ZSL. TransZero++ significantly improves its performance when AVT and VAT use the feature augmentation encoders, which shows the importance of refining the visual feature to alleviate the cross-dataset bias. If we incorporate the encoder of the standard Transformer without feature augmentation, TransZero again achieves poor results compared to its full model, *i.e.*, the acc/harmonic mean drops by 1.9%/2.0% and 2.3%/1.3% on CUB and SUN, respectively. When TransZero++ is without the decoders in AVT and VAT, its performance decreases dramatically on all datasets.

Analysis of Loss Functions. As shown in Table 3, we further conduct ablation studies to evaluate the effects of different loss functions, *i.e.*, semantical collaborative loss (including $\mathcal{L}_{\text{f\_SCL}}$ and $\mathcal{L}_{\text{p\_SCL}}$), self-calibration loss (*i.e.*, $\mathcal{L}_{\text{SC}}$) and attribute regression loss (*i.e.*, $\mathcal{L}_{\text{AR}}$). TransZero++ with sigle sub-net (*i.e.*, TransZero++(VAT) and TransZero++(VAT)) performs achieves significant gains on CUB and SUN when it no use the semantical collaborative loss ($\mathcal{L}_{\text{SCL}} = \mathcal{L}_{\text{f\_SCL}} + \mathcal{L}_{\text{p\_SCL}}$). For example, TransZero++(VAT) and TransZero++(AVT) achieves the gains of 2.1% and 3.0% in harmonic mean on CUB, respectively. This shows that the semantic

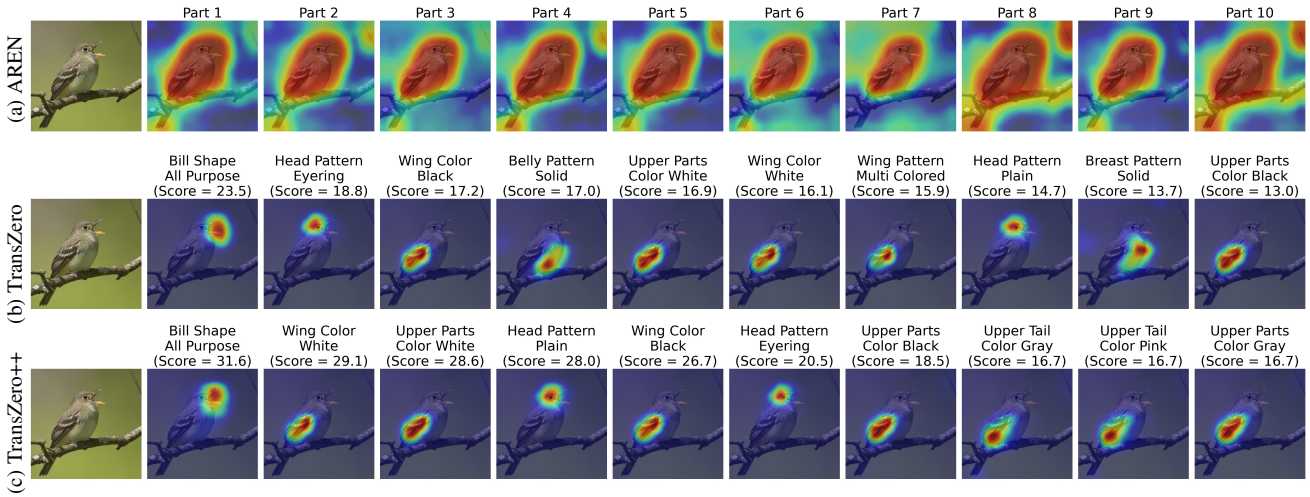


Fig. 4: Visualization of top-10 attention maps for the attention-based method (*i.e.*, AREN [32]), TransZero [20] (Conference version) and our TransZero++. Results show that TransZero localizes some important object attributes with low confident scores for representing region features, while AREN is failed. Furthermore, our TransZero++ discovers more valuable attributes that exist in the corresponding image with high confident scores compared to the TransZero. More results are presented in the Project Website. (Best viewed in color)

TABLE 4: Ablation studies for different components of TransZero++ on the CUB and SUN datasets. "FAE" is the feature augmentation encoder, "FA" means feature augmentation, and "DEC" denotes visual-semantic decoder.

Method	CUB				SUN			
	acc	U	S	H	acc	U	S	H
TransZero++ w/ ℓ_1	53.6	42.6	48.9	37.8	64.3	51.4	32.7	40.0
TransZero++ w/ KL($p_1 \| \| p_2$)	76.4	67.8	65.4	70.4	65.0	46.3	37.6	41.5
TransZero++ w/ KL($p_2 \| \| p_1$)	76.3	68.3	65.3	71.6	65.0	46.3	38.1	41.8
TransZero++ w/ JSD	76.1	67.2	63.2	71.8	64.8	46.2	37.9	41.6
TransZero++ w/ ℓ_2	**78.3**	67.5	73.6	**70.4**	**67.6**	48.6	37.8	**42.5**

collaborative learning is effective for enabling TransZero++ to learn more intrinsic semantic knowledge for ZSL. The self-calibration mechanism can effectively alleviate the seen-unseen bias problem [31], [33], [34], resulting in improvements in the harmonic mean of 11.5% and 3.8% on CUB and SUN, respectively. The attribute regression constraint further improves the performance of TransZero++ by directing $\mathcal{M}_1$ and $\mathcal{M}_1$ to conduct effective visual-semantic mapping. Furthemore, the two semantical collaborative losses (*i.e.*, $\mathcal{L}_{\text{f\_SCL}}$ and $\mathcal{L}_{\text{p\_SCL}}$) encourage TransZero++ to conduct desirable semantical collaborative learning.

Analysis of Distance Metrics for Semantical Collaborative Losses. As shown in Table 4, we conduct ablation studies to evaluate the effects of distance metrics for semantical collaborative losses (feature-level semantical collaborative loss ($\mathcal{L}_{\text{f\_SCL}}$) and prediction-level semantical collaborative loss ($\mathcal{L}_{\text{p\_SCL}}$)), *i.e.*, ℓ_1, ℓ_2, KL($p_1||p_2$), KL($p_2||p_1$), and JSD. Results show that TransZero++ performs very poorly using ℓ_1 distance for calculating the semantical collaborative losses. When TransZero++ use the KL($p_1||p_2$, KL($p_2||p_1$, or JSD to compute the semantical collaborative losses, it achieves consistent good performance almost. Thus, the symmetric and asymmetric distances for semantical collaborative losses do not make any difference. Interestingly, TransZero++ achieves the best results using ℓ_2 distance. As such, we take ℓ_2 to compute $\mathcal{L}_{\text{p\_SCL}}$ and $\mathcal{L}_{\text{f\_SCL}}$.

4.3 Qualitative Results

Here, we present the visualizations of attention maps and t-SNE [73] to intuitively show the effectiveness of our TransZero++.

4.3.1 *Visualization of Attention Maps.*

To intuitively show the effectiveness of our TransZero++ at learning attribute-relevant visual features, we visualize the attention maps learned by the existing attention-based methods (*e.g.*, AREN [32]) and TransZero++. As shown in Fig. 4, AREN simply learns region embeddings for visual representations, *e.g.*, the whole bird body, neglecting the fine-grained semantic attribute information. In contrast, our Transzero++ learns discriminative attribute localization for visual features by assigning high positive scores to key attributes (*e.g.*, the "bill shape all-purpose" of the *Acadian Flycatcher* in Fig. 4). Thus, TransZero++ discovers the semantic-augmented embeddings that are discriminative and transferable, enabling good performance both in seen and unseen classes. Compared to TransZero [20] (Conference version), TransZero++ can discover more valuable attributes for semantic-augmented embedding representations (*e.g.*, "upper part color gray" of *Acadian Flycatcher*). Furthermore, TransZero++ gets higher confidence scores for the important attributes that exist in the images than TransZero. For example, TransZero++ gets the scores of 28.0 for attribute "head pattern plain" of the *Acadian Flycatcher*, while TransZero gets the scores of 14.7.

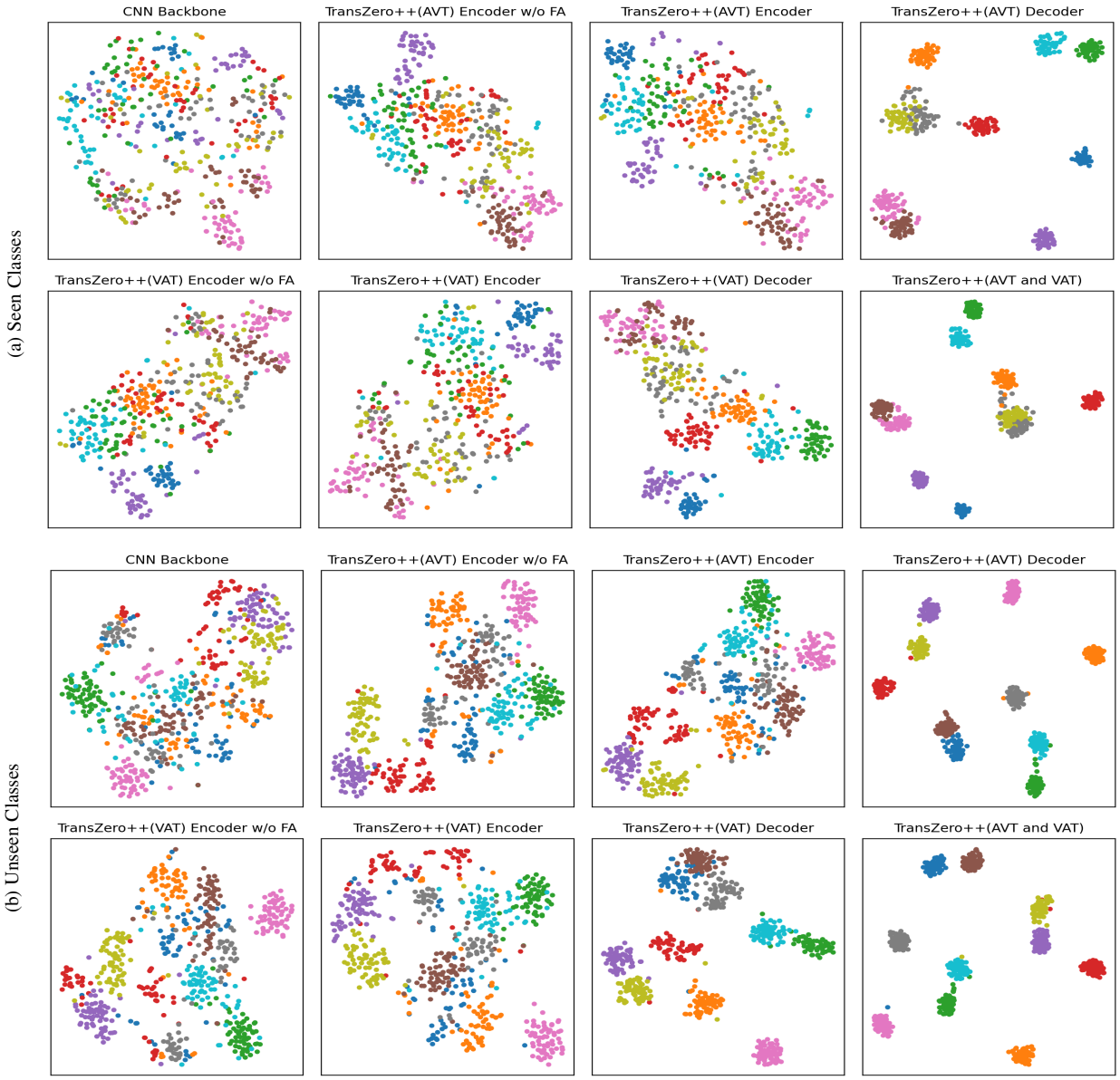


Fig. 5: t-SNE visualizations of visual features for (a) seen classes and (b) unseen classes, learned by the CNN backbone, TransZero++(AVT) encoder w/o FA, TransZero++(AVT) encoder, TransZero++(AVT) decoder, TransZero++(VAT) encoder w/o FA, TransZero++(VAT) encoder, TransZero++(VAT) decoder and TransZero++(VAT and VAT). The 10 colors denote 10 different seen/unseen classes randomly selected from CUB. Results show that our various model components in TransZero++ learn the discriminative visual feature representations, while CNN backbone (*e.g.*, ResNet101) failed. The results on SUN and AWA2 are presented in the Project Website. (Best viewed in color)

4.3.2 t-SNE Visualizations.

As shown in Fig. 5, we provide the t-SNE visualization [73] of visual features for (a) seen classes and (b) unseen classes on CUB, learned by the CNN backbone, TransZero++(AVT) encoder w/o FA, TransZero++(AVT) encoder, TransZero++(AVT) decoder, TransZero++(VAT) encoder w/o FA, TransZero++(VAT) encoder, TransZero++(VAT) decoder, TransZero++(AVT and VAT). If the standard encoder is incorporated into AVT and VAT of our TransZero++, the visual features learned by the encoder are significantly improved compared to the original visual features extracted from the CNN Backbone (*e.g.*, ResNet101). When we use the feature augmentation encoder to refine the original visual features, the quality of visual features is further enhanced. These results demonstrate that the encoder of TransZero++ effectively alleviates the cross-dataset bias problem and reduces the entangled relative geometry relationships among different regions, enabling the visual feature more discriminative and transferable. Moreover, the attribute→visual and visual→attribute decoders in AVT and VAT learn attribute-based visual features and visual-based attribute features, which are further mapped into semantic embedding space for semantic-augmented embedding representations. Since the features learned by AVT and VAT are complementary to each other, the fused semantic-augmented embedding can be further refined. As such, our TransZero++ achieves significant performance both

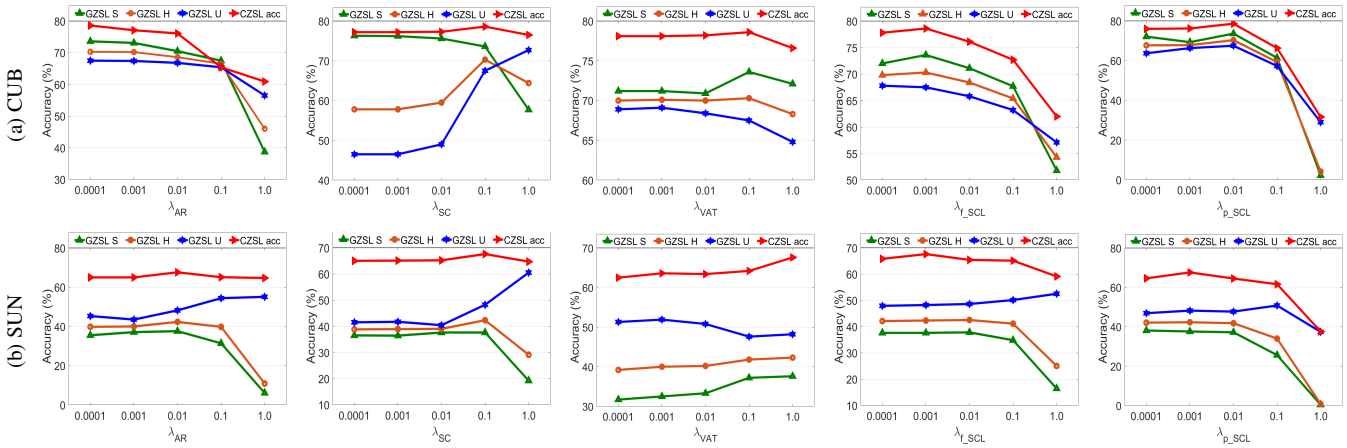


Fig. 6: The effects of loss weights that control their corresponding loss terms on CUB and SUN, *i.e.*, λ_{AR}, λ_{SC}, λ_{VAT}, $\lambda_{\text{f\_SCL}}$ and $\lambda_{\text{p\_SCL}}$.

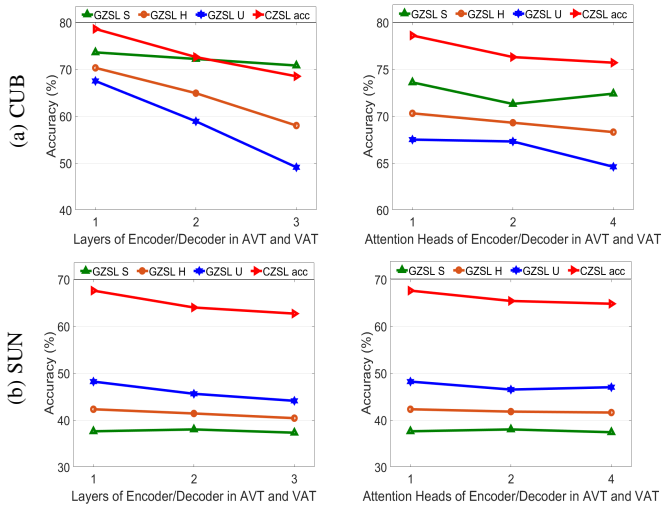


Fig. 7: The effects of different architectures for the AVT and VAT networks on CUB and SUN. We investigate the number of layers of Encoder/Decoder and attention heads in Encoder/Decoder.

in seen and unseen classes on all datasets.

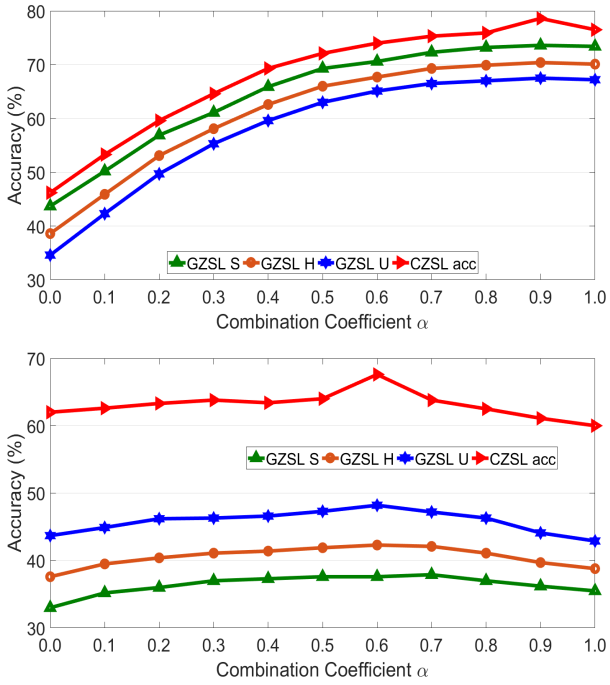


Fig. 8: The effectiveness of combination coefficients α between the AVT and VAT sub-nets on CUB (top) and SUN (bottom).

4.4 Hyperparameter Analysis

To analyse the robustness of our TransZero++ and select better hyperparameters for it. We conduct extensive experiments for evaluating the effects of loss weights (in Eq. 29 and Eq. 31) and combination coefficient (in Eq. 32).

4.4.1 *Effects of Loss Weights*

Here, we analyse the effects of loss weights that control their corresponding loss terms, *i.e.*, λ_{AR}, λ_{SC}, λ_{VAT}, $\lambda_{\text{f\_SCL}}$ and $\lambda_{\text{p\_SCL}}$. We try a wide range of these loss weights evaluated on CUB and SUN, *i.e.*, $\{0.0001, 0.001, 0.01, 0.1, 1.0\}$. Results are shown in Fig. 6. When λ_{AR}, $\lambda_{\text{f\_SCL}}$ and $\lambda_{\text{p\_SCL}}$ are set to a large value, all evaluation protocols tend to drop. Moreover, TansZero++ are relatively insensitive to λ_{SC} and λ_{VAT} when they are set to small (*e.g.*, smaller than 0.01). Based on these observations, we set $\{\lambda_{\text{AR}}, \lambda_{\text{SC}}, \lambda_{\text{VAT}}, \lambda_{\text{f\_SCL}}, \lambda_{\text{p\_SCL}}\}$ to $\{0.0001, 0.1, 0.1, 0.001, 0.01\}$ and $\{0.01, 0.1, 1.0, 0.001, 0.001\}$ for CUB and SUN datasets, repectively.

4.4.2 *Effects of Combination Coefficient*

We argue that the attribute-based visual features and visual-based attribute features learned by AVT and VAT respectively are complementary, and thus we take a combination coefficient α to fuse their corresponding semantic-augmented embeddings for desirable visual-semantic interaction (in Eq. 32). We try a wide range of α on CUB and SUN, *i.e.*, $\{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. Notably, $\alpha = 0.0$ is denoted as the TransZero++(VAT), and $\alpha = 1.0$ is denoted as the TransZero++(AVT). As shown in Fig. 8, when α is set to large relatively (*e.g.*, $\alpha > 0.5$), TransZero++ achieves better results. This demonstrates that AVT sub-net provides more desirable information for TransZero++. However, α should also not be set too large since the VAT sun-net provides additional useful information for TransZero++. Based on these results, we set α to 0.9 and 0.6 for CUB and SUN, respectively.

5 Conclusion

In this paper, we propose a novel cross attribute-guided Transformer network for ZSL, termed TransZero++. TransZero++ consists of a attribute→visual Transformer sub-net (AVT) and visual→attribute Transformer sub-net (VAT). First, AVT employs a feature augmentation encoder to improve the discriminability and transferability of visual features by alleviating the cross-dataset problem and reducing the entangled region feature relationships, respectively. Meanwhile, an attribute→visual decoder in AVT is introduced to learn the attribute localization for attribute-based visual feature representations which are locality-augmented. Secondly, VAT applied a similar feature augmentation encoder to refine the visual features, which is further fed into a visual→attribute decoder to learn the visual-based attribute features. By introducing the feature-level and prediction-level semantical collaborative losses for optimization, our TransZero++ can learn the semantic-augmented visual embedding. Considering the attribute-based visual features and visual-based attribute features that are complementary to each other, we combine the two semantic-augmented visual embeddings learned by AVT and VAT to enable desirable visual-semantic interaction cooperated with the class semantic vectors for ZSL classification. Extensive experiments on three popular challenging benchmark datasets demonstrate the superiority of our method. We believe that our work also facilitates the development of other visual-and-language learning systems, *e.g.*, image captioning [66], natural language for visual reasoning [74].

Acknowledgements

This work is partially supported by NSFC (61772220,62006244), Special projects for technological innovation in Hubei Province (2018ACA135), Key R&D Plan of Hubei Province (2020BAB027) and 2020-2022 Young Elite Scientist Sponsorship Program from China Association for Science and Technology YESS20200140.

References

[1] H. Larochelle, D. Erhan, and Y. Bengio, "Zero-data learning of new tasks," in *AAAI*, 2008, pp. 646–651.
[2] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *NeurIPS*, 2009, pp. 1410–1418.
[3] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *CVPR*, 2009, pp. 951–958.
[4] C. H. Lampert, S. Harmeling, and H. Nickisch, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 453–465, 2014.
[5] Z. Fu, T. Xiang, E. Kodirov, and S. Gong, "Zero-shot learning on semantic class prototype graph," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 2009–2022, 2018.
[6] FuYanwei, M. HospedalesTimothy, Xiang-tao, and GongShaogang, "Transductive multi-view zero-shot learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 2332–2345, 2015.
[7] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *NeurIPS*, 2013.
[8] Y. Xian, S. Sharma, B. Schiele, and Z. Akata, "F-vaegan-d2: A feature generating framework for any-shot learning," in *CVPR*, 2019, pp. 10 267–10 276.
[9] Y. Shen, L. Liu, F. Shen, and L. Shao, "Zero-shot sketch-image hashing," in *CVPR*, 2018, pp. 3598–3607.
[10] A. Dutta and Z. Akata, "Semantically tied paired cycle consistency for any-shot sketch-based image retrieval," *International Journal of Computer Vision*, pp. 1–20, 2020.
[11] M. Bucher, T.-H. Vu, M. Cord, and P. Pérez, "Zero-shot semantic segmentation," in *NeurIPS*, 2019, pp. 468–479.
[12] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran, "Zero-shot object detection," in *ECCV*, 2018.
[13] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *CVPR*, 2016, pp. 49–58.
[14] S. Badirli, Z. Akata, G. O. Mohler, C. Picard, and M. Dundar, "Fine-grained zero-shot learning with dna as side information," in *NeurIPS*, 2021.
[15] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning — the good, the bad and the ugly," in *CVPR*, 2017, pp. 3077–3086.
[16] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *CVPR*, 2018, pp. 5542–5551.
[17] G.-S. Xie, Z. Zhang, G.-S. Liu, F. Zhu, L. Liu, L. Shao, and X. Li, "Generalized zero-shot learning with multiple graph adaptive generative networks." *IEEE transactions on neural networks and learning systems*, 2021.
[18] J. Song, C. Shen, Y. Yang, Y. Liu, and M. Song, "Transductive unbiased embedding for zero-shot learning," *CVPR*, pp. 1024–1033, 2018.
[19] G.-S. Xie, X.-Y. Zhang, Y. Yao, Z. Zhang, F. Zhao, and L. Shao, "Vman: A virtual mainstay alignment network for transductive zero-shot learning," *IEEE Transactions on Image Processing*, vol. 30, pp. 4316–4329, 2021.
[20] S. Chen, Z. Hong, Y. Liu, G.-S. Xie, B. Sun, H. Li, Q. Peng, K. Lu, and X. You, "Transzero: Attribute-guided transformer for zero-shot learning," in *AAAI*, 2022.
[21] G. Arora, V. Verma, A. Mishra, and P. Rai, "Generalized zero-shot learning via synthesized examples," in *CVPR*, 2018, pp. 4281–4289.
[22] E. Schönfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized zero- and few-shot learning via aligned variational autoencoders," in *CVPR*, 2019, pp. 8239–8247.
[23] S. Chen, G.-S. Xie, Y. Yang Liu, Q. Peng, B. Sun, H. Li, X. You, and L. Shao, "Hsva: Hierarchical semantic-visual adaptation for zero-shot learning," in *NeurIPS*, 2021.
[24] S. Chen, W. Wang, B. Xia, Q. Peng, X. You, F. Zheng, and L. Shao, "Free: Feature refinement for generalized zero-shot learning," in *ICCV*, 2021.
[25] Y. Shen, J. Qin, and L. Huang, "Invertible zero-shot recognition flows," in *ECCV*, 2020.
[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.
[28] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. J. Belongie, and P. Perona, "Caltech-ucsd birds 200," *Technical Report CNS-TR-2010-001, Caltech,*, 2010.
[29] Z. Chen, Y. Luo, R. Qiu, S. Wang, Z.-Y. Huang, J. Li, and Z. Zhang, "Semantics disentangling for generalized zero-shot learning," in *ICCV*, 2021.
[30] G.-S. Xie, L. Liu, X. Jin, F. Zhu, Z. Zhang, Y. Yao, J. Qin, and L. Shao, "Region graph embedding network for zero-shot learning," in *ECCV*, 2020.
[31] D. Huynh and E. Elhamifar, "Fine-grained generalized zero-shot learning via dense attribute-based attention," in *CVPR*, 2020, pp. 4482–4492.
[32] G.-S. Xie, L. Liu, X. Jin, F. Zhu, Z. Zhang, J. Qin, Y. Yao, and L. Shao, "Attentive region embedding network for zero-shot learning," in *CVPR*, 2019, pp. 9376–9385.
[33] Y. Zhu, J. Xie, Z. Tang, X. Peng, and A. Elgammal, "Semantic-guided multi-attention localization for zero-shot learning," in *NeurIPS*, 2019.
[34] W. Xu, Y. Xian, J. Wang, B. Schiele, and Z. Akata, "Attribute prototype network for zero-shot learning," in *NeurIPS*, 2020.
[35] Y. Yu, Z. Ji, Y. Fu, J. Guo, Y. Pang, and Z. Zhang, "Stacked semantics-guided attention model for fine-grained zero-shot learning," in *NeurIPS*, 2018.
[36] Y. Liu, J. Guo, D. Cai, and X. He, "Attribute attention for semantic disambiguation in zero-shot learning," in *ICCV*, 2019, pp. 6697–6706.
[37] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *CVPR*, 2012, pp. 2751–2758.
[38] Y. Li, J. Zhang, J. Zhang, and K. Huang, "Discriminative learning of latent features for zero-shot recognition," in *CVPR*, 2018, pp. 7463–7471.
[39] Y. Yu, Z. Ji, J. Han, and Z. Zhang, "Episode-based prototype generating network for zero-shot learning," in *CVPR*, 2020, pp. 14 032–14 041.
[40] S. Min, H. Yao, H. Xie, C. Wang, Z. Zha, and Y. Zhang, "Domain-aware visual bias eliminating for generalized zero-shot learning," in *CVPR*, 2020, pp. 12 661–12 670.
[41] Z. Han, Z. Fu, S. Chen, and J. Yang, "Contrastive embedding for generalized zero-shot learning," in *CVPR*, 2021.
[42] Y.-Y. Chou, H.-T. Lin, and T.-L. Liu, "Adaptive and generative zero-shot learning," in *ICLR*, 2021.

[43] S. Narayan, A. Gupta, F. Khan, C. G. M. Snoek, and L. Shao, "Latent embedding feedback and discriminative features for zero-shot classification," in *ECCV*, 2020.

[44] C. Yan, X. Chang, Z. Li, Z. Ge, W. Guan, L. Zhu, and Q. Zheng, "Zeronas: Differentiable generative adversarial networks search for zero-shot learning." *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[45] Y.-H. H. Tsai, L.-K. Huang, and R. Salakhutdinov, "Learning robust visual-semantic embeddings," in *ICCV*, 2017, pp. 3591–3600.

[46] Y. Atzmon, F. Kreuk, U. Shalit, and G. Chechik, "A causal view of compositional zero-shot recognition," in *NeurIPS*, 2020.

[47] T. Chen, T. Pu, Y. Xie, H. Wu, L. Liu, and L. Lin, "Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning." *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[48] Y. Liu, L. Zhou, X. Bai, Y. Huang, L. Gu, J. Zhou, and T. Harada, "Goal-oriented gaze estimation for zero-shot learning," in *CVPR*, 2021.

[49] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[50] M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Intriguing properties of vision transformers," in *NeurIPS*, 2021.

[51] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. Khan, and M. Shah, "Transformers in vision: A survey," *arXiv preprint arXiv:2101.01169*, 2021.

[52] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "Metaformer is actually what you need for vision," *arXiv preprint arXiv:2111.11418*, 2021.

[53] M. Ott, S. Edunov, D. Grangier, and M. Auli, "Scaling neural machine translation," in *WMT*, 2018.

[54] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[55] V. Gabeur, C. Sun, A. Karteek, and C. Schmid, "Multi-modal transformer for video retrieval," in *ECCV*, 2020.

[56] X. Zhang, X. Sun, Y. Luo, J. Ji, Y. Zhou, Y. Wu, F. Huang, and R. Ji, "Rstnet: Captioning with adaptive attention on visual and non-visual words," in *CVPR*, 2021.

[57] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *CVPR*, 2020, pp. 10 575–10 584.

[58] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *ICCV*, 2019, pp. 4633–4642.

[59] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *CVPR*, 2020, pp. 10 968–10 977.

[60] T. Batra and D. Parikh, "Cooperative learning with visual attributes," *arXiv preprint arXiv: 1705.05512*, 2017.

[61] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NeurIPS*, 2017.

[62] Y. Ge, D. peng Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," in *ICLR*, 2020.

[63] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," *CVPR*, pp. 4320–4328, 2018.

[64] Y. Zhai, Q. Ye, S. Lu, M. Jia, R. Ji, and Y. Tian, "Multiple expert brainstorming for domain adaptive person re-identification," in *ECCV*, 2020.

[65] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014.

[66] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *NeurIPS*, 2019.

[67] D. T. Huynh and E. Elhamifar, "Compositional zero-shot learning via fine-grained dense feature composition," in *NeurIPS*, 2020.

[68] Z. Yue, T. Wang, H. Zhang, Q. Sun, and X. Hua, "Counterfactual zero-shot and open-set visual recognition," in *CVPR*, 2021.

[69] L. Chen, H. Zhang, J. Xiao, W. Liu, and S. Chang, "Zero-shot visual recognition using semantics-preserving adversarial embedding networks," in *CVPR*, 2018, pp. 1043–1052.

[70] J. Li, X. Lan, Y. Liu, L. Wang, and N. Zheng, "Compressing unknown images with product quantizer for efficient zero-shot classification," in *CVPR*, 2019, pp. 5458–5467.

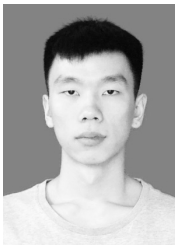
[71] Y. L. Cacheux, H. Borgne, and M. Crucianu, "Modeling inter and intra-class relations in the triplet loss for zero-shot learning," in *ICCV*, 2019, pp. 10 332–10 341.

[72] H. Jiang, R. Wang, S. Shan, and X. Chen, "Transferable contrastive network for generalized zero-shot learning," in *ICCV*, 2019, pp. 9764–9773.

[73] L. V. D. Maaten and G. E. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[74] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *ICCV*, 2015.

Shiming Chen is currently a full-time Ph.D. student in the School of Electronic Information and Communications, Huazhong University of Sciences and Technology (HUST), China. His research results have expounded in prominent conferences and prestigious journals, such as NeurIPS, ICCV, AAAI, IJCAI, IEEE TEVC. He serves as the reviewer for prestigious journals such as *IEEE Transactions on Image Processing, IEEE Transactions on Systems, Man, and Cybernetics: Systems, IEEE Transactions on Industrial Informatics, Information Fusion, Information Sciences, and Applied Soft Computing*. His current research interests span computer vision and machine learning with a series of topics, such as generative modeling and learning, zero-shot learning, and domain adaptation.



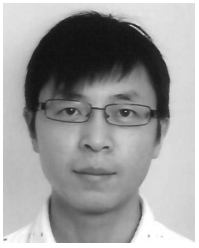
Ziming Hong is currently pursuing the M.Sc. degree in the School of Electronic Information and Communications(EIC), Huazhong University of Sciences and Technology(HUST), China. He received the B.E. degree in the School of Information Engineering, Wuwan University of Technology(WHUT), in 2019. His current research interests include graph learning and computer vision.

Guo-Sen Xie received the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2016. He was with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, UAE. His research results have expounded in 20+ publications at prestigious journals and prominent conferences, such as IEEE T-IP, T-NNLS, T-CSVT, NeurIPS, CVPR, ICCV, ECCV, AAAI. His research interests include computer vision and machine learning.

Jian Zhao is currently an Assistant Professor with the Institute of North Electronic Equipment, Beijing, China. He received his Ph.D. degree from the National University of Singapore (NUS) in 2019 under the supervision of Assist. Prof. Jiashi Feng and Assoc. Prof. Shuicheng Yan. He is the SAC of VALSE, the member of the CSIG committee for elite young professionals, the member of the board of directors of BSIG, the committee member of CSIG-BVD, the PaddlePaddle developer expert, and the member of the expert committee/arbitration of China artificial intelligence competition. He has served as the guest editor of PRL and Electronics, the presentation chair of the CICAI'21, the session chair of the ACM MM'21, and the invited reviewer of NSFC, T-PAMI, IJCV, NeurIPS (one of the top 30% highest-scoring reviewers of NeurIPS 2018), CVPR, etc. He has received the "2020-2022 Young Elite Scientist Sponsorship Program" from China Association for Science and Technology, and the "2021-2023 Beijing Young Elite Scientist Sponsorship Program" from Beijing Association for Science and Technology. His main research interests include deep learning, pattern recognition, computer vision and multimedia. He has published over 40 cutting-edge papers (e.g., T- PAMI, IJCV, NeurIPS, CVPR, etc.). He has received the nomination for the USERN Prize 2021, and won the Lee Hwee Kuan Award (Gold Award) on PREMIA'19 and the "Best Student Paper Award" on ACM MM'18 as the first author. He has won the top-3 awards several times in worldwide competitions on face recognition, human parsing, and pose estimation as the first author.

Xinge You (Senior Member, IEEE) is currently a Professor with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan. He received the B.S. and M.S. degrees in mathematics from Hubei University, Wuhan, China, in 1990 and 2000, respectively, and the Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, in 2004. His research results have expounded in 60+ publications at prestigious journals and prominent conferences, such as IEEE T-PAMI, T-IP, T-NNLS, NeurIPS, CVPR, ICCV, ECCV, IJCAI. He served/serves as an Associate Editor of the *IEEE Transactions on Cybernetics*, *IEEE Transactions on Systems, Man, Cybernetics:Systems*. His current research interests include image processing, wavelet analysis and its applications, pattern recognition, machine earning, and computer vision.

Shuicheng Yan (Fellow, IEEE) is currently the director of Sea AI Lab (SAIL) and group chief scientist of Sea. He is an Fellow of Academy of Engineering, Singapore, IEEE Fellow, ACM Fellow, IAPR Fellow. His research areas include computer vision, machine learning and multimedia analysis. Till now, he has published over 600 papers in top international journals and conferences, with Google Scholar Citation over 40,000 times and H-index 105. He had been among "Thomson Reuters Highly Cited Researchers" in 2014, 2015, 2016, 2018, 2019. Dr. Yan's team has received winner or honorable-mention prizes for 10 times of two core competitions, Pascal VOC and ImageNet (ILSVRC), which are deemed as "World Cup" in the computer vision community. Also his team won over 10 best paper or best student paper prizes and especially, a grand slam in ACM MM, the top conference in multimedia, including Best Paper Award, Best Student Paper Award and Best Demo Award.



Ling Shao (Fellow, IEEE) is the CTO and Chief Scientist of the National Center for Artificial Intelligence (NCAI), Saudi Data and Artificial Intelligence Authority (SDAIA), Riyadh, Saudi Arabia. He received the Ph.D. degree in Computer Vision, University of Oxford, England, in 2013. His research results have expounded in 40+ publications at prestigious journals and prominent conferences, such as IEEE T-PAMI, IJCV, T-IP, NeurIPS, ICCV, CVPR, ECCV. He has been an Associate Editor of *IEEE Transactions on Image Processing*, *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Circuits and Systems for Video Technology*, *IEEE Transactions on Cybernetics*, and several other journals. He has edited four books and several special issues for journals such as IJCV. He has organized a number of international workshops with top conferences including ICCV, ECCV and ACM MM. He was the General Chair for BMVC 2018, has been an Area Chair or Program Committee member for many conferences, including ICCV, CVPR, ECCV and ACM MM.

He was selected as a Highly Cited Researcher by the Web of Science in 2018-2020. He is a Fellow of the IEEE/IAPR/IET/British Computer Society. His research interests include computer vision and machine learning.