TransZero: Attribute-Guided Transformer for Zero-Shot Learning

Authors

> Chen S, Hong Z, Liu Y, et al. Transzero: Attribute-Guided Transformer for zero-shot learning[C]//AAAI. 2022, 2: 3.

Shiming Chen 陈使明

PhD student

1037 Luoyu Road,
National Anti-counterfeit Engineering Research Center,
Huazhong University of Science and Technology (HUST),
Wuhan, China, 430074

Email: shimingchen **at** hust **dot** edu **dot** cn; gchenshiming **at** gmail **dot** com

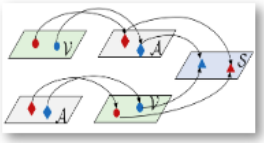
Biography

I come from Meizhou, Guangdong, where is known as "*Hakka capital*" and " *City of Football* ".
I am currently a Third-year PhD student in School of Electronic Information and Communications at Huazhong University of Science and Technology, under the supervision by Prof. Xinge You. I'm also visiting at Trustworthy Machine Learning Lab (TML Lab), University of Sydney, working with Prof. Tongliang Liu. My current research interests span computer vision and machine learning with a series of topics, such as ***zero-shot learning***, ***generative modeling and learning***, and ***visual-and-language learning***.

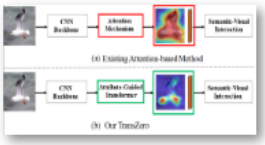
Publications

Conference Papers

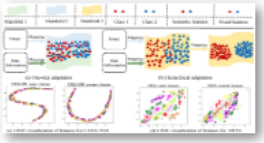
Semantic Compression Embedding for Generative Zero-Shot Learning.
Ziming Hong*, **Shiming Chen**[*#], Guo-Sen Xie, Wenhan Yang, Jian Zhao, Yuanjie Shao, Qinmu Peng, Xinge You[#] (*:co-first author; #:corresponding author)
*The 31th International Joint Conference on Artificial Intelligence (**IJCAI**), 2022: 956-963. (CCF-A)*



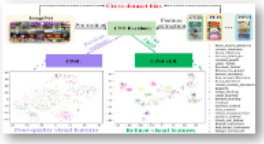
MSDN: Mutually Semantic Distillation Network for Zero-Shot Learning. **[PDF] [arXiv] [Code]**
Shiming Chen, Ziming Hong, Guo-Sen Xie, Wenhan Yang, Qinmu Peng, Kai Wang, Jian Zhao, Xinge You.
*IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2022: 7612-7621. (CCF-A)*



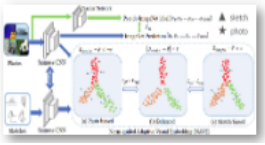
TransZero: Attribute-guided Transformer for Zero-Shot Learning. **[PDF] [arXiv] [Code]**
Shiming Chen, Ziming Hong, Yang Liu, Guo-Sen Xie, Baigui Sun, Hao Li, Qinmu Peng, Ke Lu, Xinge You.
*Thirty-Sixth AAAI Conference on Artificial Intelligence (**AAAI**), 2022: 330-338. (CCF-A)*



HSVA: Hierarchical Semantic-Visual Adaptation for Zero-Shot Learning. **[PDF] [arXiv] [Code]**
Shiming Chen, Guo-Sen Xie, Qinmu Peng, Yang Liu, Baigui Sun, Hao Li, Xinge You, Ling Shao.
*Annual Conference on Neural Information Processing Systems (**NeurIPS**), 2021: 16622-16634. (CCF-A)*

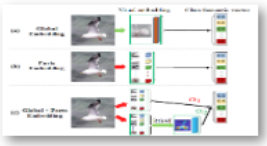


FREE: Feature Refinement for Generalized Zero-shot Learning. **[PDF] [arXiv] [Code]**
Shiming Chen, Wenjie Wang, Beihao Xia, Qinmu Peng, Xinge You, Feng Zheng, Ling Shao.
*IEEE International Conference on Computer Vision (**ICCV**), 2021: 1106-1112. (CCF-A)*

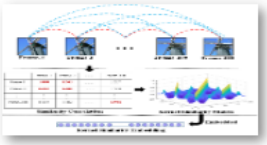


Norm-guided Adaptive Visual Embedding for Zero-Shot Sketch-Based Image Retrieval. **[PDF]**
Wenjie Wang, Yufeng Shi, **Shiming Chen**, Qinmu Peng, Feng Zheng, Xinge You
*The 30th International Joint Conference on Artificial Intelligence (**IJCAI**), 2021. (CCF-A)*

Journal Papers

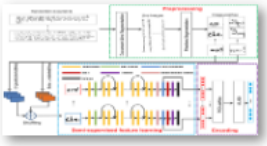


GNDAN: Graph Navigated Dual Attention Network for Zero-Shot Learning. **[Code] [PDF]**
Shiming Chen, Ziming Hong, Guo-Sen Xie, Xinge You, Weiping Ding and Ling Shao.
*IEEE Transactions on Neural Networks and Learning Systems (**TNNLS**), to appear, 2022. (SCI, IF=14.255)*

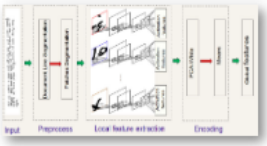


Kernelized Similarity Learning and Embedding for Dynamic Texture Synthesis. **[Code] [arXiv]**
Shiming Chen, Peng Zhang, Guo-sen Xie, Zehong Cao, Qinmu Peng, Wei Yuan, Xinge You.
*IEEE Transactions on Systems, Man and Cybernetics: Systems (**TSMCA**), to appear, 2022. (SCI, IF=11.471)*

CDE-GAN: Cooperative Dual Evolution Based Generative Adversarial Network. **[PDF] [arXiv]**
Shiming Chen, Wenjie Wang, Beihao Xia, Xinge You, Qinmu Peng, Zehong Cao, Weiping Ding.
*IEEE Transactions on Evolutionary Computation (**TEVC**), 25:986-1000, 2021. (SCI, IF=16.497)*



Semi-Supervised Feature Learning for Improving Writer Identification. **[Code]**
Shiming Chen, Yisong Wang, Chin-Teng Lin, Weiping Ding, Zehong Cao.
*Information Sciences (**INS**), 482:156-170, 2019. (SCI, IF=8.233)*



A Robust Offline Writer Identification Method. **[Code]**
Shiming Chen, Yisong Wang.
ACTA AUTOMATICA SINICA (自动化学报), 46(1):108-116, 2020. **(In Chinese, CAA-A,CCF-A,卓越期刊)**

Xinge You

关注

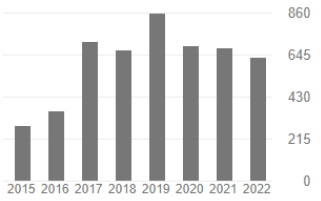
Professor of School of Electronics Information and Communications, Huazhong University of Science
在 mail.hust.edu.cn 的电子邮件经过验证 - 首页
Computer Vision Pattern Recognition Machine Learning Wavelet Analysis and its Ap...

创建我的个人资料

标题	引用次数	年份
Segmentation of retinal blood vessels using the radial projection and semi-supervised approach EI检索 SCI升级版 计算机科学1区 SCI基础版 工程技术2区 JCI 1.92 简介 SCI Q1 SCIIF(5) 7.299 SCIIF 8.52 CCF B SCU 计算机科学B CUG 工程技术T2 XJU 一区 NJU B XDU 1类贡献度 SWJTU A++ CUFE AAA SWUFE A X You, Q Peng, Y Yuan, Y Cheung, J Lei Pattern recognition 44 (10-11), 2314-2324	400	2011
Multiscale patch-based contrast measure for small infrared target detection EI检索 SCI升级版 计算机科学1区 SCI基础版 工程技术2区 JCI 1.92 简介 SCI Q1 SCIIF(5) 7.299 SCIIF 8.52 CCF B SCU 计算机科学B CUG 工程技术T2 XJU 一区 NJU B XDU 1类贡献度 SWJTU A++ CUFE AAA SWUFE A Y Wei, X You, H Li Pattern Recognition 58, 216-226	314	2016
Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning XY Jing, X Zhu, F Wu, X You, Q Liu, D Yue, R Hu, B Xu Proceedings of the IEEE Conference on Computer Vision and Pattern ...	237	2015
Shape matching and classification using height functions EI检索 SCI升级版 计算机科学3区 SCI基础版 工程技术3区 JCI 0.92 简介 SCI Q2 SCIIF(5) 3.615 SCIIF 4.76 CCF C SCU 计算机科学C CUG 工程技术T3 XJU 三区 XDU 2类贡献度 SWJTU A CUFE AA SWUFE B J Wang, X Bai, X You, W Liu, LJ Latecki Pattern Recognition Letters 33 (2), 134-143	224	2012

引用次数 查看全部

	总计	2017 年至今
引用	5908	4245
h 指数	42	33
i10 指数	98	77



开放获取的出版物数量 查看全部

40 篇文章 36 篇文章
无法查看的文章 可查看的文章
根据资金授权书

Motivation

- 基于生成的方法
  - 通过利用生成模型来生成 unseen classes 的样本，从而将 ZSL 转换为监督分类问题
  - ***缺点***
    - 依赖于全局视觉特征，不足以表示类的细粒度信息
    - 视觉表征有限，导致 visual-semantic 交互效果不好
- 基于注意力的方法
  - 试图借助语义信息学习更多的区域判别特征
  - ***缺点***
    - 直接将 entangled region (grid) 特征用于 ZSL 分类
    - 只学习了 region embeddings，忽略了将属性进行定位

- **本文方法（TransZero）**
  - 额外利用了各个属性的 word2vec 的语义信息
  - 减少 ImageNet 和 ZSL benchmarks 间的跨数据集的偏差
  - 减少区域特征间的 entangled relationships，以改善对于 unseen classes 的迁移性
  - 将属性定位到图像中最相关的位置，以学习局部增强的视觉特征，并进行 visual-semantic 交互

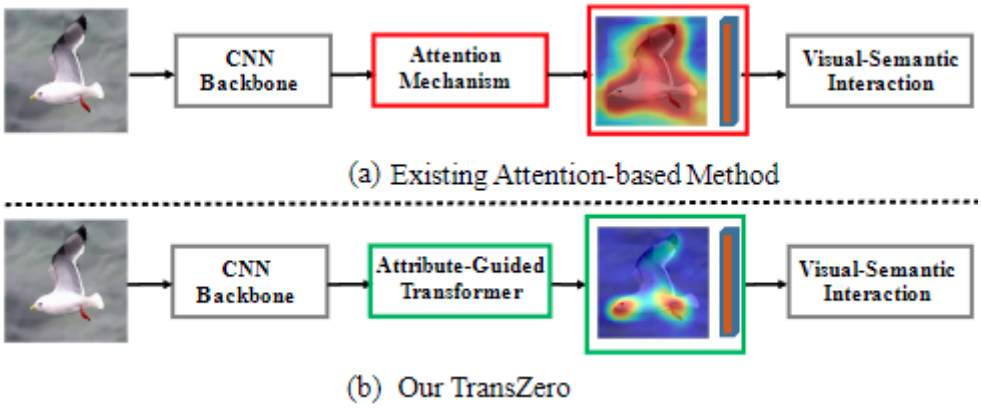


Figure 1: Motivation illustration. (a) Existing attention-based ZSL methods simply learn region embeddings (e.g., the whole bird body), neglecting the transferability and discriminative attribute localization (e.g., the distinctive bird body parts) of visual features; (b) Our TransZero reduces the entangled relationships among region features to improve their transferability and localizes the object attributes to represent discriminative region features, enabling significant visual-semantic interaction.

Framework

TransZero 结构如下：

- Attribute-Guided Transformer (AGT)
- Visual-Semantic Embedding Network (VSEN)

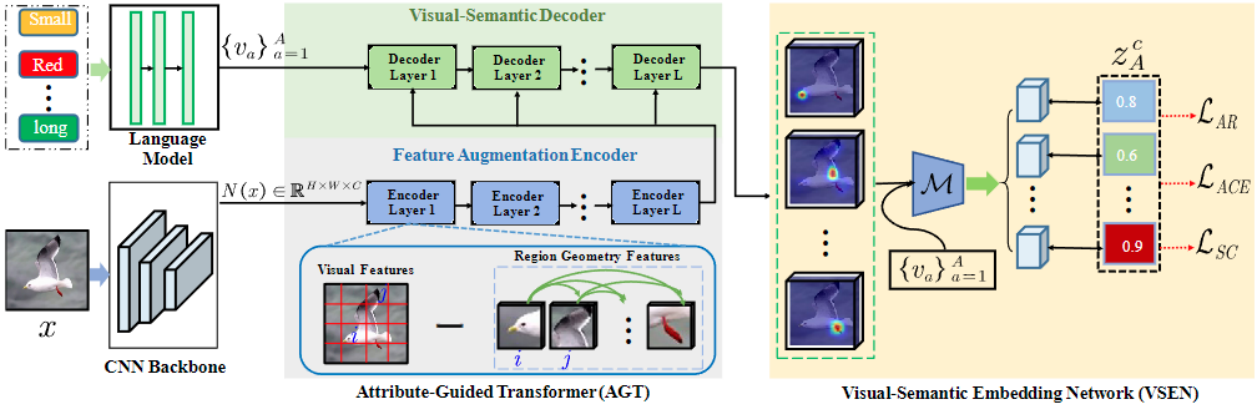


Figure 2: The architecture of the proposed TransZero model. TransZero consists of an attribute-guided Transformer (AGT) and a visual-semantic embedding network (VSEN). AGT includes a feature augmentation encoder that alleviates the cross-dataset bias between ImageNet and ZSL benchmarks and reduces the entangled geometry relationships between different regions for improving the transferability from seen to unseen classes, and a visual-semantic decoder that learns locality-augmented visual features based on the semantic attribute information. VSEN is used to enable significant visual-semantic interaction.

Attribute-Guided Transformer

Feature Augmentation Encoder

- 计算第 i 个 grid 的相对中心坐标，构建第 i 和 j 个 grid 的 region geometry features G_{ij}

To learn relative geometry features (Herdade et al. 2019; Zhang et al. 2021), we first calculate the relative center coordinates $(v_i^{\text{cen}}, t_i^{\text{cen}})$ based on the pair of 2D relative positions of the i-th grid $\left\{\left(v_i^{\min}, t_i^{\min}\right), \left(v_i^{\max}, t_i^{\max}\right)\right\}$:

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

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

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

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.

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.

- feature-augmented scaled dot-product attention
  - 将视觉特征减去 region geometry features 得到 augmented features

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:

$$Q^e = UW_q^e, K^e = UW_k^e, V^e = UW_v^e, \tag{6}$$

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

$$U \leftarrow U + Z_{aug}, \tag{8}$$

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. $U \in \mathbb{R}^{IIW \times C}$ 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.

Visual-Semantic Decoder

- 结构与标准 Transformer Decoder 相同
- 语义属性 $\mathcal{V}_A$ 作为 queries、encoder 的输出 U 作为 keys 和 values
- 得到局部增强的视觉特征 F

Visual-Semantic Decoder. Following the standard Transformer (Vaswani et al. 2017), our visual-semantic decoder takes a multi-head self-attention layer and feed-forward network (FFN) to build the decoder layer. The decoding process continuously incorporates visual information under the guidance of semantic attribute features $\mathcal{V}_A$. Thus, our visual-semantic 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 as keys (K_t^d) and values (V_t^d) and a set of learnable semantic embeddings $\mathcal{V}_A$ as queries (Q_t^d). It is defined as:

$$Q_t^d = \mathcal{V}_A W_{qt}^d, K_t^d = U W_{kt}^d, V_t^d = U W_{vt}^d, \tag{9}$$

$$\text{head}_t = \text{softmax}\left(\frac{Q_t^d K_t^{d^\top}}{\sqrt{d^d}}\right) V_t^d, \tag{10}$$

$$\hat{F} = \|_{t=1}^{T} (\text{head}_t) W_o, \tag{11}$$

where W_{qt}^d, W_{kt}^d, W_{vt}^d are the learnable weights, d^d is a scaling factor, and $\|$ is a concatenation operation. Then, an FFN with two linear transformations followed a ReLU activation in between is applied to the attended features $\hat{F}$:

$$F = ReLu\left(\hat{F} W_1 + b_1\right) W_2 + b_2, \tag{12}$$

where W_1, W_2, b_1 and b_2 are the weights and biases of the linear layers respectively, and F are the locality-augmented visual features.

Visual-Semantic Embedding Network

- 将局部增强的视觉特征 F 映射到语义空间

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

$$\psi(x_i) = \mathcal{M}(F) = \mathcal{V}_A^{\top} W F, \tag{13}$$

where W is an embedding matrix that embeds F into the semantic attribute space. In essence, $\psi(x_i)[a]$ 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 $\mathcal{V}_A = \{v_a\}_{a=1}^{A}$, TransZero attains a mapped semantic embedding $\psi(x_i)$.

Loss

Attribute Regression Loss

为了促进 VSEN 准确地将视觉特征映射到其相应的语义空间中，最小化属性 z^c 与 $\psi(x_i^s)$ 之间的均方损失：

$$\mathcal{L}_{AR} = -\frac{1}{n_b}\sum_{i=1}^{nb} \|\psi(x_i^s) - z^c\|_2^2$$

Attribute-Based Cross-Entropy Loss

为了促进嵌入后的视觉特征与对应的类的属性有最高的 compatibility score，计算 $z^{\hat{c}}$ 与 $\psi(x_i^s)$ 间的交叉熵损失：

$$\mathcal{L}_{ACE} = -\frac{1}{n_b}\sum_{i=1}^{nb} \log \frac{\exp(\psi(x_i^s) \times z^c)}{\sum_{\hat{c}\in\mathcal{C}^s} \exp(\psi(x_i^s) \times z^{\hat{c}})}$$

Self-Calibration Loss

$\mathcal{L}_{AR}$ 和 $\mathcal{L}_{ACE}$ 都在优化 seen classes，为了减少对于 seen classes 的过度拟合，引入了一个自校准损失：

$$\mathcal{L}_{SC} = -\frac{1}{n_b}\sum_{i=1}^{nb}\sum_{c'=1}^{\mathcal{C}^u} \log \frac{\exp(\psi(x_i^s) \times z^{c'} + II_{[c'\in\mathcal{C}^u]})}{\sum_{\hat{c}\in\mathcal{C}} \exp(\psi(x_i^s) \times z^{\hat{c}} + II_{[c'\in\mathcal{C}^u]})}$$

其中 $II_{[c'\in\mathcal{C}^u]}$ 是指示函数（即，当 $c \in \mathcal{C}^u$ 时为 1，否则为 -1）。

直观地说，$\mathcal{L}_{SC}$ 鼓励在训练期间将非零概率分配给 unseen classes，这允许 TransZero 在给定来自 unseen classes 的测试样本时为真正的 unseen classes 产生（大）非零概率。

整体损失函数

$$\mathcal{L}_{total} = \mathcal{L}_{ACE} + \lambda_{AR}\mathcal{L}_{AR} + \lambda_{SC}\mathcal{L}_{SC}$$

其中 λ_{AR} 和 λ_{SC} 是控制其相应损失项的权重。

Zero-Shot Prediction

After training TransZero, we first obtain the embedding features of a test instance x_i in the semantic space i.e., $\psi(x_i)$. Then, we take an explicit calibration to predict the test label of x_i, which is formulated as:

$$c^* = \arg \max_{c \in \mathcal{C}^u/\mathcal{C}} \psi(x_i) \times z^c + \mathbb{I}_{[c \in \mathcal{C}^u]}. \quad (18)$$

Here, $\mathcal{C}^u/\mathcal{C}$ corresponds to the CZSL/GZSL setting respectively.

Experiments

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 (Song et al. 2018)	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 (Li et al. 2018)	67.5	26.4	**81.6**	39.9	–	–	–	–	65.5	9.8	87.4	17.6
SGMA* (Zhu et al. 2019)	71.0	36.7	71.3	48.5	–	–	–	–	68.8	37.6	87.1	52.5
AREN* (Xie et al. 2019)	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* (Liu et al. 2019)	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* (Xu et al. 2020)	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												
f-CLSWGAN (Xian et al. 2018)	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-D2 (Xian et al. 2019)	61.0	48.4	60.1	53.6	**64.7**	45.1	38.0	41.3	71.1	57.6	70.6	63.5
OCD-CVAE (Keshari, Singh, and Vatsa 2020)	–	44.8	59.9	51.3	–	44.8	**42.9**	**43.8**	–	59.5	73.4	65.7
E-PGN (Yu et al. 2020)	**72.4**	52.0	61.1	56.2	–	–	–	–	**73.4**	52.6	83.5	64.6
Composer (Huynh and Elhamifar 2020b)	69.4	56.4	63.8	59.9	62.6	**55.1**	22.0	31.4	**71.5**	**62.1**	77.3	**68.8**
GCM-CF (Yue et al. 2021)	–	61.0	59.7	60.3	–	47.9	37.8	42.2	–	60.4	75.1	67.0
FREE (Chen et al. 2021a)	–	55.7	59.9	57.7	–	47.4	37.2	41.7	–	60.4	75.4	67.1
HSVA (Chen et al. 2021b)	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 (Chen et al. 2018)	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 (Li et al. 2019)	–	43.2	51.4	46.9	–	35.1	35.3	35.2	–	31.7	70.9	43.8
IIR (Cacheux, Borgne, and Crucianu 2019)	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 (Jiang et al. 2019)	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 (Min et al. 2020)	–	53.2	60.2	56.5	–	45.0	37.2	40.7	–	**63.6**	70.8	67.0
DAZLE* (Huynh and Elhamifar 2020a)	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 (Ours)	**76.8**	**69.3**	68.3	**68.8**	**65.6**	**52.6**	33.4	40.8	70.1	61.3	82.3	**70.2**

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 visual-semantic decoder.

Method	CUB				SUN			
	acc	U	S	H	acc	U	S	H
TransZero w/o FAE	67.3	61.0	53.1	56.8	61.2	55.7	22.5	32.1
TransZero w/o FA	74.0	66.7	66.3	66.5	63.8	49.5	31.4	38.5
TransZero w/o DEC	62.3	53.3	54.1	53.7	58.3	35.0	28.8	31.6
TransZero w/o $\mathcal{L}_{SC}$	74.8	47.1	75.5	58.1	64.2	42.4	33.4	37.4
TransZero w/o $\mathcal{L}_{AR}$	74.5	65.9	68.8	67.3	64.1	47.2	33.3	39.1
TransZero (full)	**76.8**	69.3	68.3	**68.8**	**65.6**	52.6	33.4	**40.8**

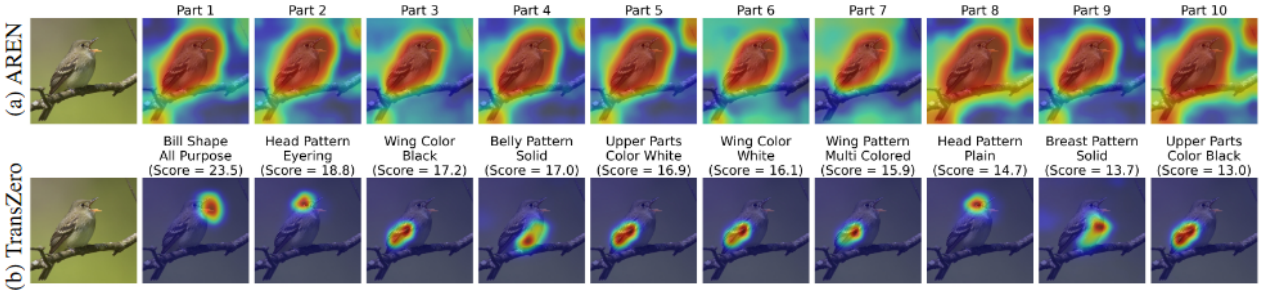


Figure 3: Visualization of attention maps for the attention-based method (i.e, AREN (Xie et al. 2019)) and our TransZero.

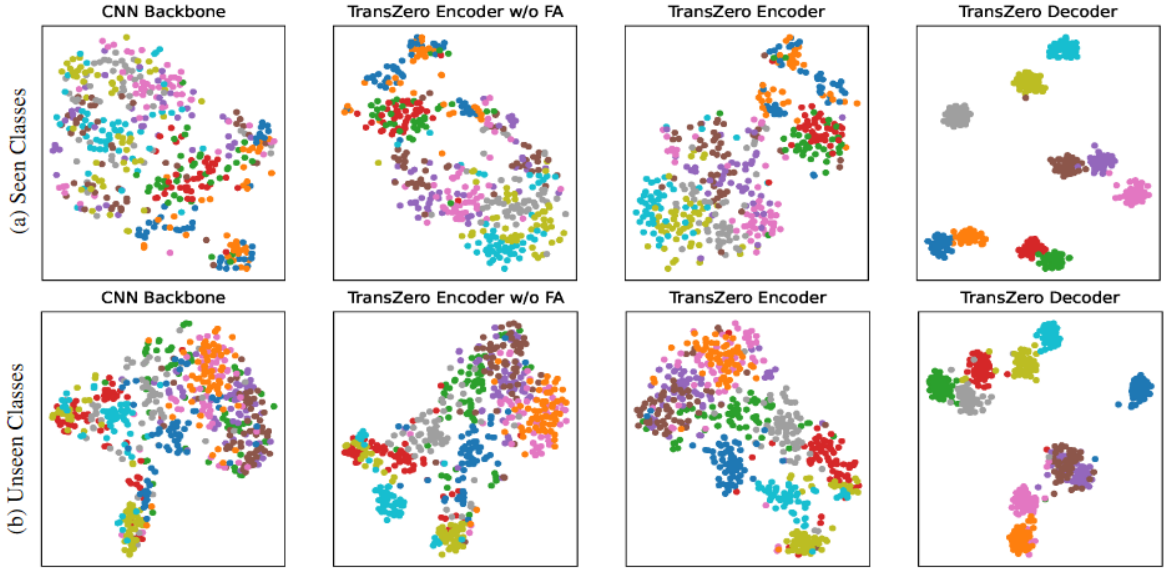
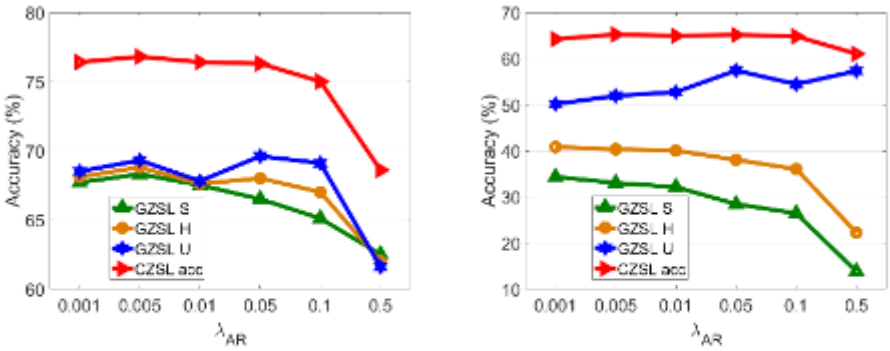
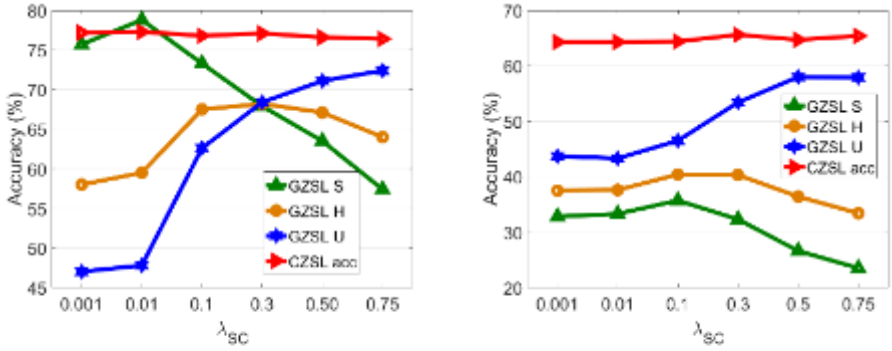


Figure 4: t-SNE visualizations of visual features for (a) seen classes and (b) unseen classes, learned by the CNN backbone, TransZero encoder w/o FA, TransZero encoder, and TransZero decoder. The 10 colors denote 10 different seen/unseen classes randomly selected from CUB.



(a) CUB (b) SUN

Figure 5: The effects of λ_{AR}.



(a) CUB (b) SUN

Figure 6: The effects of λ_{SC}.