Supplementary Materials: CREST: Cross-modal Resonance through Evidential Deep Learning for Enhanced Zero-Shot Learning

Anonymous Authors

1 INTRODUCTION

The content of our supplementary material is organized as follows:

(i.) In Section 2, we present a detailed derivation of the formulas pertaining to the loss (*i.e.* $\mathcal{L}_{ACC}$) associated with Evidential Deep Learning as discussed in the main text.

(ii.) In Section 3, we provide detailed statistics analysis for attributes to illustrate the attribute distribution imbalance and co-occurrence in real world.

(iii.) In Section 4, we provide a brief introduction to each of the 17 compared baselines and conduct a detailed analysis of the competitive ones.

2 THE DERIVATIONS OF $\mathcal{L}_{ACC}$

As defined in the main document, we have:

$$\mathcal{L}_{ACC}\left(\boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{m}}\right) = \mathcal{L}_{ACE}\left(\boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{m}}\right) + \lambda_t \mathcal{L}_{KL}\left(\boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{m}}\right). \tag{1}$$

Next, the derivations of $\mathcal{L}_{ACE}$ and $\mathcal{L}_{KL}$ will be given in detail respectively.

2.1 The Derivations of $\mathcal{L}_{ACE}$

In this section, we provide a detailed derivation of the unimodal adaptive cross-entropy loss, $\mathcal{L}_{ACE}$, employed in our CREST. Considering a parametrized Dirichlet distribution $\boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{m}}$ associated with a unimodal output from Visual Grounding Transformer (VGT) or Attribute Grounding Transformer (AGT), namely $\boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{V}}$ or $\boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{A}}$, the ground truth $\boldsymbol{y}_n$, and the density function $D(\boldsymbol{p}_n \mid \boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{m}})$ referenced in the text, we elucidate the derivation as follows:

$$\begin{aligned}
\mathcal{L}_{ACE}\left(\boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{m}}\right) &= \int \left[\sum_{j=1}^{K} -y_{nj} \log(p_{nj}^{m})\right] \frac{1}{B(\alpha_n^m)} \prod_{j=1}^{K} {p_{nj}^{m}}^{\alpha_{nj}^{m}-1} \, d\boldsymbol{p}_{\boldsymbol{n}}^{\boldsymbol{m}} \\
&= \sum_{j=1}^{K} -y_{nj} \int \log(p_{nj}^{m}) \frac{1}{B(\alpha_n^m)} \prod_{j=1}^{K} {p_{nj}^{m}}^{\alpha_{nj}^{m}-1} \, d\boldsymbol{p}_{\boldsymbol{n}}^{\boldsymbol{m}} \\
&= \sum_{j=1}^{K} -y_{nj} \mathbb{E}[\log p_{nj}^{m}] \\
&= \sum_{j=1}^{K} -y_{nj} (\psi(\alpha_{nj}^{m}) - \psi(S_n^m)) \\
&= \sum_{j=1}^{K} y_{nj} (\psi(S_n^m) - \psi(\alpha_{nj}^{m})),
\end{aligned} \tag{2}$$

where S_n^m represents the sum of Dirichlet parameters with respect to a specific modality for instance n. For the expected value of $\log p_{nj}^{m}$, [14] provides:

$$\begin{aligned}
\mathbb{E}[\log p_{nj}^{m}] &= \frac{\partial}{\partial \alpha_{nj}^{m}} \left(\sum_{j=1}^{K} \log \Gamma\left(\tilde{\alpha}_{nj}^{m}\right) - \log \Gamma\left(\sum_{j=1}^{K} \tilde{\alpha}_{nj}^{m}\right)\right) \\
&= \psi\left(\tilde{\alpha}_{nj}^{m}\right) - \psi\left(\sum_{j=1}^{K} \tilde{\alpha}_{nj}^{m}\right) \\
&= \psi\left(\tilde{\alpha}_{nj}^{m}\right) - \psi\left(S_n^m\right),
\end{aligned} \tag{3}$$

2.2 The derivation of $\mathcal{L}_{KL}$

As defined in the main document, the two density functions defined for calculating the Kullback-Leibler (KL) divergence are as follows:

$$D\left(\boldsymbol{p}_{\boldsymbol{n}}^{\boldsymbol{m}} \mid \tilde{\boldsymbol{\alpha}}_{\boldsymbol{n}}^{\boldsymbol{m}}\right) = \frac{\Gamma\left(\sum_{j=1}^{K} \tilde{\alpha}_{nj}^{m}\right)}{\prod_{j=1}^{K} \Gamma\left(\tilde{\alpha}_{nj}^{m}\right)} \prod_{j=1}^{K} {p_{nj}^{m}}^{\tilde{\alpha}_{nj}^{m}-1}$$

and

$$D\left(\boldsymbol{p}_{\boldsymbol{n}}^{\boldsymbol{m}} \mid \mathbf{1}\right) = \Gamma(K).$$

Consequently, the KL divergence loss, $\mathcal{L}_{KL}$, is derived as:

$$\begin{aligned}
\mathcal{L}_{KL}\left(\boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{m}}\right) &= KL\left[D\left(\boldsymbol{p}_{\boldsymbol{n}}^{\boldsymbol{m}} \mid \tilde{\boldsymbol{\alpha}}_{\boldsymbol{n}}^{\boldsymbol{m}}\right) \| D\left(\boldsymbol{p}_{\boldsymbol{n}}^{\boldsymbol{m}} \mid \mathbf{1}\right)\right] \\
&= \mathbb{E}\left(\log \frac{D\left(\boldsymbol{p}_{\boldsymbol{n}}^{\boldsymbol{m}} \mid \tilde{\boldsymbol{\alpha}}_{\boldsymbol{n}}^{\boldsymbol{m}}\right)}{D\left(\boldsymbol{p}_{\boldsymbol{n}}^{\boldsymbol{m}} \mid \mathbf{1}\right)}\right) \\
&= \mathbb{E}\left(\log\left(\frac{\Gamma\left(\sum_{j=1}^{K} \tilde{\alpha}_{nj}^{m}\right)}{\Gamma(K) \prod_{j=1}^{K} \Gamma\left(\tilde{\alpha}_{nj}^{m}\right)} \prod_{j=1}^{K} {p_{nj}^{m}}^{\tilde{\alpha}_{nj}^{m}-1}\right)\right) \\
&= \log\left(\frac{\Gamma\left(\sum_{j=1}^{K} \tilde{\alpha}_{nj}^{m}\right)}{\Gamma(K) \prod_{j=1}^{K} \Gamma(\tilde{\alpha}_{nj}^{m})}\right) + \mathbb{E}\left[\log\left(\prod_{j=1}^{K} p_{ij}^{\tilde{\alpha}_{nj}^{m}-1}\right)\right] \\
&= \log\left(\frac{\Gamma\left(\sum_{j=1}^{K} \tilde{\alpha}_{nj}^{m}\right)}{\Gamma(K) \prod_{j=1}^{K} \Gamma\left(\tilde{\alpha}_{nj}^{m}\right)}\right) + \sum_{j=1}^{K} (\tilde{\alpha}_{nj}^{m} - 1) \mathbb{E}\left(\log p_{nj}^{m}\right) \\
&= \log\left(\frac{\Gamma\left(\sum_{j=1}^{K} \tilde{\alpha}_{nj}^{m}\right)}{\Gamma(K) \prod_{j=1}^{K} \Gamma\left(\tilde{\alpha}_{nj}^{m}\right)}\right) + \sum_{j=1}^{K} (\tilde{\alpha}_{nj}^{m} - 1) [\psi\left(\tilde{\alpha}_{nj}^{m}\right) \\
&\quad - \psi\left(\sum_{j=1}^{K} \tilde{\alpha}_{nj}^{m}\right)],
\end{aligned} \tag{4}$$

where $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{n}}^{\boldsymbol{m}} = \boldsymbol{y}_n + (1 - \boldsymbol{y}_n) \odot \boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{m}}$ represents the Dirichlet parameters after excluding unreliable evidence from $\boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{m}}$. The functions $\Gamma(\cdot)$ and $\psi(\cdot)$ correspond to the gamma and digamma functions, respectively.

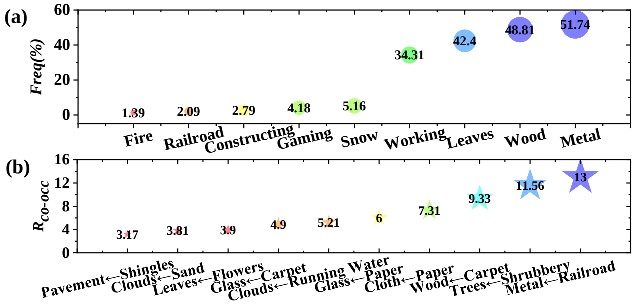

Figure 1: Statistical analysis of attribute distribution imbalance and co-occurrence on SUN and AWA2 benchmarks

3 STATISTICS ANALYSIS OF ATTRIBUTES IN REAL WORLD

From Figure 1, we can observe data pertaining to the co-occurrence of features as well as the imbalanced distribution of features. We define the frequency of occurrence of these features in a manner consistent with the metric established by [7]:

$$Freq(a) = \frac{\sum_{c' \in C} \mathbb{I}\left[a \in \mathcal{A}^{c'}\right]}{|C|} \times 100\% \, , \tag{5}$$

where $\mathbb{I}[a \in \mathcal{A}^{c'}]$ is 1 if $a \in \mathcal{A}^{c'}$, and 0 otherwise.

To quantify the extent of attribute co-occurrence, we establish the following parameter, aligning with [7]:

$$R_{co-occ}\left(a_i \leftarrow a_j\right) = \frac{\sum_{c' \in C} \mathbb{I}\left[a_i \in \mathcal{A}^{c'} \& a_j \in \mathcal{A}^{c'}\right]}{\sum_{c' \in C} \mathbb{I}\left[a_i \notin \mathcal{A}^{c'} \& a_j \in \mathcal{A}^{c'}\right]} . \tag{6}$$

Consider the SUN2012 dataset, where the attributes "hotel" and "room" are observed to co-occur 109 times, while the attribute "hotel" appears independently merely 12 times. Employing the formula introduced earlier, we deduce that $R_{co-occ}(room \leftarrow hotel) = 9.08$, and the frequency of "hotel" is calculated as $Freq(hotel) = 121/1224 \times 100\% \approx 9.89\%$. This outcome underscores the pronounced variance in the distribution of attribute types within the dataset.

4 BASELINES

In our comparative analysis, we select 17 representative or state-of-the-art models from the period of 2020-2023, predominantly introduced at top-tier conferences. A brief introduction as well as a detailed competitive analysis will be given in this section.

4.1 Brief Introduction

The brief introduction of baselines is shown as follows:

TF-VAEGAN [15]: Introduces a feedback mechanism to iteratively refine feature synthesis and ensure semantic consistency across training, generation, and classification stages in zero-shot learning (ZSL).

Composer [10]: A generative model that constructs fine-grained features for unseen classes by compositing attributes from seen classes, increasing feature diversity and specificity.

APN [18]: Enhances ZSL by fusing global and local image features and integrating attribute localization to facilitate knowledge transfer to unseen classes.

DVBE [13]: Addresses biased recognition in generalized ZSL by developing dual visual representations: a semantic-free for seen classes and a semantic-aligned for robust transfer to unseen classes.

DAZLE [11]: A fine-grained generalized ZSL framework that utilizes dense attribute-based attention to align attribute features with semantic vectors, enhancing classification of unseen classes.

RGEN [17]: Incorporates region-based relation reasoning into ZSL to capture local image region relationships, enhancing performance through joint training of attention and reasoning branches.

CE-GZSL [9]: Combines a feature generation model with an embedding model, leveraging class-wise and instance-wise supervision to improve zero-shot classification performance.

GCM-CF [19]: A counterfactual method in generalized ZSL that generates counterfactual samples to balance the classification of seen and unseen classes, enhancing decision boundary accuracy.

FREE [5] : Employs a Feature Refinement module with a self-adaptive margin center loss and semantic cycle-consistency loss to enhance visual features and reduce cross-dataset bias in generalized ZSL.

HSVA [6]: A hierarchical approach that aligns semantic and visual domains through structured and distributional adaptation, significantly outperforming common space learning methods in zero-shot tasks.

AGZSL [8]: Introduces Image-Adaptive Semantics (IAS) and generative meta-learning with virtual classes to mitigate intra-class variations and adapt semantic features for classifying unseen classes in ZSL.

GEM-ZSL [12]: Introduces a gaze estimation module to mimic human gaze for focusing on discriminative attributes in ZSL, optimizing attribute localization and feature representation.

MSDN [3]: Utilizes mutually reinforcing attention sub-networks for distilling intrinsic semantic representations between visual and attribute features, enhancing knowledge transfer to unseen classes.

TransZero [2] : An attribute-guided Transformer network that refines visual features and localizes discriminative attributes, enhancing the learning of visual-semantic interactions for ZSL.

TransZero++ [1]: Employs an attribute→visual Transformer (AVT) and a visual→attribute Transformer (VAT) to learn attribute-based visual features and visual-based attribute features to enhance ZSL with a cross attribute-guided Transformer network, refining visual features and accurately localizing object attributes for robust semantic knowledge representation.

DUET [7]: A ViT-based framework that first integrates pre-trained language models to address attribute imbalances and co-occurrences in ZSL with an end-to-end multi-modal learning paradigm.

DSP [4]: A general method that tackles the visual-semantic domain shift in generative ZSL by evolving semantic prototypes to match real prototypes, enhancing classifier training and performance.

4.2 Competitive Baselines Analysis

As the proposed method is non-generative and involves learning bidirectional cross-modal features, we select methods with similar learning frameworks and competitive capabilities for the subsequent analysis.

TransZero++: This method not only enhances feature extraction through a geometry-aware attention module but also employs AVT and VAT to learn attribute-based visual features and visual-based attribute features, respectively. It uses fine-grained features supplemented with feature-level and prediction-level semantical collaborative losses to achieve synergy in a dual Transformer framework. Nevertheless, this approach neglects the fact that the fine-grained regional feature space does not necessarily align with the label space. Actually, the imbalanced distribution and co-occurrence of attributes often intensify the vision variability at the instance level due to objective conditions. This exacerbation further impacts the alignment between the latent space learned by the model and the label space. Additionally, this approach lacks an straightforward uncertainty quantifying method that directly measures the model's epistemic uncertainty, which limits its interpretability.

DUET: The method introduces a cross-modal semantic grounding network that allows fine-grained semantic mapping between images and textual attributes and undergo classification in an end-to-end manner. Additionally, an attribute-level contrastive learning strategy is employed to effectively address issues of attribute imbalance and co-occurrence, sharpening the model's ability to discriminate subtle visual differences. However, it overlooks the instance-level vision variability and lacks an explicit uncertainty quantifying method that directly reflects the model's epistemic uncertainty. This limitation increases the challenge of recognizing hard negatives in open world.

CREST: The proposed method begins by extracting representations with a bidirectional grounding transformer without geometry enhanced attention module in TransZero++ and employs Evidential Deep Learning (EDL) [16] to directly quantify underlying epistemic uncertainty, thereby enhancing the model's resilience against hard negatives. Compared to TransZero++, this method incorporates dual learning pathways focusing on both visual-category and attribute-category alignments to better ensure a robust correlation between latent and label spaces. In contrast to DUET, it addresses the biases introduced by instance-level vision variability through Visual Instance-level Contrastive Learning (VICL), as well as the critical alignment between feature spaces and label spaces. Moreover, unlike the aforementioned methods, it can self-refine with the proposed uncertainty-informed cross-modal fusion technique, which showcase great robustness and explainability when confronted with hard negative in open world.

REFERENCES

[1] Shiming Chen, Ziming Hong, Wenjin Hou, Guo-Sen Xie, Yibing Song, Jian Zhao, Xinge You, Shuicheng Yan, and Ling Shao. 2022. TransZero++: Cross Attribute-guided Transformer for Zero-Shot Learning. (2022).

[2] Shiming Chen, Ziming Hong, Yang Liu, Guo-Sen Xie, Baigui Sun, Hao Li, Qinmu Peng, Ke Lu, and Xinge You. 2022. TransZero: Attribute-guided Transformer for Zero-Shot Learning. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*.

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

[4] Shiming Chen, Wenjin Hou, Ziming Hong, Xiaohan Ding, Yibing Song, Xinge You, Tongliang Liu, and Kun Zhang. 2023. Evolving semantic prototype improves generative zero-shot learning. In *International Conference on Machine Learning*. PMLR, 4611–4622.

[5] Shiming Chen, Wenjie Wang, Beihao Xia, Qinmu Peng, Xinge You, Feng Zheng, and Ling Shao. 2021. Free: Feature refinement for generalized zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 122–131.

[6] Shiming Chen, Guosen Xie, Yang Liu, Qinmu Peng, Baigui Sun, Hao Li, Xinge You, and Ling Shao. 2021. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. *Advances in Neural Information Processing Systems* 34 (2021), 16622–16634.

[7] Zhuo Chen, Yufeng Huang, Jiaoyan Chen, Yuxia Geng, Wen Zhang, Yin Fang, Jeff Z. Pan, and Huajun Chen. 2023. DUET: Cross-Modal Semantic Grounding for Contrastive Zero-Shot Learning. In *AAAI*. AAAI Press, 405–413.

[8] Yu-Ying Chou, Hsuan-Tien Lin, and Tyng-Luh Liu. 2021. Adaptive and Generative Zero-Shot Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=ahAUv8TI2Mz

[9] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. 2021. Contrastive embedding for generalized zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2371–2381.

[10] Dat Huynh and Ehsan Elhamifar. 2020. Compositional zero-shot learning via fine-grained dense feature composition. *Advances in Neural Information Processing Systems* 33 (2020), 19849–19860.

[11] Dat Huynh and Ehsan Elhamifar. 2020. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4483–4493.

[12] Yang Liu, Lei Zhou, Xiao Bai, Yifei Huang, Lin Gu, Jun Zhou, and Tatsuya Harada. 2021. Goal-oriented gaze estimation for zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3794–3803.

[13] Shaobo Min, Hantao Yao, Hongtao Xie, Chaoqun Wang, Zheng-Jun Zha, and Yongdong Zhang. 2020. Domain-aware Visual Bias Eliminating for Generalized Zero-Shot Learning. arXiv:2003.13261 [cs.CV]

[14] Thomas Minka. 2000. Estimating a Dirichlet distribution.

[15] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. 2020. Latent embedding feedback and discriminative features for zero-shot classification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*. Springer, 479–495.

[16] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential Deep Learning to Quantify Classification Uncertainty. arXiv:1806.01768 [cs.LG]

[17] Guo-Sen Xie, Li Liu, Fan Zhu, Fang Zhao, Zheng Zhang, Yazhou Yao, Jie Qin, and Ling Shao. 2020. Region graph embedding network for zero-shot learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 562–580.

[18] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. 2020. Attribute Prototype Network for Zero-Shot Learning. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 21969–21980. https://proceedings.neurips.cc/paper_files/paper/2020/file/fa2431bf9d65058fe34e9713e32d60e6-Paper.pdf

[19] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. 2021. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15404–15414.