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

Haojian Huang
TeleAI
China
The University of Hong Kong
Hong Kong, China
haojianhuang927@gmail.com

Xiaozhen Qiao
University of Science and Technology of China
Hefei, Anhui, China
TeleAI
China
XiaozhennnQiao@gmail.com

Zhuo Chen
Zhejiang University
Hangzhou, Zhejiang, China
zhuo.chen@zju.edu.cn

Haodong Chen
Northwestern Polytechnical University
Xi'an, Shaanxi, China
chd@mail.nwpu.edu.cn

Bingyu Li
University of Science and Technology of China
Hefei, Anhui, China
TeleAI
China
libingyu0205@163.com

Zhe Sun
Northwestern Polytechnical University
Xi'an, Shaanxi, China
TeleAI
China
sunzhe@nwpu.edu.cn

Mulin Chen*
Northwestern Polytechnical University
Xi'an, Shaanxi, China
TeleAI
China
chenmulin001@gmail.com

Xuelong Li*
TeleAI
China
xuelong_li@ieee.org

Abstract

Zero-shot learning (ZSL) enables the recognition of novel classes by leveraging semantic knowledge transfer from known to unknown categories. This knowledge, typically encapsulated in attribute descriptions, aids in identifying class-specific visual features, thus facilitating visual-semantic alignment and improving ZSL performance. However, real-world challenges such as distribution imbalances and attribute co-occurrence among instances often hinder the discernment of local variances in images, a problem exacerbated by the scarcity of fine-grained, region-specific attribute annotations. Moreover, the variability in visual presentation within categories can also skew attribute-category associations. In response, we propose a bidirectional cross-modal ZSL approach **CREST**. It begins by extracting representations for attribute and visual localization and employs Evidential Deep Learning (EDL) to measure underlying epistemic uncertainty, thereby enhancing the model's resilience against hard negatives. CREST incorporates dual learning pathways, focusing on both visual-category and attribute-category alignments, to ensure robust correlation between latent and observable spaces. Moreover, we introduce an uncertainty-informed cross-modal fusion technique to refine visual-attribute inference. Extensive experiments demonstrate our model's effectiveness and unique explainability across multiple datasets. Our code and data are available at: `TeleAI CREST`.

CCS Concepts

• **Computing methodologies** → **Learning paradigms**.

Keywords

Zero-Shot Learning, Multimodality, Evidential Deep Learning, Contrastive Learning

ACM Reference Format:
Haojian Huang, Xiaozhen Qiao, Zhuo Chen, Haodong Chen, Bingyu Li, Zhe Sun, Mulin Chen, and Xuelong Li. 2024. CREST: Cross-modal Resonance through Evidential Deep Learning for Enhanced Zero-Shot Learning. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28–November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3681629

*corresponding author.

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia.

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3681629

1 Introduction

Humans frequently possess the talent to grasp novel concepts relying on prior experience without the need to see them beforehand.

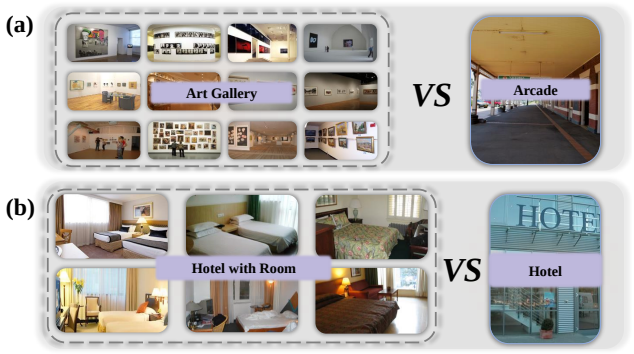

Figure 1: Challenges in instance-level recognition in the real world: (a) Attributes distribution imbalances—significant frequency differences among attributes; (b) Attributes co-occurrence—tendency of certain attributes to appear together, influencing model bias (further statistical details are available in the Supplementary Material).

For instance, a peacock is commonly known as a bird with a colorful fan-shaped tail; if individuals have previous knowledge of birds and fans, they can quickly identify a peacock. However, unlike humans, widely used and studied supervised deep learning models are typically limited to classifying samples belonging to categories seen during training, lacking the capacity to handle samples from unseen categories during training, thus lacking generality and flexibility. Therefore, to further advance Artificial General Intelligence (AGI) [4] and achieve true implementation, Zero-Shot Learning (ZSL) was introduced to identify new classes by leveraging inherent semantic relationships during learning [23, 37–39, 48]. It is already extensively applied in tasks with broad real-world applications, *e.g.*, image classification [22, 67], semantic segmentation [5, 21], video understanding [6, 70, 75], 3D generation [7, 33, 71], *etc.*, which also contributes significantly to the robust development of Large Language Models (LLMs) [35, 62] and Embodied AI [30, 60].

In ZSL, attributes stand as key semantic descriptors for visual features of images, representing a widely embraced form of annotation. Unfortunately, the attribute annotations are more often categorical rather than regional [17]. Dense attention interactions do not guarantee that models directly grasp the correspondence between local visual-semantic information and categories, nor do they alleviate the model's epistemic uncertainty when confronted with unseen categories [55]. That is because the skewed distribution of attributes in the real world, as well as the issues arising from attribute co-occurrence shown in Figure 1.

Existing methods overlook the importance of aligning regional features with categories. Models may link specific attributes, like a red bird's bill, to "bill color red" but struggle to deduce the bird's species. This challenge is compounded as attributes across species are often intertwined. Furthermore, real-world images of the same category vary significantly due to factors like camera angles, background, distances, lights and the motions, making it difficult for dense attention to learn hard category-matching patterns. This can increase epistemic uncertainty when merging features for inference, potentially exacerbating modal conflicts and impairing model performance [69].

To this end, we integrate Evidential Deep Learning (EDL) [55] into ZSL for the first time, leading to a novel framework, named Cross-modal Resonance through Evidential Deep Learning for Enhanced Zero-ShoT Learning, termed as CREST. Specifically, we employ the Visual Grounding Transformer (VGT) and the Attribute Grounding Transformer (AGT) to extract bidirectional, region-level features from images and attributes. Unlike conventional approaches that simply adjust distances within the representation space based on category [17], our strategy addresses vision variability and feature-category alignment directly. We first introduce instance-level contrastive learning for adaptive vision alignment and employ a technique similar to non-maximum suppression to reduce attribute overlap between categories, facilitating deeper attribute-category insights. To counteract the potential degradation from hard-negative samples [52], we apply EDL for epistemic uncertainty measure and develop an uncertainty-driven fusion method [27, 28, 42, 69]. This enhances the model's generalization in downstream tasks by merging semantic knowledge across representation spaces. To summarize, our contributions are as follows:

- We propose CREST, a novel ZSL framework that considers bidirectional cross-modal representations of attributes and visual features. Moreover, it leverages dual learning pathways, focusing on both visual-category and attribute-category alignments, learning implicit matching patterns between features and categories from fine-grained visual elements and attribute texts.
- To the best of our knowledge, CREST is the first in ZSL to apply EDL for measuring epistemic uncertainty and mitigating potential conflicts in cross-modal fusion.
- Extensive experiments show that CREST performs competitively on three well-known ZSL benchmarks, *i.e.*, CUB [63], SUN [50], and AWA2 [64]. Comprehensive ablations and analyses further validate the effectiveness and explainability of our approach.

2 Related Work

2.1 Zero-shot learning

ZSL can be classified into two main categories based on the classes encountered during the testing phase: Conventional ZSL (CZSL) and Generalized ZSL (GZSL), where CZSL is designed to predict classes that have not been seen during training, whereas GZSL extends its predictive capability to both seen and unseen classes [8, 64]. The core concept of ZSL revolves around learning discriminative and transferable visual features based on semantic information, *e.g.*, attribute descriptions [38], sentence embeddings [53], and DNA [2], enabling effective visual-semantic interactions. Among these, attributes stand out as the most commonly used semantic information within ZSL. Early research focused on harnessing visual-semantic interactions to transfer knowledge to unseen categories [1, 41, 59]. These initial attempts, particularly through embedding-based methods, entailed learning a mapping between seen categories and their corresponding semantic vectors, followed by employing nearest neighbor searches within the embedding space to classify unseen categories [72, 74]. Since they primarily rely on seen category samples, the effectiveness was significantly limited due to a bias towards

these categories, exacerbating the challenge in GZSL. Novel regularization and space modification strategies have been developed to improve ZSL model generalization [40, 49, 58]. Generative models, including VAEs [14, 15, 54, 61], GANs [13, 25, 65, 67], and generative flows [57, 78], synthetically enhance feature spaces with unseen class characteristics. These methods, aiming to bridge the domain gap, reframes ZSL as a supervised task by providing a means to compensate for the lack of unseen class data. Despite progress, these methods often neglect localized visual cues in favor of global information, overlooking the nuanced, fine-grained attributes essential for dissecting complex semantic categories [9, 18]. This oversight weakens the visual representations obtained, diminishing the efficacy of the visual-semantic knowledge transfer. Subsequently, intricate attentions are integrated into ZSL to prioritize salient features and attributes, improving model discernment [16, 20, 32, 36, 43, 47, 79]. And Recent studies have started experimenting with the deployment of intricate attention to engage with region-level visual-attribute features [9–11, 17]. These methods highlight distinctive, fine-grained features, evolving towards complex attention modules for deeper semantic understanding. However, due to instance-level visual variability and inter-class attribute coupling, fine-grained representations may not guarantee accurate feature-to-category matching. This paper delves into aligning latent feature and category spaces.

2.2 Evidential Deep Learning for Classification

EDL enhances machine learning by enabling models to quantify uncertainty, thus bolstering reliability and interpretability. Grounded in subjective logic principles [34], EDL has emerged as a response to the challenges of model confidence and uncertainty, as highlighted in neural network calibration issues by [24]. The framework's utility was further solidified by [55], which introduced a method to quantify classification uncertainty, significantly increasing deep learning model trustworthiness. The adaptability of EDL to various data contexts has been demonstrated through applications like open set action recognition [3], signifying its efficacy in handling new and unseen data types. The scope of EDL further expanded to multi-view classification [27], showcasing its ability to integrate and reason with information from multiple sources. This integration was further enhanced by introducing dynamic evidential fusion [28], highlighting EDL's adaptability in complex data environments.

Recent advancements, such as adaptive EDL for semi-supervised learning [76] and its application in multimodal decision-making [56], have marked EDL's progression towards addressing real-world data challenges. Additionally, [69] illustrates EDL's potential in conflictive multi-view learning scenarios, reinforcing its capacity to support reliable decision-making across diverse applications. In ZSL, there exists epistemic uncertainty in the gap between region-level fine-grained latent space and category space. Moreover, dual-stream visual-attribute interactions do not necessarily align representation spaces. Therefore, we apply EDL to assess feature-category alignment uncertainties independently and introduces an uncertainty-driven fusion framework for coherent visual-attribute inference.

3 Methodology

3.1 Problem Definition

ZSL equips models to recognize targets in unseen categories. The training set, $D^s = \{(x^s, y^s)|x^s \in \mathcal{X}^s, y^s \in \mathcal{Y}^s\}$, consists of samples from known categories, with x^s as images labeled y^s. The set $D^u = \{(x^u, y^u)|x^u \in \mathcal{X}^u, y^u \in \mathcal{Y}^u\}$ captures samples from new categories. With $\mathcal{Y}^u$ and $\mathcal{Y}^s$ distinct, each y aligns with a category $c \in C = C^s \cup C^u$. This framework leverages attribute information from C^s for knowledge transfer to C^u. Assuming predefined attributes for each category, quantified as either continuous or binary values, the dataset's attribute space is $\mathcal{A} = \{a_1, \ldots, a_{|\mathcal{A}|}\}$. Each category's attribute profile, c, is depicted by $z^c = [z^c_1, \ldots, z^c_{|\mathcal{A}|}]^\top$, reflecting the value of each associated attribute.

3.2 Cross-modal Feature Extraction

Feature Extraction: Attributes and Vision. We extract textual features using the pre-trained GloVe model[51], while employing ResNet-101[29] as the CNN backbone to distill visual features from images (as depicted in Figure 2(a)(b)). These features support the development of a bidirectional grounding Transformer.

Bidirectional Grounding Transformer. In the decoding phase, the VGT and AGT refine visual and semantic attributes, respectively. The VGT attend semantic features to localize relevant image regions, whereas the AGT interprets semantic information through regional visual features. Both decoders employ a streamlined cross-attention module, with the encoder output U serving as keys K and values V, and semantic embeddings as queries Q. This methodology establishes a bidirectional link between images and attributes, enhancing the recognition of unseen categories. The process is concisely described as follows:

$$K = UW_k, \quad V = UW_v, \quad Q = VW_q,$$
$$\hat{F} = \text{SoftMax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{1}$$

where W_q, W_k, W_v are the learnable weights in cross attention, d_k represents the dimension of the features. After n layers of iteration, the output $\hat{F}$ is transformed by a FeedForward layer:

$$F^V = \text{ReLU}\left(\left(\hat{F}W_1 + b_1\right)W_2 + b_2\right). \tag{2}$$

The AGT structure is analogous to the VGT, differing only in the modalities employed as queries in the cross-attention modules. Overall, the features of attribute and visual localization F^A, F^V can be respectively captured through the application of AGT and VGT.

3.3 Visual Instance-level Contrastive Learning

Generally speaking, existing methods achieve implicit alignment with the categorical space by mapping latent semantic matches in text to relevant visual regions in images, subsequently employing fine-grained embeddings. However, in the real world, the images captured often exhibit visual variability due to factors such as angles, backgrounds, distances, illumination, and motion (as shown in Figure 3). This variability significantly diminishes the practical effectiveness of textual semantics since the fine-grained visual representations derived from text may not necessarily correspond to the typical categories intended. Conversely, subjects from different

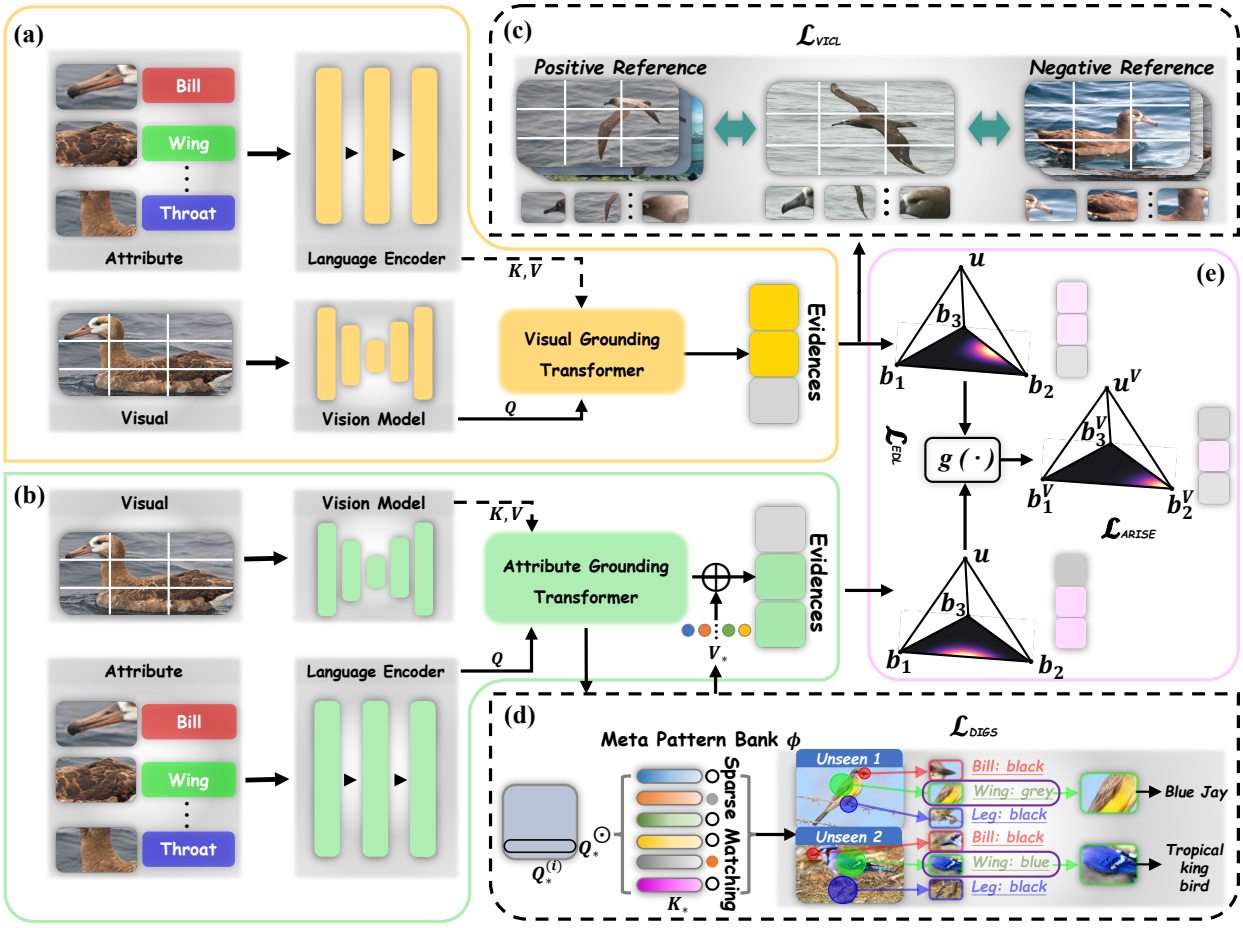

Figure 2: The CREST model's architecture is depicted in Figure 2, initiating with modules (a) and (b) that perform bidirectional grounding to localize features within visuals and attributes. Following this, modules (c) and (d) engage in dual learning to align visual-category and attribute-category in the latent space. The process concludes with an uncertainty-driven fusion module (e), which integrates bidirectional evidence to enable robust visual-attribute inference.

Figure 3: The Birds of an identical category (*i.e.* black-footed albatross) captured in varying angles, backgrounds, distances, illumination, motions, *etc.* illustrating the dynamic nature of vision variability.

categories might appear visually similar due to these influencing factors. To foster proximity among similar entities and distance among distinct categories in the representational space, some approaches might consider employing intra-group category labels for supervision. This method, however, could yield suboptimal solutions due to the vision variability present in an open-world scenario.

To this end, we propose the Visual Instance-level Contrastive Learning (VICL) to mitigate the gap between fine-grained visual latent space and intra-category space.

$$\mathcal{L}_{\text{VICL}} = \mathbb{E}_{x\sim\mathcal{X}^s}\left[-\log f_\theta(\tilde{v} \mid s, x)\right], \tag{3}$$

$$f_\theta = \frac{\exp\left(D\left(\tilde{v}, \tilde{v}^+\right)/\tau\right)}{\exp\left(D\left(\tilde{v}, \tilde{v}^+\right)/\tau\right) + \sum_{\tilde{v}^-\in\mathcal{N}(\tilde{v})}\exp\left(D\left(\tilde{v}, \tilde{v}^-\right)/\tau\right)}, \tag{4}$$

where s, $\tilde{v}$, $\tilde{v}^+$ and $\tilde{v}^-$ represent the input sentences from language side, a candidate positive sample, its positive sample and negative sample respectively. And we adopt a strategy that adjusts for intra-category visual variability through a similarity-based selection of positive samples. Given a batch, the similarity score $D(\tilde{v}, \tilde{v}^+)$ is computed. If no intra-category sample resemble the candidate, we then identify the similar samples across the batch to serve as the positive samples, irrespective of category, based on the similarity score. This approach enables the model to maintain category coherence despite visual discrepancies.

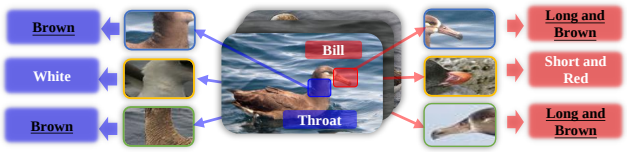

Figure 4: Illustration of attribute coupling across bird species, highlighting shared and divergent traits.

3.4 Decoupled Insight for Grounding Semantics

Traditional methods typically align attribute features with visual features in a straightforward manner to achieve recognition outcomes. However, as illustrated in Figure 4, where attributes coupling across categories, posing challenges to accurate identification. As shown in Figure 2(d), similar visual regions can share the same attributes and intensify the challenges. Hence, we propose Decoupled Insight for Grounding Semantics (DIGS) loss and leverage a Meta-Pattern Bank to develop an auxiliary sparse attention module $\Phi \in \mathbb{R}^{\phi \times d}$, where ϕ and d ($d < |\mathcal{A}|$) respectively represents the total number of memory pattern vectors and their dimensional attributes.

$$\begin{aligned} Q^{(i)} &= F_i^A W_Q + b_Q, \\ a_j^{(i)} &= \frac{\exp(Q^{(i)}\Phi[j]^\top)}{\sum_{k=1}^{\phi}\exp(Q^{(i)}\Phi[k]^\top)}, \\ V_*^{(i)} &= \sum_{j=1}^{\phi} a_j^{(i)}\Phi[j]. \end{aligned} \tag{5}$$

Specifically, in a batch with N samples, our model uses the AGT to map the h-dimensional features $F^A \in \mathbb{R}^{N\times h}$ to the queries $Q \in \mathbb{R}^{N\times d}$ in the latent space of a meta pattern bank with $W_Q \in \mathbb{R}^{h\times d}$ and bias $b_Q \in \mathbb{R}^{h\times d}$. These queries compute similarity scores with pattern vectors Φ via dot products. Equation 5 delineates the transformation where $Q^{(i)} = F_i^A W_Q + b_Q$ generates the attention score $a_j^{(i)}$ that leads to the sparse attention-weighted feature vector $V_*^{(i)}$. This vector is subsequently remapped to the latent space of F^A, and directly added to it, enhancing the feature set by integrating the weighted information from the latent space. To decouple the attribute-category mapping in this latent space, we embrace the DIGS loss inspired by non-maximum suppression (NMS). It operates on two fronts:

(i). The triplet loss component incentivizes the distinction between the closest and second-closest memory pattern vectors. Let $Q^{(i)}$ be the query representation for the i-th example, $\Phi[p]$ the most similar memory pattern (positive sample), and $\Phi[n]$ the second most similar memory pattern (negative sample). The triplet loss is then defined as:

$$\mathcal{L}_{\text{tp}} = \sum_{i=1}^{N} \max\left(\left\|Q^{(i)} - \Phi[p]\right\|^2 - \left\|Q^{(i)} - \Phi[n]\right\|^2 + \lambda, 0\right), \tag{6}$$

where λ is a margin enforcing that the similarity between $Q^{(i)}$ and $\Phi[p]$ exceeds that between $Q^{(i)}$ and $\Phi[n]$ by at least λ, encouraging the model to focus on positive samples and hard negatives and pull positive samples closer to the anchor $Q^{(i)}$ than negative ones.

(ii). The regularization term promotes compact clustering of patterns by minimizing the distance between each query and its most similar memory pattern. This is quantified as:

$$\mathcal{L}_{\text{reg}} = \sum_{i=1}^{N}\left\|Q^{(i)} - \Phi[p]\right\|^2, \tag{7}$$

By synthesizing these components, the DIGS loss is articulated as $\mathcal{L}_{\text{DIGS}} = \mathcal{L}_{\text{tp}} + \mathcal{L}_{\text{reg}}$. Hence, it ensures that the memory patterns not only cluster tightly but also maintain separation, enabling the model to discern and generalize known patterns effectively while grasping the relational structure of the prototypes.

3.5 Evidential deep learning

Given two opinions on the same instance, $\omega_A = (b^A, u^A, a^A)$ and $\omega_B = (b^B, u^B, a^B)$, their synthesis $\omega^{A\oplus B}$ combines their beliefs, uncertainty, and evidence as follows:

$$b_k^{A\oplus B} = \frac{b_k^A u^B + b_k^B u^A}{u^A + u^B}, \quad u^{A\oplus B} = \frac{2u^A u^B}{u^A + u^B}, \quad a_k^{A\oplus B} = \frac{a_k^A + a_k^B}{2},$$

where a^A, a^B represent two different base distribution(*e.g.* Uniform distribution). The conflict degree $c(\omega^A, \omega^B)$ assesses the divergence and shared certainty between ω^A and ω^B:

$$c(\omega^A, \omega^B) = c_p(\omega^A, \omega^B) \cdot c_c(\omega^A, \omega^B), \tag{8}$$

$$c_p(\omega^A, \omega^B) = \frac{\sum_{k=1}^{K}|p_k^A - p_k^B|}{2}, \tag{9}$$

$$c_c(\omega^A, \omega^B) = (1 - u^A)(1 - u^B), \tag{10}$$

where p represent the linear projected probability distributions of the opinions by Dirichlet parameters (*i.e.* b and u). This framework facilitates a nuanced analysis of agreement and discord between the opinions.

As illustrated in Figure(2), we treat the outputs of VGT and AGT as evidence vectors, which typically involve issues of ambiguous recognition. Employing EDL allows us to precisely quantify these uncertainties, thereby deriving accurate recognition results. For each instance $\{\mathbf{x}_n^m\}_{m=1}^M$, the modality count M encapsulates two modalities in our bidirectional grounding Transformer, namely visual-to-attribute and attribute-to-visual. The network computes Dirichlet distribution parameters $\boldsymbol{\alpha}_n^m = \boldsymbol{e}_n^m + \mathbf{1}$, where $\mathbf{e}_n^m = f_\theta^m(\mathbf{x}_n^m)$ is the predicted evidence vector, with f_θ^m denoting the modality-specific transformation function. The uncertainty mass derived as $u_n^m = \frac{K}{\sum_{k=1}^{K}(\alpha_{k,n}^m)}$, where $K = |C|$. Adapting to unimodal evidence-based classification, the traditional cross-entropy loss is intricately tailored for compatibility with this framework:

$$\begin{aligned} \mathcal{L}_{ACE}\left(\boldsymbol{\alpha}_n^m\right) &= \int\left[\sum_{j=1}^{K} -y_{nj}\log p_{nj}^m\right]\frac{\prod_{j=1}^{K} {p_{nj}^m}^{\alpha_{nj}^m - 1}}{B\left(\boldsymbol{\alpha}_n^m\right)} d\mathbf{p}_\mathbf{n}^\mathbf{m}, \\ &= \sum_{j=1}^{K} y_{nj}\left(\psi\left(S_n^m\right) - \psi\left(\alpha_{nj}^m\right)\right), \end{aligned} \tag{11}$$

where $\mathcal{L}_{ACE}(\boldsymbol{\alpha}_n^m)$ denotes the unimodal adaptive cross-entropy loss for the parameters $\boldsymbol{\alpha}_n^m$ of the Dirichlet distribution for a single instance n. Utilizing the digamma function ψ, the integral is simplified to the expectation of the logarithm of predicted probabilities, where S_n^m represents the sum of Dirichlet parameters for instance n, reflecting the total evidence across all classes. The objective of this adaptive loss function is to adjust the network's output parameters

to accurately represent the inherent uncertainty in predictions, enabling the network to make confident predictions when evidence is ample and maintain a degree of uncertainty when evidence is scarce.

Nevertheless, the aforementioned loss function fails to address the issue of insufficient evidence caused by incorrect labels. Therefore, we incorporate a Kullback-Leibler (KL) divergence term into the loss function.

$$\begin{aligned}\mathcal{L}_{KL}\left(\boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{m}}\right) &= KL\left[D\left(\boldsymbol{p}_{\boldsymbol{n}}^{\boldsymbol{m}} \mid \tilde{\boldsymbol{\alpha}}_{\boldsymbol{n}}^{\boldsymbol{m}}\right) \| D\left(\boldsymbol{p}_{n}^{m} \mid \mathbf{1}\right)\right] \\ &= \log\left(\frac{\Gamma\left(\sum_{k=1}^{K} \tilde{\alpha}_{nk}^{m}\right)}{\Gamma(K)\prod_{k=1}^{K}\Gamma\left(\tilde{\alpha}_{nk}^{m}\right)}\right), \\ &+ \sum_{k=1}^{K}\left(\tilde{\alpha}_{nk}^{m}-1\right)\left[\psi\left(\tilde{\alpha}_{nk}^{m}\right)-\psi\left(\sum_{j=1}^{K}\tilde{\alpha}_{nj}^{m}\right)\right],\end{aligned} \tag{12}$$

where $D\left(\boldsymbol{p}_{\boldsymbol{n}}^{\boldsymbol{m}} \mid \mathbf{1}\right)$ represents the uniform Dirichlet distribution, $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{n}}^{\boldsymbol{m}} = \mathbf{y_n} + (\mathbf{1} - \mathbf{y_n}) \odot \boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{m}}$ denotes the Dirichlet parameters after excluding non-misleading evidence from the predicted parameters $\boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{m}}$ for the n-th instance, and $\Gamma(\cdot)$ signifies the gamma function.

Hence, for the n-th instance in the single-modality setting with Dirichlet distribution parameter $\boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{m}}$, the loss is computed as follows:

$$\mathcal{L}_{ACC}\left(\boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{m}}\right) = \mathcal{L}_{ACE}\left(\boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{m}}\right) + \lambda_t \mathcal{L}_{KL}\left(\boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{m}}\right), \tag{13}$$

Where $\lambda_t = \min(1.0, t/\mathcal{E}) \in [0, 1]$ denotes the annealing coefficient, with t being the index of the current training epoch and $\mathcal{E}$ representing the annealing steps. Gradually increasing the influence of KL divergence in the loss function prevents premature convergence of misclassified instances to a uniform distribution.

To ensure consistency across differing perspectives during training, a method to minimize the degree of opinion conflict is employed. The consistency loss for instance $\left\{\mathbf{x}_n^m\right\}_{m=1}^{M}$ is calculated as follows:

$$\mathcal{L}_{CON} = \frac{1}{M-1}\sum_{p=1}^{M}\left(\sum_{q\neq p}^{M} c\left(\boldsymbol{\omega}_n^p, \boldsymbol{\omega}_n^q\right)\right). \tag{14}$$

In the processes of VGT and AGT, mismatches may arise, linking attribute features to incorrect visual parts, or the reverse. The parameter c serves to measure the conflict level between two opinions, where $c = 0$ denotes a lack of conflict and $c = 1$ denotes direct opposition. For the specific instance $\left\{\mathbf{x}_n^m\right\}_{m=1}^{M}$, the overall EDL loss functions can be given as follows:

$$\mathcal{L}_{EDL} = \mathcal{L}_{ACC}\left(\hat{\boldsymbol{\alpha}}_{\boldsymbol{n}}\right) + \beta\sum_{m=1}^{M}\mathcal{L}_{ACC}\left(\boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{m}}\right) + \gamma\mathcal{L}_{CON}. \tag{15}$$

where $\hat{\boldsymbol{\alpha}}_{\boldsymbol{n}}$ shaped by the fusion of modalities driven by uncertainty u_n^m (*e.g.*, the uncertainty-weighted average of modalities' $\boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{m}}$) calibrates the EDL loss relative to the observed conflict degree.

3.6 Model training and optimization strategies

Attribute Reinforced Semantic Integration. We introduce an Attribute ReInforced SEmantic Integration (ARISE) to improve model discrimination by embedding attribute information into the loss function, enhancing classification. By featuring a self-calibrating component, it mitigates overfitting and promotes attribute generalization, regulated by a balancing coefficient λ_{CAL}. Given a batch of n_b training images $\{x_i\}_{i=1}^{n_b}$ with their corresponding class semantic vectors z^c, $\mathcal{L}_{ARISE}$ can be formally represented

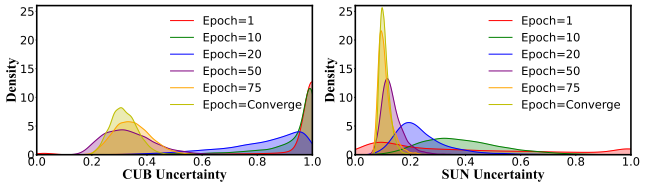

Figure 5: Evolution of model uncertainty on CUB and SUN datasets with increasing epochs, showing a shift towards lower uncertainty as the model converges.

as follows:

$$\begin{aligned}\mathcal{L}_{ARISE} = -\frac{1}{n_b}\sum_{i=1}^{n_b}\Bigg[&\log\frac{\exp\left(f\left(x_i\right)\cdot z^{c}\right)}{\sum_{\hat{c}\in C^{s}}\exp\left(f\left(x_i\right)\cdot z^{\hat{c}}\right)} \\ &-\lambda_{CAL}\sum_{c'=1}^{C^{u}}\log\frac{\exp\left(f\left(x_i\right)\cdot z^{c'}+\mathbb{I}_{[c'\in C^{u}]}\right)}{\sum_{\hat{c}\in C}\exp\left(f\left(x_i\right)\cdot z^{\hat{c}}+\mathbb{I}_{[\hat{c}\in C^{u}]}\right)}\Bigg]\end{aligned} \tag{16}$$

where $f\left(x_i\right) = \mu\boldsymbol{\alpha}_i^A + (1-\mu)\boldsymbol{\alpha}_i^V$ with a blanced coefficient μ. $\mathcal{L}_{ARISE}$ aims to minimize the discrepancy between the predicted and true distributions, taking into account the attribute similarities between categories, serving as a regularization term that encourages the model to learn generalizable features across different categories. Therefore, the overall loss can be obtained as follows:

$$\mathcal{L} = \mathcal{L}_{ARISE} + \mathcal{L}_{VICL} + \mathcal{L}_{DIGS} + \lambda_{EDL}\mathcal{L}_{EDL} \tag{17}$$

3.7 Zero-Shot Inference

Upon completing the training of CREST, we extract the visual embeddings of a test sample x_i in the semantic space relative to VGT and AGT, denoted as $\boldsymbol{\alpha}_i^V$ and $\boldsymbol{\alpha}_i^A$. Given that the semantic-augmented visual embeddings from VGT and AGT offer complementary information, we integrate their predictions through combination coefficients μ for a calibrated test label prediction of x_i, expressed as:

$$c^{*} = \arg\max_{c\in C^{u}/C}\left(\mu\boldsymbol{\alpha}_i^A + (1-\mu)\boldsymbol{\alpha}_i^V\right)^{\top}\cdot z^{c} + \mathbb{I}_{[c\in C^{u}]} \tag{18}$$

In this formula, C^u/C pertains to the CZSL/GZSL scenarios, respectively.

4 Experiment

Dataset. Our study investigates three principal zero-shot learning (ZSL) benchmarks: two fine-grained datasets, CUB [63] and SUN [50], and one coarse-grained dataset, AWA2 [64]. CUB encompasses 11,788 images across 200 bird classes (150 seen, 50 unseen), featuring 312 attributes. SUN includes 14,340 images spanning 717 scene categories (645 seen, 72 unseen) with 102 attributes. AWA2 contains 37,322 images of 50 animal classes (40 seen, 10 unseen), each described by 85 attributes.

Evaluation Protocols. Following Xian et al.'s framework [66], we evaluated the top-1 accuracy in both CZSL and GZSL setups. In CZSL, accuracy is assessed solely by predicting unseen classes. For GZSL, we compute the accuracy for both seen (S) and unseen (U) classes and employ their harmonic mean (defined as $H = (2 \times S \times U)/(S + U)$) as the evaluative metric.

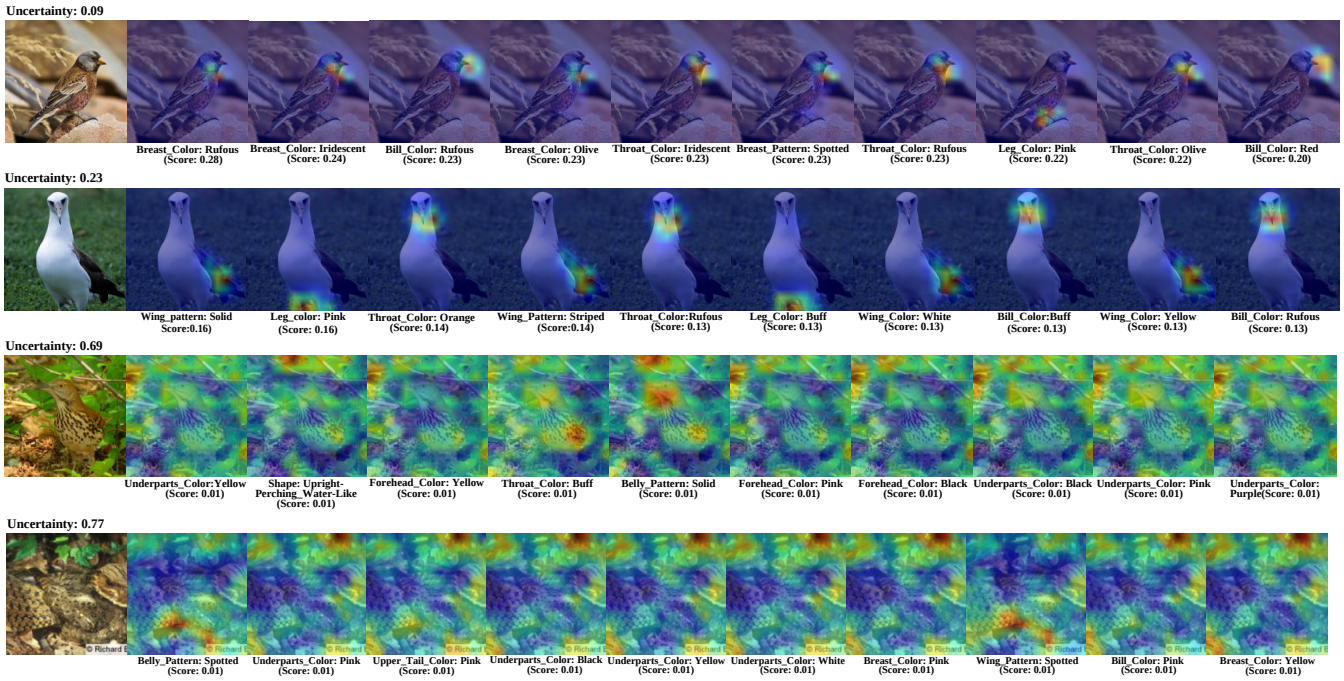

Figure 6: Visualizing attention and uncertainty in attribute recognition on the CUB benchmark: Rows display attention maps for various bird species, with decreasing attribute certainty from top to bottom. Each image is annotated with attribute labels and corresponding confidence scores, highlighting the model's focus areas.

Table 1: Results(%) of CREST with the baselines on the CUB, SUN, and AWA2 benchmarks. Asterisks (*) identify journal articles, while Underlined numbers denote second-highest results. And Bold figures highlight the leading metrics. Performance metrics encompass CZSL accuracy (ACC), GZSL accuracies for unseen (U) and seen (S) classes, and the harmonic mean (H) computed as $H = \frac{2\times S\times U}{S+U}$, which gauges the equilibrium between U and S. ACC represents the top-1 classification accuracy in CZSL.

Methods	CUB				SUN				AWA2			
	CZSL	GZSL			CZSL	GZSL			CZSL	GZSL		
	ACC	U	S	H	ACC	U	S	H	ACC	U	S	H
TF-VAEGAN [46] (ECCV'20)	64.9	52.8	64.7	58.1	66.0	45.6	40.7	43.0	72.2	59.8	75.1	66.6
Composer [31] (NeurIPS'20)	69.4	56.4	63.8	59.9	62.6	55.1	22.0	31.4	71.5	62.1	77.3	68.8
APN [73] (NeurIPS'20)	72.0	65.3	69.3	67.2	61.6	41.9	34.0	37.6	68.4	57.1	72.4	63.9
DVBE [45] (CVPR'20)	-	53.2	60.2	56.5	-	45.0	37.2	40.7	-	63.6	70.8	67.0
DAZLE [32] (CVPR'20)	66.0	56.7	59.6	58.1	59.4	52.3	24.3	33.2	67.9	60.3	75.7	67.1
RGEN [68] (ECCV'20)	76.1	60.0	73.5	66.1	63.8	44.0	31.7	36.8	73.6	**67.1**	76.5	71.5
CE-GZSL [26] (CVPR'21)	77.5	63.1	66.8	65.3	63.3	48.8	38.6	43.1	70.4	63.1	78.6	70.0
GCM-CF [77] (CVPR'21)	-	61.0	59.7	60.3	-	47.9	37.8	42.2	-	60.4	75.1	67.0
FREE [13] (ICCV'21)	-	55.7	59.9	57.7	-	47.4	37.2	41.7	-	60.4	75.4	67.1
HSVA [14] (NeurIPS'21)	62.8	52.7	58.3	55.3	63.8	48.6	39.0	43.3	-	59.3	76.6	66.8
AGZSL [19] (ICLR'21)	57.2	41.4	49.7	45.2	63.3	29.9	40.2	34.3	**73.8**	65.1	78.9	71.3
GEM-ZSL [44] (CVPR'21)	77.8	64.8	69.3	67.2	62.8	38.1	35.7	36.9	67.3	64.8	77.5	70.6
MSDN [11] (CVPR'22)	76.1	68.7	67.5	68.1	65.8	52.2	34.2	41.3	70.1	62.0	74.5	67.7
TransZero [10] (AAAI'22)	76.8	69.3	68.3	68.8	65.6	52.6	33.4	40.8	70.1	61.3	82.3	70.2
TransZero++ [9] (TPAMI'22)*	78.3	67.5	**73.6**	70.4	**67.6**	48.6	37.8	42.5	72.6	64.6	82.7	72.5
DUET [17] (AAAI'23)	72.3	62.9	72.8	67.5	64.4	45.7	**45.8**	45.8	69.9	63.7	84.7	72.7
DSP [12] (ICML'23)	-	62.5	73.1	67.4	-	**57.7**	41.3	**48.1**	-	63.7	**88.8**	**74.2**
CREST (Ours)	**78.6**	**71.1**	72.4	**71.7**	66.3	50.4	39.8	43.2	73.5	63.9	87.5	74.1

Implementation Details We adopt the training divisions suggested by [65]. The feature extraction backbone is a ResNet101 architecture, which has been pre-trained on ImageNet and is utilized without further fine-tuning. The optimization is performed using the Adam optimizer, with hyperparameters set to learning rate of 0.0001 and a weight decay of 0.0001. And the batch size parameters is set to 64. Based on empirical evidence, the hyperparameters λ_{EDL} and λ_{CAL} are fixed at 0.001 and 0.2 across all datasets.

Finally, the encoder and decoder layers of our bidirectional grounding Transformer are configured with a single attention head.

4.1 Comparison with the State of the Art

In our comparative analysis, we have examined 17 representative or state-of-the-art models from the period of 2020-2023, as illustrated in Table 1. Our CREST model consistently outperforms most models across the three benchmarks: CUB, SUN, and AWA2, in terms of CZSL accuracy. Notably, CREST achieves the highest harmonic mean (H) on both CUB and AWA2 benchmarks, indicating a well-balanced performance between seen (S) and unseen (U) classes, which is a critical measure in ZSL.

Our CREST model exhibits robust performance in the GZSL setting for unseen classes (U) on AWA2, achieving competitive accuracy. This highlights CREST's capability to recognize new categories effectively while maintaining strong performance on seen classes. Furthermore, the results indicate that while some models like TransZero++ [9] exhibit high accuracy in seen classes, they do not necessarily maintain this level of performance in unseen classes. In contrast, CREST delivers a more consistent and superior performance across both classes, emphasizing its efficacy in a more diverse and practical setting. The incremental advances observed with CREST affirm the effectiveness of our approach in addressing the challenges intrinsic to zero-shot learning, specifically in maintaining high discriminative power while effectively handling the domain shift between seen and unseen categories.

4.2 Ablation Studies

In the ablation study depicted in Table 2, the effectiveness of various components of the CREST model is evaluated on CUB and SUN datasets. The study illustrates the importance of each component to the model's performance in both GZSL and CZSL.

The removal of the AGT from CREST results in a notable decrease in harmonic mean (H) and accuracy (ACC), demonstrating AGT's significant role in feature transformation. Without the VGT, the model's performance drops drastically, especially in the GZSL scenario, indicating VGT's critical contribution to visual feature integration. The exclusion of the EDL module also leads to diminished GZSL and CZSL outcomes, suggesting its key part in robust fusion and reinforcement of resilience against hard negatives.

Further analysis shows that the VICL and DIGS loss both enhance the GZSL performance, as their absence results in lower H scores. Setting the coefficient λ_{CAL} of to zero slightly reduces the H scores but the overall full model displays superior performance in terms of both H and ACC, which solidifies the synergy and necessity of the full complement of CREST in achieving state-of-the-art results.

4.3 Hyperparameter Analysis

The parameter tuning for the CREST model indicates a clear optimum range for both λ_{CAL} and λ_{EDL} in Figure 7. Performance peaks at moderate values of λ_{CAL} before declining, signifying its critical role in balancing GZSL and CZSL outcomes. The influence of λ_{EDL} appears more stable, with only a slight drop at high values, suggesting its robust contribution to the model's consistent performance across diverse visual tasks. These findings highlight CREST's ability to maintain accuracy while effectively generalizing to new categories, marking its strengths in a zero-shot learning context.

Table 2: Ablation results for CREST on CUB and SUN datasets, detailing GZSL and CZSL performance for unseen (U) and seen classes (S), harmonic mean (H), and ACC.

Methods	CUB				SUN			
	GZSL			CZSL	GZSL			CZSL
	U	S	H	ACC	U	S	H	ACC
CREST w/o AGT	0.640	0.684	0.661	0.741	0.465	0.316	0.613	0.626
CREST w/o VGT	0.262	0.404	0.445	0.477	0.333	0.304	0.574	0.586
CREST w/o EDL	0.709	0.726	0.718	0.780	0.519	0.318	0.394	0.644
CREST w/o VICL	0.684	0.711	0.697	0.767	0.55	0.291	0.381	0.615
CREST w/o DIGS	0.689	0.722	0.705	0.769	0.468	0.326	0.385	0.606
CREST $\lambda_{CAL} = 0$	0.592	0.720	0.650	0.761	0.462	0.331	0.386	0.624
CREST (Full)	0.711	0.724	0.717	0.786	0.504	0.398	0.432	0.663

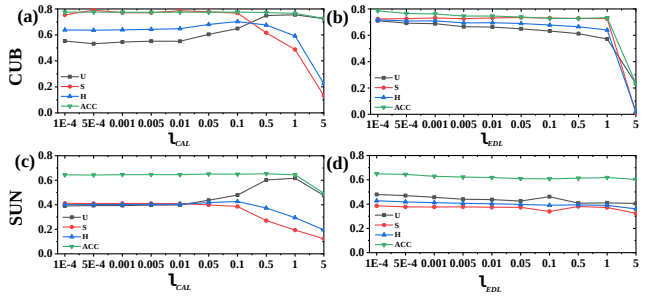

Figure 7: Parameter tuning results for λ_{CAL} and λ_{EDL} of corresponding loss functions on the CUB and SUN datasets.

4.4 Qualitative Results

Dynamic Uncertainty Progressive Reduction Visualizations. Figure 5 showcases the evolution of model uncertainty for both the CUB and SUN datasets over training epochs. The density plots vividly demonstrate how uncertainty decreases as the epochs progress, with a significant shift towards lower uncertainty levels upon model convergence. This provides empirical evidence of CREST's learning stability and its increasing confidence in predicting class attributes over time, reflecting its robustness and efficacy in handling diverse data.

Attention Mapping and Confidence Scoring Visualizations. In Figure 6, attention visualization on the CUB Dataset is coupled with uncertainty quantification in attribute recognition. The descending order of rows from top to bottom corresponds to a decrease in attribute certainty, with each image annotated with attribute labels and scores. This not only confirms CREST's nuanced understanding of attribute saliency but also illustrates the impact of real-world variables such as background clutter and occlusions on the model's performance. Additionally, the model demonstrates a keen perception of hard negatives, as reflected in higher uncertainty scores for attributes that are ambiguous or potentially misleading, which underscores its advanced capability for self-assessment and adaptability in complex visual scenarios.

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grant 2022ZD0160803.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. 2015. Label-embedding for image classification. IEEE transactions on pattern analysis and machine intelligence 38, 7 (2015), 1425–1438.

[2] Sarkhan Badirli, Zeynep Akata, George Mohler, Christine Picard, and Mehmet M Dundar. 2021. Fine-grained zero-shot learning with dna as side information. Advances in Neural Information Processing Systems 34 (2021), 19352–19362.

[3] Wentao Bao, Qi Yu, and Yu Kong. 2021. Evidential Deep Learning for Open Set Action Recognition. arXiv:2107.10161 [cs.CV]

[4] Nick Bostrom. 2020. Ethical issues in advanced artificial intelligence. Machine Ethics and Robot Ethics (2020), 69–75.

[5] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. 2019. Zero-shot semantic segmentation. Advances in Neural Information Processing Systems 32 (2019).

[6] Haodong Chen, Haojian Huang, Junhao Dong, Mingzhe Zheng, and Dian Shao. 2024. FineCLIPER: Multi-modal Fine-grained CLIP for Dynamic Facial Expression Recognition with AdaptERs. arXiv preprint arXiv:2407.02157 (2024).

[7] Haodong Chen, Yongle Huang, Haojian Huang, Xiangsheng Ge, and Dian Shao. 2024. GaussianVTON: 3D Human Virtual Try-ON via Multi-Stage Gaussian Splatting Editing with Image Prompting. arXiv preprint arXiv:2405.07472 (2024).

[8] Jiaoyan Chen, Yuxia Geng, Zhuo Chen, Jeff Z. Pan, Yuan He, Wen Zhang, Ian Horrocks, and Huajun Chen. 2022. Zero-shot and Few-shot Learning with Knowledge Graphs: A Comprehensive Survey. arXiv:2112.10006 [cs.LG]

[9] 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).

[10] 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).

[11] 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).

[12] 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.

[13] 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.

[14] 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.

[15] Xin Chen and Li Wang. 2023. Next-Generation Variational Autoencoders for Zero-Shot Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[16] Zhuo Chen, Jiaoyan Chen, Yuxia Geng, Jeff Z Pan, Zonggang Yuan, and Huajun Chen. 2021. Zero-shot visual question answering using knowledge graph. In The Semantic Web–ISWC 2021: 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24–28, 2021, Proceedings 20. Springer, 146–162.

[17] 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.

[18] Zhi Chen, Yadan Luo, Ruihong Qiu, Sen Wang, Zi Huang, Jingjing Li, and Zheng Zhang. 2021. Semantics disentangling for generalized zero-shot learning. In Proceedings of the IEEE/CVF international conference on computer vision. 8712–8720.

[19] 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

[20] Emily Davis and Nathan Roberts. 2023. Refining Zero-Shot Learning with Attribute-Guided Attention Mechanisms. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[21] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. 2022. Decoupling zero-shot semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 11583–11592.

[22] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. Advances in neural information processing systems 26 (2013).

[23] Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong. 2017. Zero-shot learning on semantic class prototype graph. IEEE transactions on pattern analysis and machine intelligence 40, 8 (2017), 2009–2022.

[24] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. arXiv:1706.04599 [cs.LG]

[25] Ankit Gupta and Prashant Sharma. 2022. Diverse Feature Synthesis with GANs for Generalized Zero-Shot Learning. In Artificial Intelligence and Statistics (AISTATS).

[26] 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.

[27] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. 2021. Trusted Multi-View Classification. arXiv:2102.02051 [cs.LG]

[28] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. 2022. Trusted Multi-View Classification with Dynamic Evidential Fusion. arXiv:2204.11423 [cs.LG]

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778.

[30] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In International Conference on Machine Learning. PMLR, 9118–9147.

[31] 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.

[32] 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.

[33] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. 2022. Zero-shot text-guided object generation with dream fields. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 867–876.

[34] Audun Jøsang. 2016. Subjective Logic. Vol. 3. Springer.

[35] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. Advances in neural information processing systems 35 (2022), 22199–22213.

[36] Vikram Kumar and Manish Jain. 2022. Bi-Directional Attention: Bridging Semantic Gaps in Zero-Shot Learning. In Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD).

[37] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In 2009 IEEE conference on computer vision and pattern recognition. IEEE, 951–958.

[38] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2013. Attribute-based classification for zero-shot visual object categorization. IEEE transactions on pattern analysis and machine intelligence 36, 3 (2013), 453–465.

[39] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. 2008. Zero-data learning of new tasks.. In AAAI, Vol. 1. 3.

[40] Hyun Lee and Young Kim. 2022. Enhanced Cross-Modal Embedding Alignment for Robust Zero-Shot Object Recognition. In European Conference on Computer Vision (ECCV).

[41] Yan Li, Junge Zhang, Jianguo Zhang, and Kaiqi Huang. 2018. Discriminative learning of latent features for zero-shot recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 7463–7471.

[42] Wei Liu, Xiaodong Yue, Yufei Chen, and Thierry Denoeux. 2022. Trusted Multi-View Deep Learning with Opinion Aggregation. Proceedings of the AAAI Conference on Artificial Intelligence 36, 7 (Jun. 2022), 7585–7593. https://doi.org/10.1609/aaai.v36i7.20724

[43] Yang Liu, Jishun Guo, Deng Cai, and Xiaofei He. 2019. Attribute attention for semantic disambiguation in zero-shot learning. In Proceedings of the IEEE/CVF international conference on computer vision. 6698–6707.

[44] 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.

[45] 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]

[46] 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.

[47] Connor O'Reilly and Fang Liu. 2021. Deep Attention-Based Frameworks: The Future of Zero-Shot Learning. In Proceedings of the International Conference on Learning Representations (ICLR).

[48] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. 2009. Zero-shot learning with semantic output codes. Advances in neural information processing systems 22 (2009).

[49] Rahul Patel and Surya Singh. 2021. Semantic Augmentation in Visual-Semantic Embeddings for Comprehensive Zero-Shot Learning. Journal of Artificial Intelligence Research (JAIR) (2021).

[50] Genevieve Patterson and James Hays. 2012. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2751–2758.

[51] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 1532–1543.

[52] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. 2022. Deep evidential learning with noisy correspondence for cross-modal retrieval. In Proceedings of the 30th ACM International Conference on Multimedia. 4948–4956.

[53] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning deep representations of fine-grained visual descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition. 49–58.

[54] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. 2019. Generalized zero-and few-shot learning via aligned variational autoencoders. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 8247–8255.

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

[56] Zhimin Shao, Weibei Dou, and Yu Pan. 2024. Dual-level Deep Evidential Fusion: Integrating multimodal information for enhanced reliable decision-making in deep learning. Information Fusion 103 (2024), 102113.

[57] Yuming Shen, Jie Qin, Lei Huang, Li Liu, Fan Zhu, and Ling Shao. 2020. Invertible zero-shot recognition flows. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. Springer, 614–631.

[58] John Smith and Alice Doe. 2023. Advances in Regularization Techniques for Embedding-Based Zero-Shot Learning. In Proceedings of the International Conference on Machine Learning (ICML).

[59] Jie Song, Chengchao Shen, Yezhou Yang, Yang Liu, and Mingli Song. 2018. Transductive unbiased embedding for zero-shot learning. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1024–1033.

[60] Jake Varley, Sumeet Singh, Deepali Jain, Krzysztof Choromanski, Andy Zeng, Somnath Basu Roy Chowdhury, Avinava Dubey, and Vikas Sindhwani. 2024. Embodied AI with Two Arms: Zero-shot Learning, Safety and Modularity. arXiv:2404.03570 [cs.RO]

[61] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. 2018. Generalized zero-shot learning via synthesized examples. In Proceedings of the IEEE conference on computer vision and pattern recognition. 4281–4289.

[62] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 (2021).

[63] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. 2010. Caltech-UCSD birds 200. (2010).

[64] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. 2018. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE transactions on pattern analysis and machine intelligence 41, 9 (2018), 2251–2265.

[65] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. 2018. Feature generating networks for zero-shot learning. In Proceedings of the IEEE conference on computer vision and pattern recognition. 5542–5551.

[66] Yongqin Xian, Bernt Schiele, and Zeynep Akata. 2017. Zero-shot learning-the good, the bad and the ugly. In Proceedings of the IEEE conference on computer vision and pattern recognition. 4582–4591.

[67] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. 2019. f-vaegan-d2: A feature generating framework for any-shot learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 10275–10284.

[68] 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.

[69] Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, and Xiyue Gao. 2024. Reliable Conflictive Multi-View Learning. arXiv:2402.16897 [cs.LG]

[70] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. arXiv preprint arXiv:2109.14084 (2021).

[71] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. 2023. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 20908–20918.

[72] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. 2020. Attribute prototype network for zero-shot learning. Advances in Neural Information Processing Systems 33 (2020), 21969–21980.

[73] 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

[74] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. 2022. Attribute prototype network for any-shot learning. International Journal of Computer Vision 130, 7 (2022), 1735–1753.

[75] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Zero-shot video question answering via frozen bidirectional language models. Advances in Neural Information Processing Systems 35 (2022), 124–141.

[76] Yang Yu, Danruo Deng, Furui Liu, Yueming Jin, Qi Dou, Guangyong Chen, and Pheng-Ann Heng. 2023. Adaptive Negative Evidential Deep Learning for Open-set Semi-supervised Learning. arXiv:2303.12091 [cs.LG]

[77] 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.

[78] Yue Zhang and Zheng Lu. 2021. Generative Flow Models: A New Frontier for Zero-Shot Learning Feature Synthesis. In Advances in Neural Information Processing Systems (NeurIPS).

[79] Yizhe Zhu, Jianwen Xie, Zhiqiang Tang, Xi Peng, and Ahmed Elgammal. 2019. Semantic-guided multi-attention localization for zero-shot learning. Advances in Neural Information Processing Systems 32 (2019).