electronics

MDPI

Article

FFusion: Feature Fusion Transformer for Zero-Shot Learning

Wenjin Tao, Jiahao Xie, Zhinan An and Xianjia Meng *

School of Information Science and Technology, Northwest University, Xi'an 710127, China; taowenjin@stumail.nwu.edu.cn (W.T.)
* Correspondence: xianjiam@nwu.edu.cn

Abstract: Zero-shot learning (ZSL) typically leverages semantic knowledge and textual descriptions of classes to forge connections between seen and unseen classes. ZSL can classify new categories of data unseen in the training set. Prior research has focused on aligning image features with their corresponding auxiliary information, overlooking the limitation whereby individual features may not capture the full spectrum of information inherent in the original image. Additionally, there are concerns regarding the bias in the predicted results towards seen classes during the testing phase. In this work, we introduce a novel approach termed Feature Fusion Transformer Network (FFusion), which decomposes visual and semantic features into a multitude of functionally distinct features, fusing them to form new representations that better encapsulate the information content of the original images during training. Furthermore, we implement novel loss functions to balance the model's focus on unseen and seen classes within the ZSL framework. Experiments show that our model achieves accuracy of 76.8% in the CUB dataset, while reducing the bias between seen and unseen classes to a mere 0.2%.

Keywords: zero-shot learning; feature fusion; semantic features; visual features

1. Introduction

Zero-shot learning (ZSL), which is a cutting-edge deep learning approach [1–3], seeks to overcome the reliance on labeled data [4–6], tackling the challenge of classifying unseen classes [7,8]. ZSL emulates the human capacity for the recognition of novelty, identifying unseen classes during inference by leveraging the inherent associations between seen and unseen classes gleaned from training data [9,10]. Imagine a bird researcher who is well versed in the characteristics of many bird species but has never seen a Common Raven. However, knowing that it has a large, slightly curved beak and black, glossy feathers, they can easily identify one on first sight by matching these features with their existing knowledge. The key idea of ZSL is to employ semantic information as a core link between seen and unseen classes [11,12]. ZSL distills visual features from seen classes by training a dataset of seen classes, combining them with semantic attributes, and then extends this to accurately recognize and classify unseen classes, thus enhancing cross-class generalization [13–16]. ZSL requires a meticulous algorithm design during the training phase to match the visual image information of the seen classes with the semantic information, ensuring generalization to unseen classes. In the testing phase, when encountering an instance of an unseen class, an ideal ZSL model discerns the pertinent semantic features to predict its attributes accurately. Therefore, two crucial challenges must be overcome in ZSL: (1) accurately associating visual information with semantic information; (2) reducing the overfitting of the model to the seen classes' data during the training phase and increasing the generalization of the model to unseen classes.

check for updates

Academic Editor: Chiman Kwan

Received: 17 January 2025
Revised: 23 February 2025
Accepted: 24 February 2025
Published: 26 February 2025

Citation: Tao, W.; Xie, J.; An, Z.; Meng, X. FFusion: Feature Fusion Transformer for Zero-Shot Learning. *Electronics* **2025**, *14*, 926. https://doi.org/10.3390/electronics14050926

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Early ZSL methods usually establish a mapping space. Visual and semantic information of the same class is embedded in each other in this space. In these methods, the mappings used do not fully reflect the complex relationship between visual and semantic information, so ZSL performs well on seen classes but has limited effectiveness on unseen classes [17]. In response to this challenge, ALE [18], CRnet [19], TFPL [20], and DVBE [21] employ global visual features derived from the characteristics of seen classes. However, the ALE, CRnet, DVBE, and TFPL systems each target distinct core challenges within the realm of ZSL, with ALE aiming to bolster the image classification performance via label embedding, CRnet and DVBE striving to alleviate the biases inherent in ZSL, and TFPL working to augment the recognition capabilities for unseen classes through deep feature learning and projection function learning. Nonetheless, they exhibit commonality: they all leverage the global features of images to attain their respective objectives. However, by relying exclusively on global features, these models cannot precisely capture the local features that directly correspond to semantic attributes. For example, if a class's distinctive feature is the color of its eyes, global features capture the general appearance and not the nuanced details. Consequently, employing global features for embedding may not accurately represent this unique attribute, potentially resulting in inferior outcomes.

Due to the constraints of global visual features, the mapping between visual and semantic features remains insufficiently distinct. Consequently, several methods integrate attention mechanisms into models to identify local image attributes. S^2GA [22] integrates a stacked semantic guidance attention mechanism to enhance the learning of more discriminative local visual features. AREN [23] is capable of extracting more nuanced local feature representations under the guidance of attention and contrast loss mechanisms, yet it does not adequately represent the interrelationships among local features. Subsequently, AREN further uses region map learning to refine local inferences of visual features [24]. SGMA [25] utilizes a weakly supervised learning strategy and incorporates multiple attention mechanisms for the precise localization of visual features. APN [26] augments the model's capacity for visual–semantic interaction through the fusion of local and global features. Nonetheless, these methods are not without their drawbacks: (1) the integrated attention mechanism struggles to achieve the precise localization of local features, hindering effective representation; (2) these methods exhibit biases when classifying both seen and unseen classes. To address these issues, we propose the following solutions: (1) we introduce a novel feature representation method that synthesizes deep and shallow visual and semantic information for a more holistic local feature representation; (2) we implement a novel loss function that accounts for class distribution disparities and diminishes the reliance on seen classes via regularization, thereby enhancing the model's capacity to recognize unseen classes. The main contributions of this paper are summarized as follows:

- We introduce FFusion, featuring a novel feature fusion module that amalgamates key local, mixed, and global features. The feature fusion approach within FFusion mitigates the reliance on a singular feature source for different features to complement each other.
- The feature fusion technique provides a clear conceptual framework that facilitates an understanding of how the model integrates information from various feature sources.
- We implement a debiasing loss function designed to equitably distribute the model's focus between seen and unseen classes. This function diminishes the model's dependence on seen classes, thereby enhancing the recognition of unseen classes through judicious balancing.

- We execute an extensive experimental study to demonstrate the efficacy of the proposed FFusion model, with comparative analyses against leading approaches demonstrating our model's competitiveness.

The article is organized as follows. Section 2 presents the related work, reviewing existing methods and research progress in the field of ZSL. Section 3 details our proposed method, including the design of the feature fusion module and the debiasing loss function. Section 4 provides the experimental results, which demonstrate the effectiveness of our method on multiple standard datasets. Section 5 discusses the experimental results, analyzes the limitations of our method, and puts forward directions for future improvement. Finally, Section 6 summarizes the main contributions and conclusions of this paper.

2. Related Work

2.1. Zero-Shot Learning

The primary objective of embedded ZSL is to devise an efficient mapping strategy for the visual–semantic space within the embedding continuum. Upon encountering an entirely new class of data, the model ascertains their class by aligning the visual features with the most pertinent semantic attributes within the unseen classes' embedding space [27,28]. ALE [18] employs a ranking loss function to obtain a discriminative mapping that contrasts visual and semantic features, thereby facilitating semantic embedding. Conversely, SJE [29] and ESZSL [4] each introduce novel loss functions specifically tailored to ZSL learning. SJE employs a bilinear mapping approach to effectively integrate visual features with categorical semantic information. ESZSL employs a squared loss function to refine the bilinear contrast mechanism and explicitly regularizes the objective function. CMT [30], in contrast, develops visual–semantic nonlinear mappings through a neural network architecture featuring two hidden layers. These methods devise new mapping spaces or loss functions for the construction of visual attribute spaces. However, all of these methods mainly focus on training using global features, ignoring the problem wherein global features cannot describe local fine-grained details, resulting in poor visual–semantic consistency. Consequently, a subset of research has initiated the employment of attention mechanisms to discern local features, with methods such as S^2GA [22], AREN [23], and SGMA [25] integrating local and global features. S^2GA leverages a stacked, semantically guided attention mechanism to bolster the training of local features, thereby promoting visual–semantic congruence. AREN further explores more localized feature representations, facilitated by attention mechanisms and contrastive loss. To address the underdeveloped representation of inter-feature relationships, AREN incorporates a region map learning strategy. SGMA adopts a weakly supervised learning approach to accurately localize visual features through the integration of diverse attention mechanisms. These methods generally exhibit limitations in representing local features, as the employed attention mechanisms may not be sufficiently potent for local attribute identification. Concurrently, the attention mechanisms are less adept in extracting pivotal common semantic knowledge from visual attributes. To address the issue of inadequate local feature expression, we introduce the weighting of visual features based on the interaction matrix with attribute features, yielding a local feature representation conducive to feature fusion.

2.2. Transformer Model

Since its introduction by Vaswani et al. [31], Transformer has emerged as the dominant approach for Natural Language Processing (NLP) tasks, leveraging its self-attention mechanism to capture long-range dependencies across sequences. Several methodologies, including those of Huang et al. [32] and Cornia et al. [33], have demonstrated the

effectiveness of the self-attention mechanism. Within the realm of computer vision, this paradigm has been integrated via non-local blocks and their derivatives, further being operationalized within the scope of visual Transformer models. For example, Vision Transformer (ViT) [34] was the first endeavor in this domain, attaining performance on par with that of conventional Convolutional Neural Network (CNN) frameworks, contingent upon substantial training data. Subsequent endeavors, exemplified by methods such as DeiT [35], have augmented the performance via sophisticated training regimens and knowledge distillation techniques. At present, the research community is dedicated to amalgamating the strengths of Transformer and CNN to more effectively harness local information [36]. Building on this foundation, our proposed Transformer-based FFusion holistically incorporates features across various dimensions and modalities to address the shortcomings associated with inadequate attention to global and local features.

3. Proposed Method

3.1. Problem Definition

In generalized zero-shot learning (GZSL), the training data for seen classes are denoted as $D_{seen} = \{X_i^s, Y_i^s\}$, encompassing C_{seen} classes in total. Here, $X_i^s \in X$ signifies training sample i, while Y_i^s signifies its corresponding class. Furthermore, a dataset exists for unseen classes, denoted as $D_{unseen} = \{X_i^u, Y_i^u\}$, with $X_i^u \in X$ and Y_i^u representing the unseen class samples and their respective labels. In total, there are C_{unseen} unseen classes. These data are excluded from the training process, being utilized solely for the assessment of the model performance during the testing phase. A class $c \in C_{seen}$, part of the union $C_{seen} \cup C_{unseen} = C$, possesses a semantic vector z^c encompassing attributes α, represented in the form of $z^c = \{z_1, z_2, \ldots, z_\alpha\}$. Semantic vector sets are used to describe different attributes. Employing semantic information, we forge connections between seen and unseen classes, aiming to transfer knowledge from seen to unseen classes. We harness attribute descriptions within the semantic vector set to discern the semantic attribute information for each attribute, represented as $G_\alpha = \{u_\beta\}_{\beta=1}^{\alpha}$, utilizing the language model GloVe [37].

3.2. Overview

In this study, we introduce a Feature Fusion Transformer Network (FFusion) designed for ZSL. As depicted in Figure 1, FFusion is composed of a Feature Fusion Transformer (FFT) and a Feature Interaction Network (FIN). Initially, FFusion necessitates two input types, the visual global features F_g and the semantic features F_{sem}, sourced from the CNN backbone and GloVe, respectively. The FFT distills multi-dimensional features from both the visual and semantic domains, encompassing global features F_g, local features F_l, mixed features F_{mix}, and key local features F_{max}. It integrates these multi-dimensional features to generate a composite feature, which is a rich, high-level representation that encapsulates the intrinsic characteristics of the original features across their respective dimensions. Following the acquisition of the fused visual features via the FFT, we proceed to map these features onto the semantic embedding space. The mapping function, denoted as M, within the FIN correlates the integrated features with the semantic features F_{sem} to ascertain the confidence score for the α-th attribute of image X_i. Moreover, we optimize our model's performance by incorporating the attribute regression loss $\mathcal{L}_{AR}$, the attribute cross-entropy loss $\mathcal{L}_{ACE}$, and the debiasing loss $\mathcal{L}_{DEB}$.

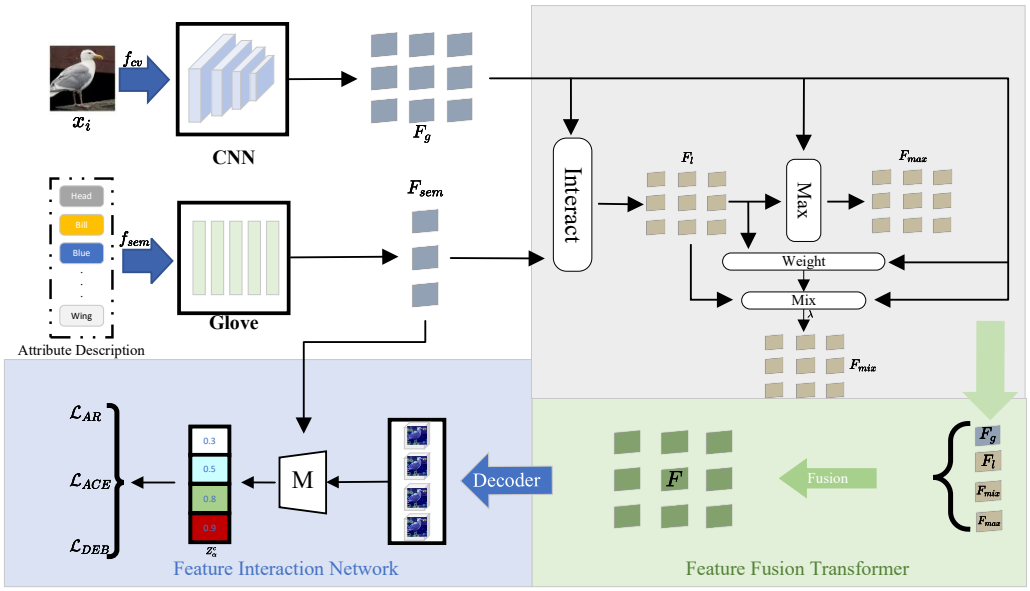

Figure 1. The overall framework diagram of FFusion.

3.3. Feature Fusion Transformer (FFT)

In ZSL, local and global features are both crucial, yet they bring distinct challenges depending on the situation. Global features offer a macroscopic view of an image, encompassing attributes like the overall shape, size, and color distribution, facilitating the identification of key categorical distinctions. Nonetheless, global features might lack the requisite detail precision. If the disparities in the global features among classes are subtle, or if intricate detail is essential for classification, global features might fail to offer adequate discriminative cues, thereby diminishing the classification accuracy. Local features focus on discrete regions within an image, capturing granular details like texture and contours. They are particularly valuable in discerning specific objects or attributes, such as the recognition of distinct avian features like the beak and feathers. Nonetheless, local features alone cannot encapsulate the entirety of a class. Furthermore, the extraction of local features is contingent upon the semantic attributes of the targeted regions, potentially overlooking other salient image information. To learn the underlying relationship between global (F_g) and local (F_l) features, we employ a CNN to derive visual global features F_g from the raw visual data f_{cv} and utilize GloVe to extract semantic features F_{sem} from the raw semantic data f_{sem}. The formula used is as follows:

$$(F_g, F_{sem}) = (CNN(f_{cv}), GloVe(f_{sem})) \tag{1}$$

Our model is designed to classify unseen classes; therefore, augmenting the features' expressive power is essential. Through the integration of visual (image) features with semantic (class attribute) features, the combined feature set's expressive capacity is augmented. Specifically, we compute an interaction feature a_i by assessing the interplay between visual and semantic features. Subsequently, we apply normalization to the interaction feature a_i to derive a weight vector A_i, facilitating the conversion of interaction features into weights for subsequent calculations. The normalization formula is presented below:

$$A_i = Softmax(a_i) = \frac{e^{a_i}}{\sum_j e^{a_j}} \tag{2}$$

In this formula, j is the sum across all class elements, with the summation aggregating all elements of the last dimension.

Subsequently, we apply the feature weight vector A_i to the global visual features F_g, yielding a new set of weighted visual features, designated as the refined local features F_l. By assigning weights to the visual features, the model synthesizes information from the visual and semantic domains, attaining a higher resolution in feature representation. This granular feature information enhances the model's capacity to capture the subtleties of the image. The corresponding formula, expressing this operation, is presented below:

$$F_l = \sum_{r=1}^{R} A_r \cdot F_g^r \tag{3}$$

where r is the index in the summation process, ranging from 1 to R. R represents the size of the last dimension of the features.

We obtain both global and local features; however, leveraging them effectively is a challenge. Consequently, to enable the model to adaptively discern the significance of each feature and promote knowledge transfer from seen to unseen classes [38], we determine the feature weight λ, enabling the dynamic modulation of the global and local feature contributions. This equilibrium is vital for zero-shot classification, enabling the model to adjust the feature amalgamation flexibly in response to the unique characteristics of various unseen classes. Specifically, we calculate the feature weight λ to balance the global features F_g and the local features F_l. The incorporation of feature weight enhances the model's discrimination of novel classes. This is attributed to the weight's capacity to enable the model to identify and leverage pivotal features common to all classes, thereby contributing to classification. Subsequently, the feature weight λ are applied to synthesize a mixed feature set F_{mix} through the weighted aggregation of the global features F_g and the local features F_l, enabling the model to reconcile disparate feature contributions and amalgamate their similarities and significance, thereby enriching the information available for ZSL. The formula defining F_{mix} is presented as follows:

$$F_{mix} = \lambda \cdot F_g + (1 - \lambda) \cdot F_l \tag{4}$$

Owing to the cross-dataset bias [15] between ImageNet and the BML benchmark, coupled with distributional inconsistencies between the training and test datasets, the model must exhibit resilience against noise and extraneous feature interference. By identifying the most salient features, the model avoids undue preoccupation with extraneous features in training, thereby enhancing its ability to discern the data's intrinsic characteristics. Specifically, the key local feature F_{max} is determined through element-wise maximization between the original and the weighted visual features, effectively pinpointing the peak response at each spatial location. This approach is conducive to retaining features with high relevance to the data while discarding the irrelevant ones. The formulation defining F_{max} is presented as follows:

$$F_{max} = max(F_g, F_l) \tag{5}$$

In our previous efforts, we have procured four distinct feature representations, aiming to equip the model with an encompassing feature representation capable of capturing distinctive information from varied feature sources. Our methodology effectively amplifies the model's proficiency in identifying previously unencountered classes and diminishes the model's dependence on a single type of feature. Ultimately, we integrate the global feature F_g, local feature F_l, mixed feature F_{mix}, and key local feature F_{max} to derive the ultimate feature representation F. The fusion of these features can attenuate the disparities among them, preserving the aggregate information and diminishing the model's dependency on any one feature. Furthermore, this fusion strategy is likely to enhance the model's

adaptability across varied data distributions. The corresponding formula, encapsulating this strategy, is delineated below:

$$F = \frac{1}{n}(F_g + F_l + F_{max} + F_{mix}) \tag{6}$$

where $\frac{1}{n}$ is the fusion coefficient, the value of which is ascertained through empirical optimization to attain optimal feature integration.

3.4. Feature Interaction Network (FIN)

After the FFT, we acquire the definitive feature fusion F, and, to optimize its capabilities, we integrate the semantic feature F_{sem}, augmenting the model's categorization proficiency. Specifically, once the feature fusion is generated, it is subsequently mapped onto the semantic embedding space. Furthermore, the semantic attribute vector $G_\alpha = \{u_\beta\}_{\beta=1}^{\alpha}$ within the semantic feature F_{sem} serves as a support vector, facilitated by the mapping function M, enhancing the visual feature F with semantic attribute information to achieve alignment. The fused feature fusion F is mapped to a semantic embedding space that is congruent with the semantic attribute information G_α via the mapping function M. This mapping procedure not only ensures the strict alignment of the feature representations with the semantic attributes of the classes, but also equips the model with the means to accurately classify them within a high-dimensional semantic space, facilitated by the transformation of the embedding matrix W. The FIN's output, the attribute score $\theta(x_i)$, quantifies the confidence level for each attribute within the image x_i, offering a direct basis for inference based on attribute classes. The FIN is engineered to expedite knowledge transfer from seen to unseen classes. Through the manipulation of features within the semantic embedding space, the model leverages the attribute information from seen classes to deduce the attributes of unseen classes.

$$\theta(x_i) = M(F) = \sum_j u_{\beta,j}(WF)_j \tag{7}$$

where W is the embedding matrix responsible for mapping F into the semantic attribute space. $u_{\beta,j}$ denotes the j th component of the vector u_β. Essentially, $\theta(x_i)[\alpha]$ denotes the attribute score indicative of the confidence level for the presence of the α-th attribute in image x_i. Considering a collection of semantic attribute vectors $G_\alpha = \{u_\beta\}_{\beta=1}^{\alpha}$, FFusion derives the corresponding semantic embedding matrix $\theta(x_i)$.

3.5. Model Optimization

To effectively optimize and train a model proficient in zero-shot classification, it must accurately predict the attributes of seen classes and efficiently classify unseen classes, mitigating biases stemming from data distribution inconsistencies simultaneously. We have integrated three distinct loss functions into the training of FFusion: the attribute regression loss, attribute cross-entropy loss, and debiasing loss.

3.5.1. Attribute Regression Loss

To motivate the model to map visual features accurately to their semantic embeddings, we implement the attribute regression loss $\mathcal{L}_{AR}$ as a regularizing factor for FFusion. The primary objective of the attribute regression loss is to guarantee the accurate prediction of the presence probability for each attribute within the image. By reducing the discrepancy between the predicted attribute scores and actual attribute values, the model refines its visual feature mappings to more closely mirror the semantic attributes of the image. In this context, visual–semantic mapping is regarded as a regression task, where we minimize the

mean squared error between the ground-truth attribute set z^c of size m_α and the embedded attribute scores $\xi(x_i^s)$:

$$\mathcal{L}_{AR} = -\frac{1}{m_\alpha}\sum_{i=1}^{m_\alpha}(\xi(x_i^s) - z^c)^T(\xi(x_i^s) - z^c) \tag{8}$$

3.5.2. Attribute-Based Cross-Entropy Loss

Given the presence of visually relevant attributes in an image, which aligns the image embedding closely with its semantic vector z^c, we refine the parameters of the FFusion model utilizing the attribute-based cross-entropy loss $\mathcal{L}_{ACE}$. The attribute-based cross-entropy loss enhances the model's discernment of nuanced inter-class differences and directs it to differentiate among the semantic attributes among classes [39]. This approach enhances the model's precision in categorization by optimizing the congruence between image representations and pertinent class-specific semantic vectors, while concurrently reducing the correspondence with extraneous class vectors. That is, class logits are derived by initially computing the dot product of visual embeddings against each class semantic vector. This computation is executed to ensure that the image achieves the maximum compatibility score with its corresponding class semantic vector. Considering a batch of m_α training images $\sum_{i=1}^{m_\alpha}$ and their corresponding class semantic vectors z^c, we define $\mathcal{L}_{ACE}$ as

$$\mathcal{L}_{ACE} = -\frac{1}{m_\alpha}\sum_{i=1}^{m_\alpha}\log\frac{\exp(\xi(x_i^s)\times z^c)}{\sum_{\hat{c}\in C^s}\exp(\xi(x_i^s)\times z^{\hat{c}})} \tag{9}$$

where C^s is the set of seen classes within the complete class sample set C.

3.5.3. Debiasing Loss

A major challenge in ZSL is mitigating the model's propensity for bias towards visible classes in the training dataset—that is, since the model is trained exclusively on visible classes, the accuracy for these classes is inevitably much higher than for unseen classes during testing. Furthermore, complementing the aforementioned attribute regression and cross-entropy losses, we have implement a debiasing loss function $\mathcal{L}_{DEB}$ aimed at attenuating this bias, ensuring uniformity in the model's predictive distributions for both seen and unseen classes. The use of $\mathcal{L}_{DEB}$, seeking to ensure consistency in both the mean and variance of distributions, can more effectively balance the reliance on scores between seen and unseen classes. Additionally, by minimizing the disparities in the distribution of the means and variances of the predicted scores across the two domains, the debiasing loss aids in bolstering the model's discriminate capabilities for unseen classes [40]. The formulation of the debiasing loss function is presented as follows:

$$\mathcal{L}_{DEB} = \frac{1}{m_\alpha}\sum_{i=1}^{m_\alpha}(|\bar{X}_{s,i} - \bar{X}_{u,i}| + |\sigma_{s,i} - \sigma_{u,i}|) \tag{10}$$

where $\bar{X}_s$ and σ_s denote the mean and variance of the predicted scores for seen classes, and $\bar{X}_u$ and σ_u denote those for unseen classes.

3.5.4. Overall Optimization Objective

To achieve multifaceted model optimization, we diminish the reliance on the training data distribution and foster a more generalized feature representation. We amalgamate the attribute regression loss, attribute-based cross-entropy loss, and debiasing loss into a

unified comprehensive loss function. Ultimately, we articulate the overarching optimization objective as:

$$\mathcal{L} = \lambda_{AR}\mathcal{L}_{AR} + \mathcal{L}_{ACE} + \lambda_{DEB}\mathcal{L}_{DEB} \tag{11}$$

where the weighting coefficients λ_{AR} and λ_{DEB} are employed to modulate the influence of disparate loss components.

4. Experiments

4.1. Experimental Setup

To thoroughly assess the efficacy of our proposed model across datasets with diverse levels of complexity, we perform experiments on three well-established ZSL benchmark datasets: CUB [41], AWA2 [17], and SUN [42]. These datasets encompass a spectrum of class divisions, ranging from fine-grained to coarse-grained, offering a variety of testing scenarios for our model.

- The CUB dataset comprises 11,788 images across 200 bird species, categorized into 150 seen (training/test set = 100/50) and 50 unseen classes, with 312 attributes detailing the semantic information of the images, capturing fine-grained features such as color and shape. This presents the model with an extensive fine-grained classification challenge.
- The AWA2 dataset contains 37,322 images across 50 animal classes, with 40 seen (training/test set = 27/13) and 10 unseen classes, utilizing 85 attributes to describe the dataset's animal attributes, thereby offering the associated images and attributes.
- The SUN dataset exhibits greater diversity, encompassing 717 scene classes and 14,340 images, posing a more complex challenge for classification tasks, with 645 seen (training/test set = 580/65) and 72 unseen classes. Each class within the SUN dataset features 102 attribute descriptions, focusing on scene classification and attribute recognition, offering a substantial collection of scene images and annotations that is ideal for scene recognition and attribute learning tasks. It represents a coarse-grained dataset suitable for ZSL.

4.2. Evaluation Protocols

We employ two distinct evaluation frameworks: conventional zero-shot learning (CZSL) and generalized zero-shot learning (GZSL).

- Within the CZSL framework, we evaluate the ability of the model to classify unseen classes using separate sample sets for training and testing. Specifically, the training set is composed exclusively of samples from seen classes, while the test set consists solely of samples from unseen classes. Initially, the mean accuracy per class is determined, followed by dividing the aggregate total by the class count; this process indicates the model performance through the $top-1$ accuracy metric, labeled as acc.
- Within the GZSL framework, we evaluate the classification performance across both seen and unseen classes, using both seen and unseen samples in the test set. Therefore, we calculate the mean $top-1$ accuracy per class for seen classes, denoted as S, and for unseen classes, denoted as U. Furthermore, besides calculating these accuracies individually, we employ the harmonic mean, denoted as H, as an integrative metric to equitably balance the classification performance between seen and unseen classes.

$$H = 2\frac{SU}{(S+U)} \tag{12}$$

In this context, U denotes the classification accuracy achieved for the unseen classes, while S signifies that for the seen classes.

We employ the dataset segmentation approach outlined in Xian et al. [14] for the extraction of feature maps, utilizing a ResNet101 network pre-trained on ImageNet as our CNN backbone. In creating the trainable components of our model, we leverage the stochastic gradient descent (SGD) algorithm, fine-tuned with hyperparameters that included momentum of 0.9 and a weight decay factor of 0.0001, thereby enhancing the model's efficacy. The learning rate is set to 0.0001, complemented by a batch size of 50 to ensure stable training dynamics. For the attribute regression and debiasing components, we designate coefficients of 0.003 and 0.2 for λ_{AR} and λ_{DEB}, applying these to all datasets.

4.3. Comparison with State-of-the-Art Methods

The sophistication of our FFusion model is effectively showcased in the context of CZSL. As evidenced in Table 1, FFusion achieves optimal accuracy (*acc*) of 76.8% and competitive accuracy of 64.4% for CZSL on the CUB and SUN datasets, respectively, and these results highlight the model's ability to deal with classifications of varying coarse and fine granularities. Compared to BGSNet [43], the state-of-the-art method on the CUB dataset, FFusion attains a significant improvement of over 3.5%. Despite achieving sub-optimal results on the SUN dataset in comparison to leading methods such as AREES [44] and SupVFD [45], FFusion's performance remains within a competitive range. On the AWA2 dataset, FFusion exhibits upper- to mid-level performance of 65%.

Regarding the GZSL scenario, Table 1 delineates the performance metrics of various methods across the three datasets, thereby highlighting their efficacy in this domain. The results presented within the GZSL framework offer insights into the comparative performance of diverse methods. While state-of-the-art methods excel in terms of their performance on seen classes, they often struggle with unseen classes. In contrast, our proposed FFusion method exhibits pronounced advantages in balancing the accuracy between seen and unseen classes. For instance, FFusion achieves a performance rate that exceeds the state of the art by 0.2% in terms of the harmonic mean (H) on CUB and delivers superior outcomes on the SUN dataset.

Table 1. Comparison of CZSL/GZSL experimental results between FFusion and other advanced zero-sample image classification methods on CUB, SUN, and AWA2 datasets.

Method	CUB				AWA2				SUN			
	acc (%)	*U* (%)	*S* (%)	*H* (%)	*acc* (%)	*U* (%)	*S* (%)	*H* (%)	*acc* (%)	*U* (%)	*S* (%)	*H* (%)
QFSL [13]	58.8	33.3	48.1	39.4	63.5	52.1	72.8	60.7	56.2	30.9	18.5	23.1
APN [26]	72.0	65.3	69.3	67.2	68.4	57.1	72.4	63.9	61.6	41.9	34.0	37.6
f-CLSWGAN [14]	57.3	43.7	57.7	49.7	68.2	57.9	61.4	59.6	60.8	42.6	36.6	39.4
Composer [46]	69.4	56.4	63.8	59.9	71.5	62.1	77.3	68.8	62.6	55.1	22.0	31.4
SP-AEN [47]	55.4	34.7	70.6	46.6	58.5	23.3	90.9	37.1	59.2	24.9	38.6	30.3
IIR [48]	63.8	30.4	65.8	41.2	67.9	17.6	87.0	28.9	63.5	22.0	34.1	26.7
TCN [28]	59.5	52.6	52.0	52.3	71.2	61.2	65.8	63.4	61.5	31.2	37.3	34.6
TFPL [20]	65.8	51.1	46.9	48.9	-	-	-	-	62.9	31.4	30.2	30.7
DVBE [21]	-	53.2	60.2	56.5	-	63.6	70.8	67.0	-	45.0	37.2	40.7
DCR-GAN [49]	61.0	55.8	66.8	60.8	-	-	-	-	63.7	47.1	38.5	42.4
DAZLE [50]	66.0	56.7	59.6	58.1	67.9	60.3	75.7	67.1	59.4	52.3	24.3	33.2
AREN [23]	71.8	38.9	78.7	52.1	67.9	15.6	92.9	26.7	60.6	19.0	38.8	25.5
LFGAA [51]	67.6	36.2	80.9	50.0	68.1	27.0	93.4	41.9	61.5	18.5	40.0	25.3
HSVA [52]	62.8	52.7	58.3	55.3	-	59.3	76.6	66.8	63.8	48.6	39.0	43.3
AREES [44]	65.7	53.6	56.9	55.2	73.6	57.9	77.0	66.1	64.3	51.3	35.9	42.2
DCA-VAE [53]	64.1	52.9	56.0	54.4	73.1	58.7	76.9	66.6	64.0	49.2	36.7	42.2
SupVFD [45]	65.1	55.8	60.3	58.0	72.6	64.9	77.1	70.5	65.6	50.6	36.8	42.6
BGSNet [43]	73.3	60.9	73.6	66.7	69.1	61.0	81.8	69.9	63.9	45.2	34.3	39.0
FFusion (Ours)	76.8	67.5	67.3	67.4	65.0	60.1	71.9	65.5	64.4	48.6	34.3	40.2

We attribute the advantages of FFusion primarily to the following two aspects. (1) Multiple feature fusion: The approach of multiple feature fusion in FFusion effectively integrates features from different sources. This approach not only bolsters the model's capacity

to discriminate across various classes but also refines its recognition abilities in the context of GZSL. (2) The debiasing loss function: The debiasing loss function introduced by FFusion effectively addresses the bias that the model may exhibit towards seen and unseen classes during training, thereby fostering fair learning and prediction capabilities for unseen classes. Based on experimental results from the dataset, TCN [28] exhibits the smallest accuracy disparity between seen and unseen classes at 0.6% in CUB, whereas our method shows an even smaller difference of just 0.2% and significantly outperforms TCN in terms of accuracy. This comparative analysis substantiates the capabilities of FFusion in achieving an equilibrium between the seen and unseen class accuracies. In the AWA2 dataset, the variance in accuracy between seen and unseen classes is markedly evident, typically varying from 15% to 20%, with the most pronounced gap reaching 77.3%. Despite the suboptimal balancing effect on AWA2, our method still maintains a higher accuracy rate in comparison to f-CLSWGAN [14], recognized for its optimal balancing performance. In light of these results, our method demonstrates greater potential for GZSL.

4.4. Ablation Experiment

To thoroughly investigate the individual contributions of the model's elements and features, we conduct an exhaustive ablation analysis. As depicted in Table 2, we assess the impact of the FFT, local features within the FFT (F_l), key local features within the FFT (F_{max}), mixed features from the FFT (F_{mix}), the attribute regression loss ($\mathcal{L}_{AR}$), and the debiasing loss ($\mathcal{L}_{DEB}$). Excluding the FFT results in markedly diminished model performance compared to the complete model, indicated by the 14.4%/12.7% reduction in the mean accuracy/harmonic mean (acc/H) on CUB. Furthermore, we create variations of the FFT by excluding each feature subset individually—local, key local, and mixed features. Each of these partial models performed less effectively than the holistic model, with the most notable decrease in performance observed in the absence of local features, resulting in a 12.1%/9.4% decrease in the mean acc/H on the CUB dataset. Utilizing the FFT without local features leads to a significant performance decline, as local features are pivotal in highlighting visual aspects that closely relate to semantic attributes. Moreover, the debiasing loss is instrumental in addressing bias; its removal results in a pronounced bias towards seen classes, with an accuracy gap of 20.2% between seen and unseen classes. Incorporating the debiasing loss function enhances the harmonic mean for CUB by 5.6%, substantially reducing the bias. Additionally, the attribute regression component refines the model performance by facilitating precise visual–semantic mapping. Compared to the baseline model, our enhanced model not only achieves a notable improvement in accuracy but also effectively addresses the issue of bias between seen and unseen classes. This baseline model comprises only the fundamental components, i.e., the semantic embedding module, the linear mapping module, and the original loss function, without any enhancements.

Table 2. Ablation experiment on CUB dataset.

	***acc* (%)**	***U* (%)**	***S* (%)**	***H* (%)**
Baseline	59.5	49.7	59.5	54.2
w/o FFT	62.4	55.1	54.3	54.7
w/o F_l	64.7	58.3	57.8	58.0
w/o F_{max}	76.1	66.2	67.0	66.6
w/o F_{mix}	75.3	65.9	66.8	66.3
w/o $\mathcal{L}_{AR}$	74.7	67.4	63.6	65.4
w/o $\mathcal{L}_{DEB}$	74.7	52.2	72.4	60.7
FFusion (full)	76.8	67.5	67.3	67.4

4.5. Hyperparametric Analysis

The impact of varying attribute regression loss weights, denoted as λ_{AR}, is illustrated in Figure 2. The parameter λ_{AR} quantifies the significance of the attribute regression loss in facilitating effective visual–semantic interaction for the model. It is designed to ensure the accurate prediction of object attributes within an image, including color, size, and shape. An extensive investigation of the λ_{AR} values on the CUB dataset has been conducted, encompassing $\lambda_{AR} = [0.001, 0.002, 0.003, 0.005, 0.01, 0.05]$. With a low λ_{AR} setting, the accuracy for unseen classes and the harmonic mean (H) tend to increase; this occurs because an overly small attribute regression loss does not sufficiently guide the model to learn attribute features, resulting in a suboptimal corrective influence on the model's attribute prediction accuracy. At a high λ_{AR} setting, the evaluation metrics are prone to decrease, as an overly large attribute regression loss leads to an excessive focus on the attribute regression task, overshadowing other critical loss functions. Taking into account the accuracies of U, S, and H collectively, the model achieves an optimal balance at $\lambda_{AR} = 0.003$, signifying that this value effectively mediates the learning process, neither disregarding the impact of other loss functions nor failing to efficiently direct the model towards attribute feature learning.

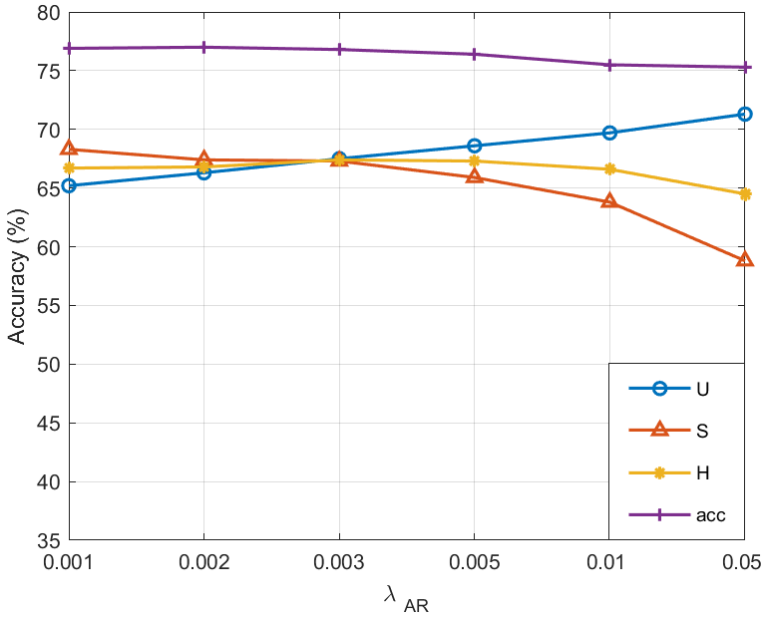

Figure 2. Hyperparameter λ_{AR} on CUB analysis.

The influence of the debiasing loss weight, denoted as λ_{DEB}, across various parameter settings is depicted in Figure 3. The parameter λ_{DEB} governs the impact of the debiasing loss, with an appropriate value enhancing its capacity to mitigate the bias between seen and unseen classes. We conduct an exhaustive examination of $\lambda_{DEB} = [0.001, 0.01, 0.1, 0.2, 0.3, 0.4]$ on the CUB dataset, aiming to identify the optimal balance. As the value of λ_{DEB} escalates, the accuracy for unseen classes increases, while that for seen classes declines. At a low λ_{DEB} setting, the accuracy for unseen classes markedly lags behind that for seen classes, as the debiasing loss's minimal value renders it ineffective in counteracting the model's inherent bias towards seen classes. At an elevated λ_{DEB} setting, the accuracy for unseen classes surpasses that for seen classes; this is a consequence of the exaggerated debiasing loss value, which disproportionately prioritizes unseen class outcomes, leading to an imbalance between the seen and unseen class accuracies. In light of the experimental outcomes, we have designated the debiasing loss weight λ_{DEB} as 0.2.

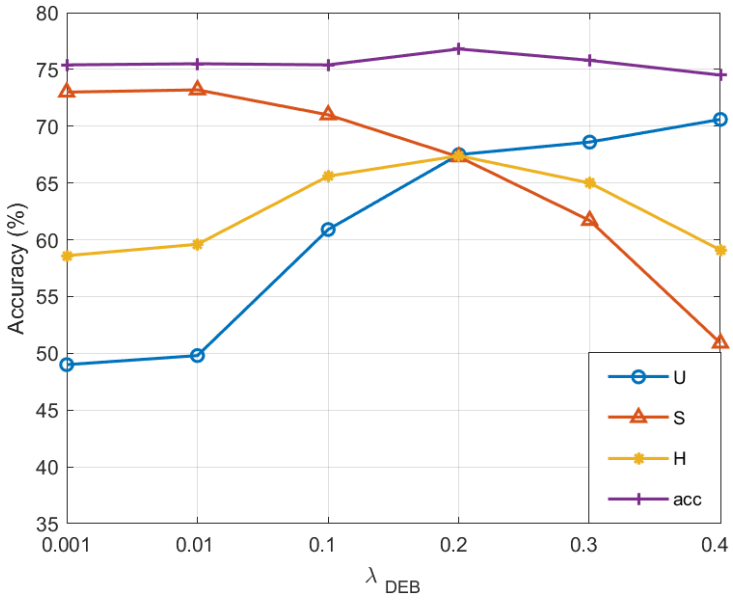

Figure 3. Hyperparameter λ_{DEB} on CUB analysis.

4.6. Qualitative Results

In this section, we randomly sample data from both the seen and unseen classes of the CUB dataset for a qualitative analysis with our pre-trained model. We identify the top five attributes with the highest scores for each instance and present their descriptions alongside corresponding images within each subfigure of Figure 4. On the seen class side, each sample possesses at least four attribute descriptions that align closely with the depicted bird's characteristics; for instance, the top five attributes for the 'Ivory Gull' exemplify this correspondence with the sample image. On the unseen class side, a minimum of three attribute descriptions per sample correspond to the bird's features within the sample images. Consequently, FFusion demonstrates substantial performance in attribute recognition across both seen and unseen classes.

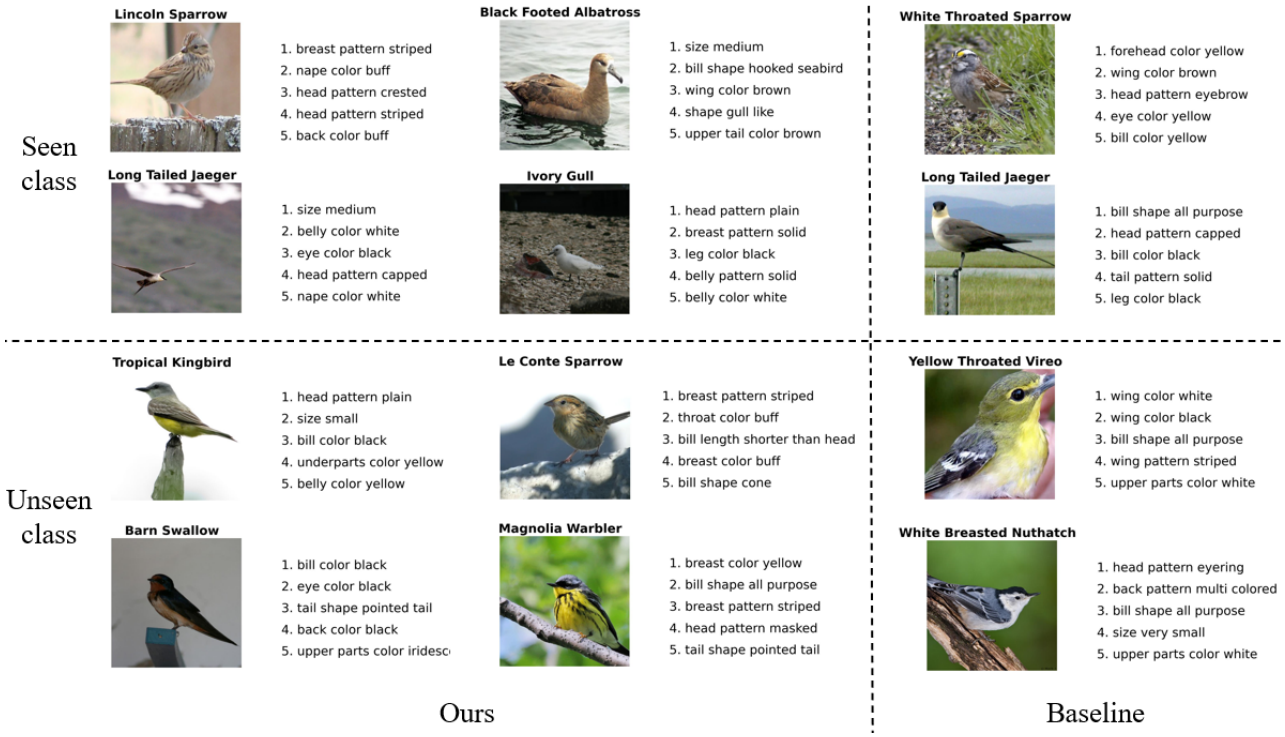

Figure 4. Qualitative result of FFusion.

In addition to this, we conducted a comparison with the unaltered baseline model. The results revealed that, while some attributes exhibit correlation, there were instances of attribute descriptive errors. For example, in the case of the white-throated sparrow, one of the attribute descriptions stated 'eye color yellow', which is factually incorrect. This comparison serves as a compelling indication of the effectiveness of the improvements that we have implemented in our model.

5. Discussion

The fundamental concept of ZSL lies in leveraging the semantic information of seen classes, such as category descriptions and attribute labels, to forge a mapping function that transitions from the semantic space to a visual space. This function enables the classification of categories during the testing phase, even if they were absent from the training phase. However, this process is fraught with challenges, including the ineffective representation of features extracted from images and the bias problem between seen and unseen classes. To address these issues, our FFusion model introduces a novel feature fusion module that amalgamates key local, mixed, and global features. This feature fusion approach within FFusion reduces the reliance on a singular feature source for different features to complement each other. Additionally, we propose a debiasing loss function designed to equitably distribute the model's focus between seen and unseen classes. This function diminishes the model's dependence on seen classes, thereby enhancing the recognition of unseen classes through judicious balancing.

Our method achieves 76.8% accuracy on the CUB dataset. However, it is limited to classifying only images of animals and scenes. In the future, we will consider applying ZSL to remote sensing and multispectral classification, which require higher classification accuracy, and extensively explore the potential of ZSL applications in these fields.

From the experimental results, we can see that the debiasing loss function performs better on CUB, but, at the same time, we notice that there is still room for improvement in the performance on the other two datasets, which provides us with a direction for our future research. We will continue to consider how to expand the scope of the debiasing loss function and achieve better results.

In addition to this, it is worth noting that some researchers have begun to apply ZSL models to video action recognition in addition to image classification [45]. The core idea of ZSL is equally applicable to video action recognition. However, despite some progress made in this area, as evidenced by the experimental results in the above-mentioned paper, there remains substantial research potential to further improve the performance and usefulness of ZSL models in the video action recognition task.

6. Conclusions

In this study, we introduce the FFusion for ZSL, addressing the limitation whereby individual features may fail to capture the full spectrum of information inherent in the original image. The FFT distills multi-dimensional features from both visual and semantic inputs, synthesizing these into a composite feature set that encapsulates the diverse attributes of the original features across various dimensions. Our method achieves accuracy of 76.8% in CUB. To mitigate the model's bias towards seen and unseen classes during the testing phase, we implement a debiasing loss function designed to ensure uniformity in the model's predictive distribution for both class types. Our model significantly reduces the bias between seen and unseen classes to a mere 0.2%. Experiments are conducted on three ZSL benchmark datasets to verify the effectiveness of the FFusion module components.

Author Contributions: Conceptualization, W.T.; methodology, W.T.; validation, W.T.; data curation, J.X.; writing—original draft, W.T. and Z.A.; writing—reviewing and editing, J.X. and X.M.; supervision, X.M.; funding acquisition, X.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Natural Science Foundation of China under Grant 62276211.

Data Availability Statement: The CUB dataset can be obtained via https://www.vision.caltech.edu/datasets/cub_200_2011/ [41]; the AWA2 dataset can be obtained via https://cvml.ista.ac.at/AwA2/ [17]; and the SUN dataset can be obtained via http://groups.csail.mit.edu/vision/SUN/ [42], accessed on 24 February 2025.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.-F. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
2. Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; Wang, B. Moment matching for multi-source domain adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1406–1415.
3. Fan, D.P.; Wang, W.; Cheng, M.M.; Shen, J. Shifting more attention to video salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8554–8564.
4. Romera-Paredes, B.; Torr, P.H.S. An embarrassingly simple approach to zero-shot learning. In Proceedings of the 32nd International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015; JMLR.org; pp. 2152–2161.
5. Buckchash, H.; Raman, B. Towards zero shot learning of geometry of motion streams and its application to anomaly recognition. *Expert Syst. Appl.* **2021**, *177*, 114916. [CrossRef]
6. Xu, J.; Zhou, L.; Zhao, W.; Fan, Y.; Ding, X.; Yuan, X. Zero-shot learning for compound fault diagnosis of bearings. *Expert Syst. Appl.* **2022**, *190*, 116197. [CrossRef]
7. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.H.; Hospedales, T.M. Learning to compare: Relation network for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1199–1208.
8. Zhu, Y.; Elhoseiny, M.; Liu, B.; Peng, X.; Elgammal, A. A generative adversarial approach for zero-shot learning from noisy texts. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1004–1013.
9. Palatucci, M.; Pomerleau, D.; Hinton, G.E.; Mitchell, T.M. Zero-shot learning with semantic output codes. In Proceedings of the 23rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 7–10 December 2009; pp. 1410–1418.
10. Huynh, D.; Elhamifar, E. A shared multi-attention framework for multi-label zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8776–8786.
11. Vyas, M.R.; Venkateswara, H.; Panchanathan, S. Leveraging Seen and Unseen Semantic Relationships for Generative Zero-Shot Learning. In Proceedings of the Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 70–86.
12. Li, X.; Xu, Z.; Wei, K.; Deng, C. Generalized zero-shot learning via disentangled representation. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; Volume 35, pp. 1966–1974.
13. Song, J.; Shen, C.; Yang, Y.; Liu, Y.; Song, M. Transductive unbiased embedding for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1024–1033.
14. Xian, Y.; Lorenz, T.; Schiele, B.; Akata, Z. Feature generating networks for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5542–5551.
15. Chen, S.; Wang, W.; Xia, B.; Peng, Q.; You, X.; Zheng, F.; Shao, L. Free: Feature refinement for generalized zero-shot learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 122–131.
16. Han, Z.; Fu, Z.; Chen, S.; Yang, J. Contrastive embedding for generalized zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2371–2381.
17. Xian, Y.; Lampert, C.H.; Schiele, B.; Akata, Z. Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 2251–2265. [CrossRef] [PubMed]

18. Akata, Z.; Perronnin, F.; Harchaoui, Z.; Schmid, C. Label-embedding for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 1425–1438. [CrossRef] [PubMed]
19. Zhang, F.; Shi, G. Co-representation network for generalized zero-shot learning. In Proceedings of the 36th International Conference on Machine Learning, Los Angeles, CA, USA, 9–15 June 2019; pp. 7434–7443.
20. Li, A.; Lu, Z.; Guan, J.; Xiang, T.; Wang, L.; Wen, J.R. Transferrable feature and projection learning with class hierarchy for zero-shot learning. *Int. J. Comput. Vis.* **2020**, *128*, 2810–2827. [CrossRef]
21. Min, S.; Yao, H.; Xie, H.; Wang, C.; Zha, Z.J.; Zhang, Y. Domain-aware visual bias eliminating for generalized zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12664–12673.
22. Yu, Y.; Ji, Z.; Fu, Y.; Guo, J.; Pang, Y.; Zhang, Z.M. Stacked semantics-guided attention model for fine-grained zero-shot learning. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, QC, Canada, 2–8 December 2018; pp. 5998–6007.
23. Xie, G.S.; Liu, L.; Jin, X.; Zhu, F.; Zhang, Z.; Qin, J.; Yao, Y.; Shao, L. Attentive region embedding network for zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9384–9393.
24. Xie, G.S.; Liu, L.; Zhu, F.; Zhao, F.; Zhang, Z.; Yao, Y.; Qin, J.; Shao, L. Region graph embedding network for zero-shot learning. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Part IV 16, pp. 562–580.
25. Zhu, Y.; Xie, J.; Tang, Z.; Peng, X.; Elgammal, A. Semantic-guided multi-attention localization for zero-shot learning. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 14943–14953.
26. Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; Akata, Z. Attribute prototype network for zero-shot learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21969–21980.
27. Verma, V.K.; Arora, G.; Mishra, A.; Rai, P. Generalized zero-shot learning via synthesized examples. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4281–4289.
28. Jiang, H.; Wang, R.; Shan, S.; Chen, X. Transferable contrastive network for generalized zero-shot learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9765–9774.
29. Akata, Z.; Reed, S.; Walter, D.; Lee, H.; Schiele, B. Evaluation of output embeddings for fine-grained image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2927–2936.
30. Socher, R.; Ganjoo, M.; Manning, C.D.; Ng, A. Zero-shot learning through cross-modal transfer. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 935–943.
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
32. Huang, L.; Wang, W.; Chen, J.; Wei, X.Y. Attention on attention for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4634–4643.
33. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10578–10587.
34. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, *arXiv:2010.11929*.
35. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 10347–10357.
36. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 568–578.
37. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
38. Yu, Y.; Ji, Z.; Han, J.; Zhang, Z. Episode-based prototype generating network for zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14035–14044.
39. Chen, S.; Hong, Z.; Liu, Y.; Xie, G.; Sun, B.; Li, H.; Peng, Q.; Lu, K.; You, X. Transzero: Attribute-guided transformer for zero-shot learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; pp. 330–338.
40. Liu, M.; Li, F.; Zhang, C.; Wei, Y.; Bai, H.; Zhao, Y. Progressive Semantic-Visual Mutual Adaption for Generalized Zero-Shot Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 15337–15346.

41. Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Perona, P. *Caltech-UCSD Birds 200*; California Institute of Technology: Pasadena, CA, USA, 2010.
42. Patterson, G.; Hays, J. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2751–2758.
43. Li, Y.; Liu, Z.; Chang, X.; McAuley, J.; Yao, L. Diversity-Boosted Generalization-Specialization Balancing for Zero-Shot Learning. *IEEE Trans. Multimed.* **2023**, *25*, 8372–8382. [CrossRef]
44. Liu, Y.; Dang, Y.; Gao, X.; Han, J.; Shao, L. Zero-Shot Learning with Attentive Region Embedding and Enhanced Semantics. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 4220–4231. [CrossRef] [PubMed]
45. Niu, C.; Shang, J.; Zhou, Z.; Yang, J. Superclass-aware visual feature disentangling for generalized zero-shot learning. *Expert Syst. Appl.* **2024**, *258*, 125150. [CrossRef]
46. Huynh, D.; Elhamifar, E. Compositional zero-shot learning via fine-grained dense feature composition. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 19849–19860.
47. Chen, L.; Zhang, H.; Xiao, J.; Liu, W.; Chang, S.F. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1043–1052.
48. Cacheux, Y.L.; Borgne, H.L.; Crucianu, M. Modeling inter and intra-class relations in the triplet loss for zero-shot learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10333–10342.
49. Ye, Z.; Hu, F.; Lyu, F.; Li, L.; Huang, K. Disentangling semantic-to-visual confusion for zero-shot learning. *IEEE Trans. Multimed.* **2021**, *24*, 2828–2840. [CrossRef]
50. Huynh, D.; Elhamifar, E. Fine-grained generalized zero-shot learning via dense attribute-based attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4483–4493.
51. Liu, Y.; Guo, J.; Cai, D.; He, X. Attribute attention for semantic disambiguation in zero-shot learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6698–6707.
52. Chen, S.; Xie, G.; Liu, Y.; Peng, Q.; Sun, B.; Li, H.; You, X.; Shao, L. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 16622–16634.
53. Liu, Y.; Gao, X.; Han, J.; Shao, L. A Discriminative Cross-Aligned Variational Autoencoder for Zero-Shot Learning. *IEEE Trans. Cybern.* **2023**, *53*, 3794–3805. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.