Depth-Aware Multi-Modal Fusion for Generalized Zero-Shot Learning

Weipeng Cao[1,2], Xuyang Yao[3], Zhiwu Xu[3], Yinghui Pan[2], Yixuan Sun[4], Dachuan Li[5,6,*], Bohua Qiu[7,8,9], Muheng Wei[7]

[1]*Guangdong Laboratory of Artificial Intelligence and Digital Economy (Shenzhen), Shenzhen, China*
[2]*National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen, China*
[3]*College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China*
[4]*Stony Brook University, New York, United States*
[5]*Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology, Shenzhen, China*
[6]*Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China*
[7]*ZhenDui Industry Artificial Intelligence Co. Ltd, Shenzhen, China*
[8]*Department of Automation, Shanghai Jiao Tong University, Shanghai, China*
[9]*Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, China*
*Corresponding author. Emails: lidc3@mail.sustech.edu.cn

2024 IEEE 22nd International Conference on Industrial Informatics (INDIN) | 979-8-3315-2747-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/INDIN58382.2024.10774475

***Abstract*—Realizing Generalized Zero-Shot Learning (GZSL) based on large models is emerging as a prevailing trend. However, most existing methods merely regard large models as black boxes, solely leveraging the features output by the final layer while disregarding potential performance enhancements from other layers. Indeed, numerous researchers have visually depicted variations in the features learned across different layers of neural networks. Motivated by this observation, we propose a Vision Transformer (ViT)-based GZSL method named Depth-Aware Multi-Modal ViT (DAM2ViT), which exploits multi-level features of ViT. DAM2ViT incorporates a multi-modal interaction block to align semantic information of categories across multiple layers, thereby augmenting the model's capacity to learn associations between visual and semantic spaces. Extensive experiments conducted on three benchmark datasets (i.e., CUB, SUN, AWA2) have showcased that DAM2ViT achieves competitive results compared to state-of-the-art methods.**

***Index Terms*—Generalized Zero-Shot Learning, Multi-Modal, Deep Learning**

I. Introduction

Deep learning technology has found extensive applications across various domains, encompassing computer vision (CV) [1], natural language processing (NLP) [2], autonomous driving vehicles [3], and more. However, the majority of existing deep learning models exhibit limitations in recognizing only those categories encountered during the training phase. This limitation presents formidable challenges when applied to open domains in real-world scenarios, where the categories present may extend beyond those included in the training set. To address this issue, Generalized Zero-Shot Learning (GZSL) technology has been introduced and garnered considerable attention. Currently, the prominent implementation pathways for GZSL comprise embedding and generative-based methods.

This study was supported by National Natural Science Foundation of China (62106150, 52272419), the Open Fund of National Engineering Laboratory for Big Data System Computing Technology (SZU-BDSC-OF2024-22), Director Fund of Guangdong Laboratory of Artificial Intelligence and Digital Economy (Shenzhen) (24420001), and Shenzhen Science and Technology Program (KJZD20230923114220042).

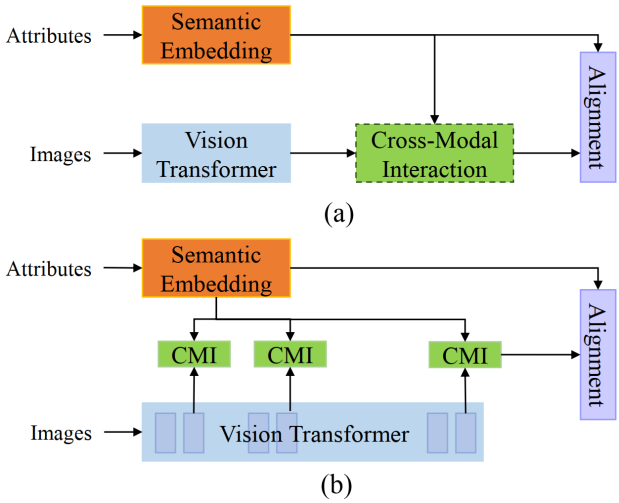

Fig. 1. (a) Conventional Transformer-based approaches, and the dashed box suggests that it may not always be present. These approaches typically rely on processing the deepest visual features. (b) Our proposed approach, where CMI stands for Cross-Modal Interaction. We utilize visual features from different depths for multi-modal interaction.

Recently, several large-scale foundation models have demonstrated significantly improved performance across numerous CV and NLP tasks compared to traditional end-to-end trained deep learning models [4]. From the fundamental standpoint of machine learning, these models have enhanced representation learning capabilities facilitated by extensive data, robust computing resources, and sophisticated algorithms, consequently enhancing the generalization abilities of downstream classifiers and regressors. Enhancing the performance of domain models built upon large foundation models has emerged as a prominent research focus. This study specifically delves into the GZSL-based models.

As elucidated earlier, embedding-based and generative

model-based methods have garnered considerable success. However, the advent of the Vision Transformer (ViT) [4] and the exploration of multi-modal large models have reinvigorated fresh perspectives for embedding-based approaches. Consequently, a plethora of Transformer-based studies have emerged [5], [6]. These endeavors aim to delve deeper into spatial information within visual features, along with attribute matching with visual data. By harnessing pre-trained ViT models and fostering interaction among diverse modalities, these studies have enhanced the accuracy of GZSL models.

While multi-modal large models such as Contrastive Language-Image Pre-training (CLIP) [7] and Segment Anything Mode (SAM) [8] exhibit remarkable generalization across diverse datasets owing to their extensive training on billion-level data, it is noteworthy that they are prone to encountering dataset overlap issues. Consequently, these models tend to lean towards open-vocabulary recognition methods rather than traditional zero-shot approaches. To address this challenge, ViT-based methods often choose to pre-train models on datasets like ImageNet-21K or even ImageNet-1K, aligning more closely with the principles of traditional ZSL.

Nevertheless, these approaches have frequently concentrated their efforts on the deepest layers of ViT, overlooking the potential impact of shallower features on matching performance. Drawing from human cognitive experience, observing an object from multiple dimensions aids in its more comprehensive and accurate identification.

Inspired by this observation, we embark on an exploration of the influence of different ViT layer features on model accuracy in this study. The disparity between previous ViT-based methods and our approach is illustrated in Fig. 1. In recent years, cross-modal feature fusion has garnered significant attention not only in zero-shot image recognition but also across various other domains. Nonetheless, previous zero-shot approaches predominantly rely on pre-trained visual and text models to generate separate features before fusion, a methodology known as late fusion. Thus, this paper delves into investigating the impact of early fusion on the recognition efficacy of the models.

Specifically, we introduce the Multi-Modal Interaction Block (M2IB), which takes visual features and semantic embeddings as inputs and yields multi-modal features. Subsequently, we utilize the attribute embeddings corresponding to categories as supervised signals to guide model training. Finally, we integrate the proposed M2IB to facilitate interaction across various layers of ViT, thereby examining its impact on widely used GZSL datasets. Drawing insights from the data derived from our experiments, we present our ultimate model: the Depth-Aware Multi-Modal ViT (DAM2ViT), which achieves State-of-the-Art (SOTA) results across a range of commonly employed GZSL datasets.

Our contributions can be summarized as follows:

- We introduce the M2IB as a facilitator for cross-layer multi-modal interactions, acting as a conduit for models to effectively utilize and communicate visual representations across various layers.
- Our proposed attribute regression loss eliminates the need for M2IB to engage in semantic embedding to attribute vector conversion. This reduction in unnecessary modal transformations enables M2IB to focus more on fostering inter-modal interactions.
- We investigate the impact of utilizing different layers of visual features on model accuracy across diverse datasets. Drawing upon these insights, we introduce the DAM2ViT, which achieves comparable results with SOTA methods.

II. Method

A. Preliminaries

GZSL models encompass the assimilation of knowledge from familiar classes to effectively tackle unfamiliar ones. We define the set of seen classes as $D_s = \{x, y, a_y | x \in X^s, y \in Y^s, a_y \in A^s\}$, where X^s, Y^s, and A^s represent the set of images, class label, and attributes associated with the known categories, respectively. Likewise, we establish the set of unknown classes as $D_u = \{x, y, a_y | x \in X^u, y \in Y^u, a_y \in A^u\}$, where X^u, Y^u, and A^u denote the set of images, categories, and attributes of the unknown classes, respectively. Noteworthy is the categorical distinction between the sets of seen and unseen classes, which are exhaustive yet mutually exclusive ($Y^s \cap Y^u = \emptyset, Y^s \cup Y^u = Y$), where Y denotes the set of all categories. Similarly, the attribute sets embody the characteristic that their collective union encapsulates all attributes ($A^s \cup A^u = A$), where A denotes the set of all attributes.

In this paper, following PSVAM [6], we utilize GloVe [9] embeddings to encode the textual attributes, yielding the corresponding embedding denoted as S.

B. Overview

Our innovative model, the DAM2ViT, integrates RGB images (represented as $X \in \mathbb{R}^{C \times H \times W}$) and attributes encoded as word embeddings $S \in \mathbb{R}^{1 \times K \times T}$ as its primary inputs. These word embeddings are constructed using GloVe, where K signifies the number of attributes in the datasets, and T denotes the dimensionality of the word vectors determined by GloVe. This approach aligns seamlessly with the methodology elucidated in Progressive Semantic-visual Mutual Adaption (PSVMA) [6].

In DAM2ViT, each transformer layer of ViT is responsible for generating visual features. Using the notations B, N, and L to denote the batch size, the number of tokens, and the token length, respectively, and $i = 1, 2, ..., I$ where I represents the total number of layers in ViT, these features are computed via cross attention with respect to S employing the proposed M2IB. This process yields $\hat{F}^i \in \mathbb{R}^{B \times N \times L}$ and $\hat{S} \in \mathbb{R}^{B \times K \times T}$. Subsequently, the obtained $\hat{F}^i$ is transformed into an attribute vector, denoted as $\hat{a_y}$, through pooling and a projection head.

C. Multi-Modal Interaction Block

The M2IB within our framework serves as a pivotal component facilitating the intricate interactions between visual and

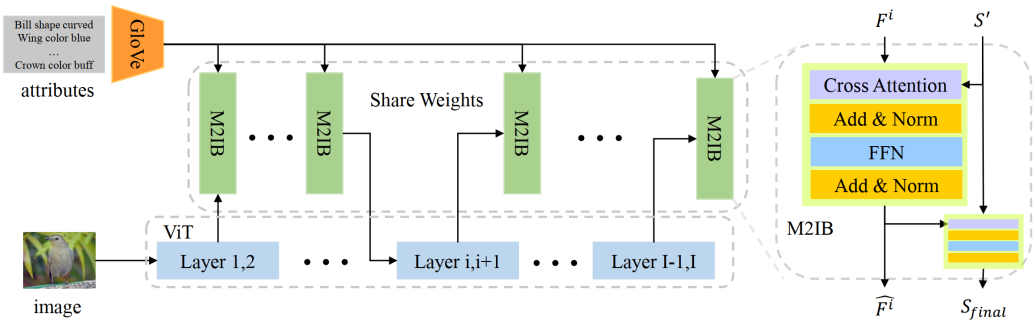

Fig. 2. Architecture of our proposed model showcasing the integration of M2IB within the ViT framework to foster intricate multi-modal interactions. The output features of M2IB serve as pivotal inputs for both the downstream ViT layer and the computation of the attribute regression loss, ensuring comprehensive information integration and effective attribute learning.

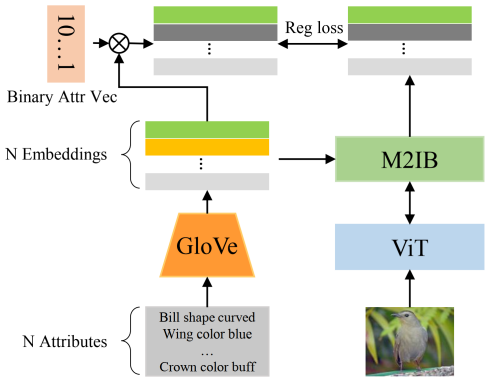

Fig. 3. Process of calculating the attribute regression loss.

semantic modalities. M2IB operates by taking visual features (F^i) extracted from a specific layer of the ViT and maps the word embeddings representing all attributes (S') as its inputs. Initially, M2IB employs cross-attention, wherein visual features act as queries and word embeddings as keys and values. This mechanism enables the queries to retrieve pertinent attribute information, thus establishing a robust linkage between the visual and semantic domains. Subsequently, token fusion is achieved through a Feed-Forward Network (FFN), generating multi-modal intermediate features (F^i_{inter}).

Following this, another cross-attention module is employed, utilizing S' as queries and F^i_{inter} as keys and values. This stage focuses on semantically embedding the query to discern attributes within the visual features more effectively. Once again, FFN is applied for token fusion, culminating in the production of the final multi-modal feature (S_{final}). Notably, when the input visual feature to M2IB originates from the final ViT layer ($i = I$), F^i_{inter} computes the attribute vector. Conversely, for other layers, F^i_{inter} is passed on to the subsequent ViT layer for further processing, ensuring a comprehensive integration of multi-modal information across the network.

Attribute Regression Loss. Following the computation of S_{final} through the M2IB, we employ a linear layer to map it to the same length as the embedding vector, resulting in $\hat{S} = f_{invs}(S_{final}) \in \mathbb{R}^{B\times K\times T}$. Simultaneously, we utilize the binary attribute vector label a_y to perform element-wise multiplication with the mapped word embeddings S, which is shown in Fig 3. Any attribute word embeddings not associated with that category are set to 0, yielding the word vector label $S_{label} = a_y \cdot S$ corresponding to that specific category. Subsequently, our attribute regression loss is defined as:

$$L_{reg} = ||\hat{S} - S_{label}||_1 \tag{1}$$

D. Depth-Aware Multi-Modal ViT

Prior research has predominantly concentrated on harnessing solely the deepest visual features extracted from the ViT for multi-modal interaction with semantic features, often overlooking the potential impact of shallower features on model classification. Recognizing this limitation, we introduce the innovative DAM2ViT framework (depicted in Fig. 2), which judiciously incorporates visual features from various layers simultaneously.

DAM2ViT achieves this by seamlessly integrating visual features from different depths through the strategic insertion of weight-sharing M2IB between multiple discrete ViT layers, thereby fostering effective multi-modal interactions. Moreover, each multi-modal interaction engenders an L_{reg}, and the cumulative influence of these regression losses collectively contributes to the model's final classification outcomes.

$$L_{regT} = \sum_{i\in\hat{I}} ||\sigma_i \hat{S}^i - S_{final}||_1 \tag{2}$$

Here, $\hat{S}^i$ signifies the outcome of the computation involving S and F^i utilizing the M2IB. The set $\hat{I}$ denotes the ViT layer subsequent to the insertion of an M2IB. We extensively explore the strategic placement of M2IB and the challenge of overfitting in Section 3.

Moreover, σ_i denotes the weights assigned to attribute regression losses corresponding to the i-th layer. These weights serve to assess the compatibility between visual features from various layers and text features. In our methodology, we render these weights trainable and constrain their sum to a predetermined value. This constraint is crucial as without it, these parameters tend to converge to excessively small values during training.

TABLE I
OVERVIEW OF COMMONLY USED BENCHMARKS FOR GZSL TASK.

Dataset	Classes	Images	Attributes
	total/seen/unseen		
CUB	200 / 150 / 50	11788	312
SUN	717 / 645 / 72	14340	102
AWA2	50 / 40 / 10	37322	85

To tackle the overfitting issue arising from multiple M2IB insertions, we adopt the strategy of directly eliminating the Penultimate Layer (PL) of ViT. While this approach effectively mitigates the overfitting concern, it also entails a certain degree of performance degradation.

E. Optimization and Inference

In our pursuit of computing the classification loss, mitigating unseen bias, and facilitating GZSL inference, we draw inspiration from the pioneering work presented in PSVMA, where we leverage L_{cls}, L_{deb}, and a dedicated inference methodology. The formulation of L_{deb} is articulated as follows:

$$L_{deb} = ||\alpha_s - \alpha_u||_2^2 + ||\beta_s - \beta_u||_2^2 \tag{3}$$

where α_s and β_s denote the mean and variance of predictions for seen class scores, respectively, while α_u and β_u correspond to the mean and variance of predictions for unseen class scores.

Our ultimate loss function is defined as:

$$L_{total} = L_{cls} + \lambda_1 L_{regT} + \lambda_2 L_{deb} \tag{4}$$

where λ_1 and λ_2 represent hyper-parameters governing the contribution of the different loss components.

In terms of inference, it is computed as follows:

$$\tilde{y} = argmax_{y_c \in Y^S \cup Y^U}(\hat{y} - \gamma \mathbb{I}_{Y^S}(y_c)) \tag{5}$$

where $\mathbb{I}_{Y^S}$ is an indicator function that equals 1 if y_c belongs to the seen class set Y^S, and 0 otherwise.

III. EXPERIMENTS

A. Experiment Settings

Datasets. In this study, we employ three commonly utilized datasets for the GZSL task: CUB [23], SUN [24], and AWA2 [25]. Table I provides a concise summary of the categories, attributes, and image samples associated with these datasets.

Metrics. We utilize common GZSL metrics, including the Top-1 accuracy for seen classes (S) and unseen classes (U), and the harmonic mean of the two ($H = (2*S*U)/(S+U)$), to assess the performance of each method on different datasets.

Implement Details. We adhere to the identical hyperparameter configurations as outlined in PSVMA and utilize a ViT-base model pre-trained on ImageNet-1k. This approach ensures the avoidance of incomplete GZSL experimental setups.

B. Comparison Study

We conducted an extensive comparison of our proposed DAM2ViT with five generative-based and ten embedding-based methods, and the experimental findings are succinctly presented in Table II. Here, 'Generative' and 'Embedding' designate generative-based and embedding-based methods, respectively. Furthermore, 'PL' in the table denotes the penultimate layer of the ViT. The proposed DAM2ViT model outperforms other compared methods across multiple metrics: achieving a remarkable 73.3% unseen class accuracy (U) and 74.5% harmonic mean score (H) on the CUB dataset. On the SUN dataset, it secures a noteworthy 61.1% U score, second only to PSVMA, while its 50.5% H score positions it as the third-best performer among the compared methods. Additionally, on the AWA2 dataset, DAM2ViT attains an impressive 75.0% U-score, surpassing other comparison methods, and a commendable 73.8% H-score, placing it third among the comparison methods. From this table, it is evident that our method exhibits superior generalization ability compared to other methods, while maintaining similar overall performance.

C. Ablation Study

Quantitative Results.

The comparison experiment results are outlined in Table III. In our experimental setup, we utilize a ViT-based model and designate each transformer layer with a specific position where the M2IB can potentially be inserted. To facilitate this, we adopt a binary encoding scheme, where 0/1 denotes the absence/presence of M2IB insertion, while 'D' represents the discarding of the ViT layer. Consequently, our model can be effectively represented by a 12-bit binary sequence comprising 0s, 1s, and Ds. It's important to note that the all-0 sequence corresponds to our baseline configuration. Moreover, the presence of a '+' at the end of this sequence indicates that the weights σ_i in Equation 2 are learnable parameters, whereas their absence suggests that all σ_i are set to 1.

From Table III, several key insights emerge: (1) The incorporation of M2IB and attribute regression loss indeed yields performance enhancements for the model. However, the magnitude of improvement is intricately linked to the placement and quantity of M2IB insertions. (2) In terms of overall performance enhancement, particularly regarding generalization ability, the location of insertion emerges as paramount, superseding the significance of the number of M2IB insertions. Furthermore, integrating additional learnable weights σ_i into the loss function based on DAM2ViT can augment the model's generalization ability. For instance, on the CUB and AWA2 datasets, this modification results in further performance enhancements. However, on the SUN dataset, it presents challenges for the model in learning complex features.

Furthermore, in Fig. 3, we present a heatmap visualization of the attention patterns for both the Baseline and the first five ViT layers of DAM2ViT (without PL). The heatmap illustrates that both models effectively focus on the object itself. However, while the Baseline exhibits a fixed region of interest within the object, our proposed model demonstrates

TABLE II
COMPARISON RESULTS OF DIFFERENT GZSL APPROACHES ON TESTING BENCHMARKS.

Type	Methods	CUB			SUN			AWA2		
		U	S	H	U	S	H	U	S	H
Generative	CE-GZSL (CVPR'21) [10]	63.9	66.8	65.3	48.8	38.6	43.1	63.1	78.6	70.0
	FREE (ICCV'21) [11]	55.7	59.9	57.7	47.4	37.2	41.7	60.4	75.4	67.1
	HSVA (NeurIPS'21) [12]	52.7	58.3	55.3	48.6	39.0	43.3	56.7	79.8	66.3
	ICCE (CVPR'22) [13]	67.3	65.5	66.4	-	-	-	65.3	82.3	72.8
	SC-EGG (IJCAI'22) [14]	64.1	73.6	68.5	45.1	43.6	44.3	60.9	89.3	72.4
Embedding	GEM-ZSL (CVPR'21) [15]	64.8	77.1	70.4	38.1	35.7	36.9	64.8	77.5	70.6
	ViT-ZSL (arXiv) [16]	67.3	75.2	71.0	44.5	55.3	49.3	51.9	90.0	68.5
	DPPN (NeurIPS'21) [17]	70.2	77.1	73.5	47.9	35.8	41.0	63.1	86.8	73.1
	IEAM-ZSL (DGAM'21) [18]	68.6	73.8	71.1	48.2	54.7	51.3	53.7	89.9	67.2
	TransZero (AAAI'22) [19]	69.3	68.3	68.8	52.6	33.4	40.8	61.3	82.3	70.2
	MSDN (CVPR'22) [20]	68.7	67.5	68.1	52.2	34.2	41.3	62.0	74.5	67.7
	DUET (AAAI'23) [5]	62.9	72.8	67.5	45.7	45.8	45.8	63.7	84.7	72.7
	PSVMA (CVPR'23) [6]	70.1	77.8	73.8	61.7	45.3	52.3	73.6	77.3	75.4
	DPN (TCSVT'23) [21]	63.7	80.6	71.2	48.3	41.4	44.5	65.2	87.6	74.8
	EMP (MM'23) [22]	70.8	78.4	74.4	47.4	36.0	40.9	62.1	87.5	72.7
	DAM2ViT (Ours)	73.3	75.7	74.5	61.1	42.9	50.5	75.0	72.6	73.8

Note: We provide the Top-1 accuracy results for unseen and seen classes, and their harmonic mean, denoted as U, S, and H, respectively. The best and second scores are highlighted in red, and blue respectively.

TABLE III
EXPERIMENT RESULTS WITH DIFFERENT COMBINATIONS OF M2IB INSERTIONS.

Model	CUB			SUN			AWA2		
	U	S	H	U	S	H	U	S	H
0000 0000 0000	64.99	**80.49**	71.91	59.72	48.60	53.59	62.84	79.13	70.05
0000 0000 0001	67.74	77.07	72.10	59.51	48.72	53.58	64.21	78.10	70.53
0000 0000 0011	67.81	77.80	72.46	60.28	49.34	**54.26**	64.68	78.13	70.77
0000 0001 0011	67.98	77.75	72.53	61.67	43.06	50.71	73.92	70.78	72.31
0000 0100 0001	67.56	77.74	72.29	60.28	49.03	54.08	65.18	78.49	71.22
0000 0000 0111	67.63	77.52	72.24	59.58	**49.38**	54.00	64.56	78.47	70.84
0001 0001 0001	68.22	77.83	72.71	60.83	48.72	54.11	65.44	78.90	71.54
0010 0100 1001	68.35	78.83	73.00	60.35	48.91	54.03	67.50	78.62	72.64
0101 0101 0101	66.40	80.20	72.68	60.07	49.22	54.11	68.29	**79.63**	73.53
0101 0101 01D1	72.76	76.11	74.40	**62.08**	42.98	50.80	74.41	70.62	72.46
0101 0101 01D1 +	**73.33**	75.71	**74.50**	61.18	42.95	50.55	**75.04**	72.66	**73.83**

the capacity to concentrate on different information locations across various layers, significantly enhancing its classification capabilities. **Analysis.** Regarding the quantization and visualization results, we propose the following hypotheses: Firstly, the placement of M2IB at different positions can influence the ability of various ViT layers to capture distinct attributes during image processing. However, the exact nature of this impact remains unclear. We speculate that the insertion of M2IB may lead to two possible outcomes: (1) The chosen insertion point prompts attribute representations in different ViT layers to diverge, resulting in increased accessible attribute information for M2IB. This, in turn, enhances M2IB's capacity to discriminate between different attributes, ultimately improving the model's ability to transfer knowledge. This improvement manifests in the form of enhanced generalization. (2) Conversely, the insertion location could lead to significant redundancy in attribute representations across different ViT layers. This redundancy may diminish the effective attribute information available to M2IB, causing the model to disproportionately focus on a subset of attributes while neglecting others. This simplifies the classification process, potentially resulting in overfitting issues.

In summary, we postulate that the choice of insertion location for M2IB plays a crucial role in shaping the behavior of the model with respect to attribute representation and classification performance. Further experimentation is needed to validate these conjectures and gain deeper insights into the effects of insertion positions on model behavior.

IV. CONCLUSIONS

In this paper, we introduce the M2IB with an attribute regression loss to facilitate cross-modal interactions between

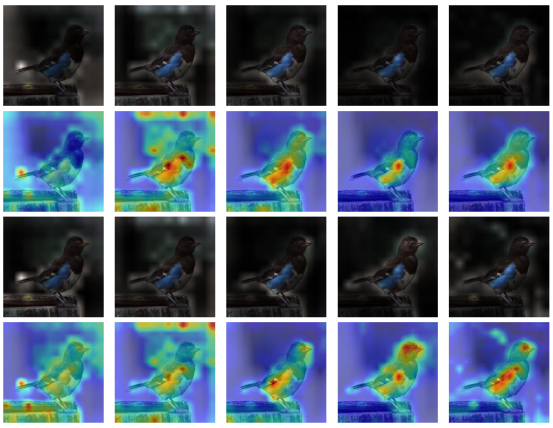

Fig. 4. Evolution of the Attention Map for the initial five layers of ViT.The Attention Maps and Heat Maps for Baseline and DAM2ViT (w/o PL) are listed from top to bottom, respectively. The incorporation of M2IB aids the model in achieving a more refined focus on the target object itself.

visual and semantic features, thereby enhancing the optimization of the model. Additionally, we propose a family of models referred to as DAM2ViT, which are constructed upon the integration of M2IB. Through our experimental evaluations, we have demonstrated the effectiveness of M2IB in improving the model's generalization performance. Furthermore, we have discovered that visual features generated by different layers of the ViT may exhibit either redundancy or complementarity. Building upon this insight, we explore the potential for further enhancing the model's capabilities by employing diverse arrangements of M2IB.

Moreover, we observed that different combinations of layers in the foundation model yield inconsistent performance improvements for the GZSL model, suggesting that the optimal layer combination method may be problem-specific. In the future, we plan to explore the integration of meta-parameters or domain knowledge related to modeling problems [26] to design adaptive layer combination methods.

REFERENCES

[1] W. Cao, Y. Wu, C. Huang, M. J. Patwary, and X. Wang, "Mff: Multi-modal feature fusion for zero-shot learning," *Neurocomputing*, vol. 510, pp. 172–180, 2022.

[2] W. Cao, C. Zhou, Y. Wu, Z. Ming, Z. Xu, and J. Zhang, "Research progress of zero-shot learning beyond computer vision," in *Algorithms and Architectures for Parallel Processing: 20th International Conference, ICA3PP 2020, New York City, NY, USA, October 2–4, 2020, Proceedings, Part II 20*. Springer, 2020, pp. 538–551.

[3] W. Cao, Y. Wu, C. Chakraborty, D. Li, L. Zhao, and S. K. Ghosh, "Sustainable and transferable traffic sign recognition for intelligent transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[5] Z. Chen, Y. Huang, J. Chen, Y. Geng, W. Zhang, Y. Fang, J. Z. Pan, and H. Chen, "Duet: Cross-modal semantic grounding for contrastive zero-shot learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 405–413.

[6] M. Liu, F. Li, C. Zhang, Y. Wei, H. Bai, and Y. Zhao, "Progressive semantic-visual mutual adaption for generalized zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 15 337–15 346.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[8] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[9] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[10] Z. Han, Z. Fu, S. Chen, and J. Yang, "Contrastive embedding for generalized zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 2371–2381.

[11] S. Chen, W. Wang, B. Xia, Q. Peng, X. You, F. Zheng, and L. Shao, "Free: Feature refinement for generalized zero-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 122–131.

[12] S. Chen, G. Xie, Y. Liu, Q. Peng, B. Sun, H. Li, X. You, and L. Shao, "Hsva: Hierarchical semantic-visual adaptation for zero-shot learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 622–16 634, 2021.

[13] X. Kong, Z. Gao, X. Li, M. Hong, J. Liu, C. Wang, Y. Xie, and Y. Qu, "En-compactness: Self-distillation embedding & contrastive generation for generalized zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 9306–9315.

[14] Z. Hong, S. Chen, G. Xie, W. Yang, J. Zhao, Y. Shao, Q. Peng, and X. You, "Semantic compression embedding for generative zero-shot learning," *IJCAI, Vienna, Austria*, vol. 7, pp. 956–963, 2022.

[15] Y. Liu, L. Zhou, X. Bai, Y. Huang, L. Gu, J. Zhou, and T. Harada, "Goal-oriented gaze estimation for zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 3794–3803.

[16] F. Alamri and A. Dutta, "Multi-head self-attention via vision transformer for zero-shot learning," *arXiv preprint arXiv:2108.00045*, 2021.

[17] C. Wang, S. Min, X. Chen, X. Sun, and H. Li, "Dual progressive prototype network for generalized zero-shot learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2936–2948, 2021.

[18] F. Alamri and A. Dutta, "Implicit and explicit attention for zero-shot learning," in *DAGM German Conference on Pattern Recognition*. Springer, 2021, pp. 467–483.

[19] S. Chen, Z. Hong, Y. Liu, G.-S. Xie, B. Sun, H. Li, Q. Peng, K. Lu, and X. You, "Transzero: Attribute-guided transformer for zero-shot learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 330–338.

[20] S. Chen, Z. Hong, G.-S. Xie, W. Yang, Q. Peng, K. Wang, J. Zhao, and X. You, "Msdn: Mutually semantic distillation network for zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 7612–7621.

[21] Y. Hu, L. Feng, H. Jiang, M. Liu, and B. Yin, "Domain-aware prototype network for generalized zero-shot learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[22] Y. Zhang and S. Feng, "Enhancing domain-invariant parts for generalized zero-shot learning," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 6283–6291.

[23] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[24] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2751–2758.

[25] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2251–2265, 2018.

[26] Z. Xie, W. Cao, and Z. Ming, "A further study on biologically inspired feature enhancement in zero-shot learning," *International Journal of Machine Learning and Cybernetics*, vol. 12, no. 1, pp. 257–269, 2021.