Proceedings of Machine Learning Research vol 278:1–10, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

HD-MF: Hierarchical Dynamic-aware Multimodal Fusion for Fine-Grained Bird Recognition

Junjing Li 2262405029@STU.SUDA.EDU.CN
School of Future Science and Engineering, Soochow University, Suzhou, 215222, China

Xing Liu 2262405016@STU.SUDA.EDU.CN
School of Future Science and Engineering, Soochow University, Suzhou, 215222, China

Jiu Luo* LUOJIU@SUDA.EDU.CN
School of Future Science and Engineering, Soochow University, Suzhou, 215222, China
**Corresponding author*

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

Fine-grained bird recognition plays a crucial role in biodiversity monitoring. Its primary challenge lies in identifying subtle inter-class visual differences and overcoming the inherent limitations of unimodal information. Audio provides crucial complementary cues, yet audiovisual fusion still faces challenges such as the semantic gap. To address these challenges, this paper proposed a hierarchical dynamic-aware multimodal fusion (HD-MF) architecture. This architecture captures locally aligned cross-modal features via its Cross-modal Spatial Interaction Module, extracts global high-order cross-modal correlations using the Factorized Bilinear Fusion Module, and dynamically integrates the outputs of these two fusion approaches through a Dynamically Adaptive Gated Fusion Unit. Evaluated on AViS, a paired audiovisual dataset constructed for this study, HD-MF achieved state-of-the-art performance. Experimental results demonstrated that HD-MF effectively integrates audiovisual complementary information, providing a novel and effective approach for enhancing fine-grained bird recognition performance.

Keywords: Fine-grained Bird Recognition, Multimodal Fusion, Audiovisual Fusion, Biodiversity Monitoring

1. Introduction

Bird species recognition plays a crucial role in biodiversity monitoring and ecological research. Unlike generic object recognition tasks, the primary challenge in fine-grained bird image classification lies in identifying subtle inter-class visual differences, such as variations in beak morphology and wing patterns (Zhang et al., 2014). In recent years, leveraging the rapid advancements in Convolutional Neural Networks (CNNs) and Transformer architectures, methods based on a single visual modality have achieved significant progress, reaching classification accuracies approaching 93% (Chou et al., 2022; Wang et al., 2024) on the widely-used public benchmark dataset CUB-200-2011 (Wah et al., 2011). However, further improvements in classification accuracy seem to be plateauing, and this progress often comes at the cost of a significant increase in the number of model parameters.

Researchers have gradually recognized the limitations of relying solely on unimodal visual data, which has spurred the development of fine-grained bird recognition methods based on multimodal data, such as multimodal strategies integrating text-image (Choudhury et al., 2024) or audio-image (Zhou et al., 2022) information.

Visual and auditory information from birds often exhibit complementary characteristics in the spatio-temporal dimensions. In complex natural monitoring environments, acquired bird images are susceptible to various factors such as occlusion, motion blur, poor illumination conditions, and background interference, leading to incomplete or noisy extracted visual representations. In contrast, birdsong frequently contains significant behavioral and environmental information; therefore, audio data can effectively supplement visual information by providing additional fine-grained discriminative features (Zhang et al., 2024). However, fusing these two heterogeneous modalities presents several challenges, primarily including the semantic gap and modality imbalance (Baltrušaitis et al., 2018).

To address the aforementioned challenges, this paper proposed a Hierarchical Dynamic-aware Multimodal Fusion (HD-MF) architecture. Its overall structure is illustrated in Figure 1. HD-MF employs a multi-level fusion strategy: image and audio data are first processed through their respective backbone networks to extract intermediate features. These features are subsequently fed into the cross-modal spatial interaction module (CSIM) and the factorized bilinear fusion module (FBFM) to extract distinct types of fused features. Specifically, CSIM utilizes a cross-modal attention mechanism to explicitly model local semantic correlations between image and audio features, thereby achieving spatial alignment and synergy to capture spatially aligned fine-grained cross-modal features. FBFM focuses on extracting high-order cross-modal interactions at the global level to capture implicit global semantic correlations between the modalities. Finally, these two types of fused features are fed into the dynamically adaptive gated fusion unit (DAGFU). This unit adaptively adjusts and integrates these two feature types via dynamically learned gating weights, optimizing the final classification decision.

In summary, the HD-MF architecture overcomes the limitations of unimodal data by effectively integrating image and audio information. By leveraging its multi-level fusion strategy, the architecture thoroughly exploits cross-modal complementary information at the spatial, global, and decision levels, thereby significantly enhancing fine-grained bird image classification performance.

2. Methodology

This section provides a detailed description of our proposed HD-MF architecture. The architecture primarily comprises the CSIM (Sec 2.1), the FBFM (Sec 2.2), and the DAGFU (Sec 2.3). Figure 1 presents an overall overview of the architecture.

2.1. Cross-modal Spatial Interaction Module

The CSIM aims to achieve alignment and interaction between cross-modal features in the spatial dimension. This module takes intermediate image features $F_{img} \in \mathbb{R}^{B\times C\times H\times W}$ and audio features $F_{\text{aud}} \in \mathbb{R}^{B\times C\times T}$ from the backbone networks as input. Since F_{aud} is sequential data, we first convert it into a pseudo-spatial representation F_{aud}^{space} to match the spatial structure of the image features F_{img} through the following operations:

$$\widetilde{F}_{aud} = Interpolate\left(F_{aud}, L = H \times W\right), \tag{1}$$

$$F_{aud}^{space} = Reshape\left(\widetilde{F}_{aud}, [B, C, H, W]\right), \tag{2}$$

where $Interpolate\left(\cdot\right)$ denotes temporal interpolation, aimed at preserving local correlations as much as possible while mapping temporal information to the pseudo-spatial domain; $Reshape\left(\cdot\right)$

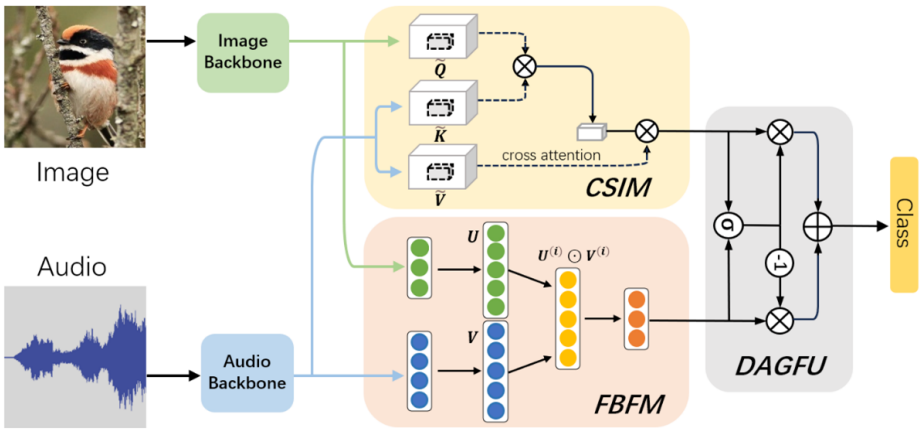

Figure 1: Overview of the proposed HD-MF architecture which consists of three parts. (1) Cross-modal Spatial Interaction Module (CSIM): Explicitly models local spatial correlations between image and audio features. (2) Factorized Bilinear Fusion Module (FBFM): Models high-order interactions between image and audio. (3) Dynamically Adaptive Gated Fusion Unit (DAGFU): Adaptively fuses the features extracted by CSIM and FBFM using dynamically learned weights.

is the tensor reshaping operation; H and W represent the spatial height and width of the image feature map, respectively.

Within the CSIM, we employ a cross-modal attention mechanism (Wei et al., 2020). Specifically, 1×1 convolutional layers $conv_q$, $conv_k$ and $conv_v$ are used to extract the Q from F_{img}, and the K, V from F_{aud}^{space}, respectively. Subsequently, Q, K and V are flattened along the spatial dimensions into sequence form, yielding $\widetilde{Q}$, $\widetilde{K}$ and $\widetilde{V}$. The spatially interactive cross-modal features F_{CSIM} are computed as follows:

$$\lambda = Softmax\left(\frac{\tilde{Q}\tilde{K}^T}{\sqrt{C'}}\right), \tag{3}$$

$$F_{CSIM} = F_{img} + \gamma \cdot \mathrm{Conv}(\lambda \tilde{V}), \tag{4}$$

where Equation (3) calculates the Scaled Dot-Product Attention weight matrix $\lambda \in \mathbb{R}^{B\times(HW)\times(HW)}$ between $\tilde{Q}$ and $\tilde{K}$, while Equation (4) utilizes this attention weight matrix λ to perform a weighted aggregation of $\tilde{V}$, which is then added to the original image features F_{img} via a residual connection to yield the spatially interactive cross-modal features $F_{CSIM} \in \mathbb{R}^{B\times C\times H\times W}$. γ is a learnable scaling parameter.

To address the common issue of spatio-temporal misalignment in cross-modal fine-grained features, CSIM leverages its attention mechanism, treating audio information as dynamic contextual cues. It adaptively performs cross-modal feature alignment and fusion within the spatial dimensions of the image feature map, thereby guiding the model to focus on audio features that are semantically highly relevant to local image regions. This helps enhance the spatial dependencies between modalities and improves the accuracy and robustness of the fused feature representation.

2.2. Factorized Bilinear Fusion Module

Inspired by related work (Yu et al., 2017), the Factorized Bilinear Fusion Module is designed to capture high-order global interactions between image and audio features. This module first applies Global Average Pooling (GAP) to the input intermediate features F_{img} and the pseudo-spatial audio features F_{aud}^{space}, aggregating them from their respective spatial (or pseudo-spatial) dimensions into global feature vectors to obtain more compact global semantic representations:

$$f_{img} = GAP\left(F_{img}\right) \in \mathbb{R}^{B\times C}, \tag{5}$$

$$f_{aud} = GAP\left(F_{aud}^{space}\right) \in \mathbb{R}^{B\times C}. \tag{6}$$

Subsequently, these global feature vectors undergo linear projection using learnable weight matrices W_u and W_v:

$$U = W_u f_{img}, \tag{7}$$

$$V = W_v f_{aud}, \tag{8}$$

where $U, V \in \mathbb{R}^{B\times(C\cdot k)}$ are the projected features, and k is a predefined decomposition factor. Finally, the global bilinear fusion features F_{FBFM} are computed by performing an element-wise product on the projected features U and V along their factor dimension, followed by summation:

$$F_{FBFM} = \sum_{i=1}^{k} U^{(i)} \odot V^{(i)}, \tag{9}$$

where $F_{FBFM} \in \mathbb{R}^{B\times C}$, and $U^{(i)}, V^{(i)} \in \mathbb{R}^{B\times C}$ respectively denote the$i-th$ sub-vectors obtained by splitting the tensors U and V along the factor dimension. $\odot$ represents the Hadamard product. In summary, the Factorized Bilinear Fusion Module, by explicitly modeling the second-order interactions between image and audio features, effectively mines the global complementary information across modalities, thereby helping to capture deeper-level global semantic correlations between them.

2.3. Dynamically Adaptive Gated Fusion Unit

The DAGFU is designed to adaptively fuse feature information from the CSIM and FBFM modules in a data-driven manner. This unit receives two inputs: the spatially interactive features after Global Average Pooling (GAP), denoted as $f_{spatial} = GAP\left(F_{CSIM}\right) \in \mathbb{R}^{B\times C}$, and the global bilinear fusion features $f_{global} = F_{FBFM} \in \mathbb{R}^{B\times C}$. Based on these two input features, the DAGFU first calculates a gating vector g:

$$g = \sigma\left(W_g\left[f_{spatial} \oplus f_{global}\right]\right), \tag{10}$$

where $\oplus$ denotes the feature concatenation operation, W_g is a learnable weight matrix, and $\sigma\left(\cdot\right)$ represents the Sigmoid activation function, ensuring that each element of the output g lies within the range $[0, 1]$. This gating vector g is then used to perform a weighted fusion of the two input features, yielding the final fused feature f:

$$f = g \odot f_{spatial} + (1 - g) \odot f_{global} \tag{11}$$

Given that the spatial interaction features $f_{spatial}$ primarily capture local spatial details, while the global bilinear fusion features f_{global} focus more on global semantic correlations, the relative importance of these two feature types may vary across different input samples. Therefore, the DAGFU employs this dynamic weight adjustment mechanism, enabling the model to dynamically modulate the fusion ratio of these two feature types based on the specific characteristics of the current input data. This adaptive strategy overcomes the limitations of fixed weighted fusion methods (such as simple averaging or summation), allowing for better preservation and utilization of the complementary advantages offered by both spatial details and global semantic information. Furthermore, through this real-time feature importance assessment and weighting, the mechanism is expected to enhance the model's robustness and generalization capability, ultimately improving multimodal classification performance.

3. Experiments

3.1. Dataset Construction

To effectively evaluate the performance of the proposed HD-MF architecture, we constructed a novel medium-scale multimodal dataset named AViS, which contains paired bird images and audio recordings. This dataset covers 61 distinct bird species. For each species, it includes approximately 200 high-quality images and around 50 audio clips in WAV format on average. Each audio clip has a duration of 5 seconds. Although the number of image and audio samples varies somewhat across species, the AViS dataset generally exhibits good inter-class balance.

3.2. Implementation Details

Image Backbone: We selected a ResNet-50 (He et al., 2016) model, a widely-used deep convolutional neural network known for its residual connections that facilitate the training of deeper architectures, pre-trained on the ImageNet (Deng et al., 2009) dataset as the image feature extractor and removed its original final fully connected layer. Input images were uniformly resized to 224x224 pixels before being fed into the network. During training, random rotation and random horizontal flipping were applied as data augmentation strategies.

Audio Backbone: We employed a wav2vec2 (Baevski et al., 2020) model, a prominent framework for self-supervised learning of speech representations directly from raw audio data, pre-trained on the *WSJ* and *Librispeech* datasets as the audio feature extractor. All audio samples were processed into WAV file format, with the sampling rate uniformly set to 16kHz, duration truncated or padded to 5 seconds, and converted to mono.

We randomly partitioned the AViS dataset into training, validation, and test sets using a 70%:15%:15 ratio. During the model training phase, we adopted a differential learning rates strategy: specifically, a lower learning rate was set for the pre-trained backbone network layers to better preserve their learned general-purpose feature representations, while a relatively higher learning rate was assigned to the HD-MF feature fusion network components to accelerate model convergence.

In the experiments, all models were trained for 50 epochs. We employed a learning rate scheduler: if the accuracy on the validation set did not improve for 5 consecutive epochs, the current learning rate was decayed by multiplying it by a factor of 0.1.

3.3. Comparative Experiments

To evaluate the performance of the proposed HD-MF architecture, we compared it against a series of baseline models and multimodal fusion methods on the AViS test set. We first assessed the performance of baseline models using only a single modality (either image or audio) to establish the performance benchmark for unimodal approaches on this fine-grained bird recognition task. Subsequently, we compared the performance against several representative multimodal feature fusion methods. These include MFB (Yu et al., 2017), known for its factorized bilinear pooling approach to capture inter-modal interactions; MulT (Tsai et al., 2019), which employs cross-modal Transformers for fusing unaligned multimodal sequences; AudioCLIP (Guzhov et al., 2022), an extension of contrastive learning paradigms like CLIP to the audio domain; and EchoTrack (Lin et al., 2024), a recent method for audio-visual fusion in dynamic scenarios, leveraging cross-modal attention. We compared these methods alongside our proposed HD-MF model. To ensure a fair comparison, all multimodal methods utilized the same pre-trained backbone networks for feature extraction as used in HD-MF. The comparative results are presented in Table 1.

Table 1: Comparison of model performance on the AViS test set. '✓' indicates usage of the corresponding modality. The best result is shown in bold.

Method	Image	Audio	Accuracy (%)
ResNet-50	✓		84.2
wav2vec2		✓	87.3
MFB	✓	✓	91.8
MulT	✓	✓	92.5
AudioCLIP	✓	✓	93.1
EchoTrack	✓	✓	93.9
HD-MF	✓	✓	**94.4**

The experimental results showed that, compared to the ResNet-50 baseline using only the image modality, our proposed HD-MF method achieved a significant performance improvement of 10.2%. Compared to the wav2vec2 baseline using only the audio modality, the performance improved by 7.1%. This strongly demonstrated the significant complementarity between bird image and audio information, and that effectively fusing information from these two modalities can significantly improve the performance of fine-grained bird classification. Furthermore, compared to other multimodal feature fusion methods, HD-MF also achieved the highest performance, which further validated the superiority of the HD-MF architecture in effectively mining and utilizing cross-modal complementary information.

3.4. Ablation Studies

To validate the effectiveness of the key components within our proposed HD-MF model and their respective contributions to the final performance, we conducted a series of ablation studies. These studies aimed to evaluate the performance of different model variants, with the results summarized in Table 2.

Table 2: Ablation study results of the HD-MF model and its variants on the AViS dataset. Feature Concatenation: Baseline, direct feature concatenation. CSIM: Model with only the CSIM module. FBFM: Model with only the FBFM module. -DAGFU: Model with CSIM + FBFM using static fusion, without the dynamic gate. HD-MF: The complete model.

Method	Accuracy (%)
Feature Concatenation	89.5
CSIM	92.3
FBFM	91.7
-DAGFU	93.8
HD-MF	**94.4**

The results demonstrated that compared to the baseline method of directly concatenating image and audio features (Feature Concatenation), individually introducing the CSIM yielded a performance improvement of 2.8%. This indicated the effectiveness of explicitly modeling the fine-grained correspondence between cross-modal data in the spatial dimension. Similarly, incorporating the FBFM alone resulted in a 2.2% performance increase over the baseline, validating the efficacy of mining high-order interactions between image and audio features at a global scale. Notably, when CSIM and FBFM were used in combination (corresponding to the "-DAGFU" model variant in Table 2, implying static fusion), the performance further improved to 93.8%. This suggested that the local spatial interaction features captured by CSIM and the global fusion features extracted by FBFM provided complementary information from different perspectives. Finally, the complete HD-MF model, constructed by introducing the DAGFU, achieved a synergistic effect through adaptive weighted fusion of these two feature types, reaching the optimal performance of 94.4% and significantly boosting the overall accuracy of bird classification.

3.5. Effectiveness with Different Backbones

To further evaluate the generalizability and robustness of the proposed HD-MF fusion framework, we designed and conducted a supplementary experiment. This experiment aimed to assess the effectiveness of the HD-MF framework in fusing features extracted by different backbones, thereby validating that its applicability is not limited to specific backbone combinations. In this experiment, the structure and parameter settings of the HD-MF fusion module were kept fixed. We selected representative and state-of-the-art image and audio backbones to replace the original ResNet-50 and wav2vec2. Specifically, the image backbones chosen were EfficientNet-B4 (Tan and Le, 2019) and Swin Transformer-T (Liu et al., 2021), while the audio backbone selected was AST (Gong et al., 2021). For all chosen backbones, we loaded their weights pre-trained on large-scale datasets and subsequently fine-tuned them on the AViS dataset. We also recorded the number of parameters for each backbone, aiming to analyze the relationship between model performance and complexity. The corresponding experimental results are summarized in Table 3.

The experimental results demonstrated that the proposed HD-MF fusion framework exhibited high effectiveness and robustness when integrating features extracted by different image and audio backbones. Even when the feature extraction backbones were replaced, the performance of the HD-MF framework remained stable within the range of 94.0% to 95.0%, showing minor and expected

Table 3: Performance comparison of HD-MF with different backbone combinations on the AViS test set. Parameter counts refer to the respective backbone models. The base versions of wav2vec and AST were used. The best result is shown in bold.

Backbone		Params (M)		Accuracy (%)
Image	Audio	Image	Audio	
ResNet-50	wav2vec2	∼25.6	∼95	94.4
EfficientNet-B4	wav2vec2	∼19.3	∼95	94.0
Swin Transformer-T	wav2vec2	∼28.3	∼95	94.8
ResNet-50	AST	∼25.6	∼87	94.2
EfficientNet-B4	AST	∼19.3	∼87	94.1
Swin Transformer-T	AST	∼28.3	∼87	**95.0**

fluctuations. Specifically, employing backbones with superior performance or a larger number of parameters generally corresponded to commensurate improvements in performance. These findings strongly confirmed that the hierarchical dynamic fusion mechanism utilized by HD-MF is capable of adapting to feature inputs from diverse sources and representations. This indicated its efficacy as a generalizable multimodal information integration strategy capable of consistently enhancing performance on the fine-grained bird recognition task.

4. Conclusion

Addressing the challenges of unimodal limitations and audiovisual fusion in fine-grained bird recognition, this paper proposed the HD-MF architecture. This architecture combines the local spatial interaction capability of CSIM, the global high-order fusion capability of FBFM, and the dynamic gating mechanism of DAGFU to effectively integrate complementary audiovisual information. Experiments demonstrated that HD-MF achieved state-of-the-art performance on the AViS dataset constructed for this study. Future work could focus on extending HD-MF to more complex application scenarios and exploring possibilities for model lightweighting.

Acknowledgments

The corresponding author Jiu Luo thanks support by the Opening Project of Guangdong Province Key Laboratory of Computational Science at the Sun Yat-sen University (No. 2024014).

References

A. Baevski, H. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33: 12449–12460, 2020. doi: 10.48550/arXiv.2006.11477.

T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018. doi: 10.1109/TPAMI.2018.2798607.

P.-Y. Chou, C.-H. Lin, and W.-C. Kao. A novel plug-in module for fine-grained visual classification. *arXiv preprint arXiv:2202.03822*, 2022. doi: 10.48550/arXiv.2202.03822.

S. Choudhury, I. Laina, C. Rupprecht, and A. Vedaldi. The curious layperson: Fine-grained image recognition without expert labels. *International Journal of Computer Vision*, 132(2):537–554, 2024. doi: 10.1007/s11263-023-01885-9.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Y. Gong, Y.-A. Chung, and J. Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. doi: 10.21437/Interspeech.2021-698.

A. Guzhov, F. Raue, J. Hees, and A. Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980, 2022. doi: 10.48550/arXiv.2106.13043.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

J. Lin, J. Chen, K. Peng, X. He, Z. Li, R. Stiefelhagen, and K. Yang. Echotrack: Auditory referring multi-object tracking for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):18964–18977, 2024. doi: 10.1109/TITS.2024.3437645.

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. doi: 10.48550/arXiv.2103.14030.

M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114, 2019. doi: 10.48550/arXiv.1905.11946.

Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the Association for Computational Linguistics Meeting*, page 6558, 2019. doi: 10.18653/v1/P19-1656.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset, 2011. Accessed 2025-05-27.

J. Wang, Q. Xu, B. Jiang, B. Luo, and J. Tang. Multi-granularity part sampling attention for fine-grained visual classification. *IEEE Transactions on Image Processing*, 33:4529–4542, 2024. doi: 10.1109/TIP.2024.3441813.

X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10941–10950, 2020. doi: 10.1109/CVPR42600.2020.01095.

Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1821–1830, 2017. doi: 10.48550/arXiv.1708.01471.

F. Zhang, C. Zhao, and B. Geng. Deep multimodal data fusion. *ACM Computing Surveys*, 56(9): 1–36, 2024. doi: 10.1145/3649447.

N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based r-cnns for fine-grained category detection. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, volume 13, pages 834–849, 2014. doi: 10.1007/978-3-319-10590-1_54.

X. Zhou, X. Song, H. Wu, J. Zhang, and X. Xu. Mavt-fg: Multimodal audio-visual transformer for weakly-supervised fine-grained recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3811–3819, 2022. doi: 10.1145/3503161.3548383.