

Mitigating Textual Noise in Multimodal FGVC via Hierarchical Semantic Purification and Multi-Stage Alignment

Menghuan Zhang, Qing Cai^{ID}, *Member, IEEE*, Fan Zhang^{ID}, Guanbin Li^{ID}, *Member, IEEE*, Junyu Dong, *Member, IEEE*, David Zhang^{ID}, *Life Fellow, IEEE*, and Sheng Chen^{ID}, *Life Fellow, IEEE*

Abstract—Fine-grained visual classification (FGVC) plays a crucial role in the realm of computer vision. Recently, multimodal FGVC methods, leveraging textual descriptions as semantic guidance, have gained considerable attention. However, current approaches often encounter two primary limitations: 1) Redundant or ambiguous textual descriptions: existing methods rely on raw or generated descriptions without filtering, introducing redundant and ambiguous semantic noise; and 2) Underutilization of hierarchical visual features: most approaches align single-layer visual features with auxiliary semantic embeddings, underutilizing hierarchical information. To address these challenges, we propose a task-oriented multimodal FGVC framework that eliminates textual redundancy while enhancing multi-layer alignment between cross-modalities. Specifically, our method comprises two key components: Hierarchical Semantic Purification (HSP) and Multi-layer Cross-Modal Alignment (MCA). The former employs a semantic distillation dictionary to eliminate redundant elements and uses a self-attention mechanism for ranking and semantic refinement. The latter establishes effective cross-modal fusion by integrating multi-layer features with purified text features, effectively combining multi-scale visual representations. Experimental results on 7 public datasets demonstrate that our proposed method outperforms existing

counterparts, contributing to advancements in fine-grained visual classification.

Index Terms—Fine-grained visual classification, multimodal learning, alignment and fusion.

I. INTRODUCTION

FINE-GRAINED visual classification (FGVC) aims to distinguish subordinate categories within the same class (e.g., bird species [1], [2], dog breeds [3], aircraft [4]), and has demonstrated significant potential in domains such as biodiversity monitoring [5], diseases subtype identification [6], [7], [8], [9], ophthalmic imaging [10], [11], security [12], and industrial defect inspection [13], [14]. The core challenge of FGVC lies in modeling subtle inter-class differences while remaining robust to large intra-class variations [15]. Traditional FGVC approaches can be divided into two paradigms: unimodal (visual-only) [16], [17], [18] and multimodal (vision-language fusion) approaches [19].

Unimodal approaches focus on extracting salient visual features and have demonstrated competitive performance by effectively capturing fine-grained information from high-resolution images [20], [21], [22], [23], [24]. However, the information provided by single-modal data is limited due to its inability to leverage complementary cross-modal information, which is crucial for robust feature representation in complex scenarios. To address these limitations, recent works have turned to multimodal learning, integrating vision and language modalities [19], [25], [26], [27], [28]. By leveraging textual descriptions to guide visual attention, these methods have demonstrated impressive performance in FGVC. As illustrated in the Fig. 1(a), existing multimodal FGVC methods still suffer from two critical challenges despite their promise.

1) **Redundant or ambiguous textual descriptions:** Many current methods utilize descriptions generated by vision-to-language generative models [29], [30], [31], [32], [33] or collected from public repositories [34]. These textual descriptions often contain visually irrelevant elements (e.g., background) or ambiguous descriptions (e.g., size, pose), which can lead to attention diffusion and degraded classification performance [19], [35]. While recent advances such as prompt tuning [36], [37], [38], [39], [40] and prompt-guided multimodal generation frameworks [41] improve input-level adaptation, they fall short in explicitly filtering noise or modeling token-level semantic relevance. 2) **Underutilization of**

Received 9 January 2026; revised 15 July 2026; accepted 22 August 2026. Date of publication 7 September 2026; date of current version 23 September 2026. This work was supported in part by the National Science Foundation of China under Grant 62622604, Grant 62601686, and Grant 62471448; in part by Shandong Provincial Natural Science Foundation under Grant ZR2024YQ004; and in part by the Taishan Scholars Youth Expert Program of Shandong Province under Grant tsqn202312109. The associate editor coordinating the review of this article and approving it for publication was Dr. Changqing Zhang. (Menghuan Zhang and Qing Cai contributed equally to this work.) (Corresponding authors: Qing Cai; Fan Zhang.)

Menghuan Zhang and Junyu Dong are with the Department of Information Science and Engineering, Ocean University of China, Qingdao 266100, China (e-mail: zhangmenghuan@stu.ouc.edu.cn; dong-junyu@ouc.edu.cn).

Qing Cai is with the Innovation School of Artificial Intelligence, Hefei University of Technology, Hefei 230009, China (e-mail: caiqing1617@gmail.com).

Fan Zhang is with the Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: fan-hti.zhang@polyu.edu.hk).

Guanbin Li is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China (e-mail: liguanbin@mail.sysu.edu.cn).

David Zhang is with the School of Artificial Intelligence and the School of Data Science, The Chinese University of Hong Kong, Shenzhen, Guangdong 518172, China (e-mail: davidzhang@cuhk.edu.cn).

Sheng Chen is with the School of Electronics and Computer Science, University of Southampton, SO17 1BJ Southampton, U.K., and also with the Department of Information Science and Technology, Ocean University of China, Qingdao 266100, China (e-mail: sqc@ecs.soton.ac.uk).

This article has supplementary downloadable material available at <https://doi.org/10.1109/TIP.2026.3729490>, provided by the authors.

Digital Object Identifier 10.1109/TIP.2026.3729490

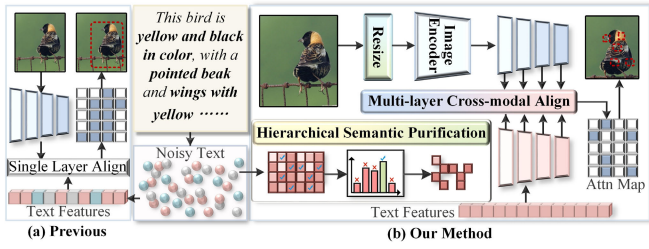


Fig. 1. Motivation: The integration of redundant textual features and single-layer visual features. The previous (Left) utilizes redundant textual inputs with the last-layer visual features, yielding diffuse attention activation with background interferences. In contrast, our method (Right) integrates the HSP and MCA modules to focus on discriminative regions (e.g., avian beaks, caudal, and pedal extremities) while suppressing environmental noise.

hierarchical visual features: Existing multimodal approaches primarily align final-layer visual features with single-layer text embeddings, neglecting rich hierarchical visual representations in deep visual backbones. This simplification hinders the model’s ability to capture cross-scale semantic correspondences, thereby limiting the effectiveness of cross-modal integration [42].

Motivated by the above discussions, we propose a task-oriented multimodal fine-grained visual classification framework as distinctively shown in Fig. 1(b), which consists of two components. 1) **Hierarchical Semantic Purification (HSP):** A text refinement module that removes redundant or ambiguous tokens based on a domain-specific dictionary and a token contribution ranking mechanism. This purification ensures that only the most semantically discriminative tokens are retained for vision-language alignment. 2) **Multi-layer Cross-modal Alignment (MCA):** A cross-modal fusion module that hierarchically integrates multi-layer visual features with dynamically reshaped multi-layer text embeddings refined by the HSP module. The text embeddings are reshaped to match multi-level visual features and fused via adaptive cross-attention, followed by a contrastive loss to regularize cross-modal feature consistency under the supervision of classification objectives.

The illustration of Fig. 1 contrasts our method with the previous one, emphasizing that our proposed method effectively suppresses irrelevant background activations and focuses attention on semantically discriminative object regions (e.g., the regions of avian beak, caudal, and pedal). Extensive experiments on 7 public FGVC datasets demonstrate that our approach consistently outperforms previous excellent unimodal and multimodal baselines, validating the effectiveness of both HSP and MCA components. To summarize, our main contributions are as follows.

- We propose a task-oriented hierarchical semantic purification strategy for multimodal fine-grained visual classification, which filters redundant and ambiguous textual tokens to retain visually grounded and discriminative descriptive cues.
- We design an effective multi-layer cross-modal alignment scheme that aligns hierarchical textual features with visual representations, enabling structurally consistent

and semantically stable multimodal fusion tailored to fine-grained discrimination.

- We conduct comprehensive experiments on 7 widely used fine-grained visual classification datasets, demonstrating the robustness and effectiveness of our method against recent state-of-the-art methods.

II. RELATED WORK

A. Unimodal Fine-Grained Visual Classification

Unimodal FGVC methods rely solely on the visual modality to distinguish fine-grained categories, which can be broadly categorized into two main categories [15]: convolutional neural network (CNN)-based approaches [43], [44], [45], [46], [47] and Transformer-based approaches [48], [49], [50], [51], [52]. CNN-based methods typically localize the most discriminative regions and extract part-level visual features to improve classification. Representative works, such as CrossX [53], DCL [54], P2P-Net [20], FDL [55], SFFF [56], AMRFF [57], and PMG-V2 [58], adopt part alignment, keypoint supervision or region feature learning to reduce intra-class variation. Transformer-based models have shown notable performance due to their global receptive fields and self-attention mechanisms. Pioneering works, such as TransFG [59], IELT [23], DoD [60], TransIFC [61], MPSA [62], FAL-ViT [63], CAMF [64], and ViT-Net [48], leverage patch-wise attention to extract rich semantics and enhance global-local feature interaction for identifying the object. Further improvements, such as the Swin Transformer [49], further improve hierarchical representation by introducing window-based attention mechanisms. More recently, semantic-aware self-supervised learning has been introduced to enhance discriminative representation learning for fine-grained visual recognition without relying on additional annotations [65], demonstrating the effectiveness of semantic supervision for feature learning.

Although unimodal methods have shown strong performance, they are still constrained by the lack of semantic guidance from other modalities, which is particularly critical in visually ambiguous or low-resolution settings.

B. Multimodal Fine-Grained Visual Classification

Multimodal learning has attracted substantial attention in FGVC due to its ability to leverage complementary information across modalities [15]. Compared with unimodal approaches, integrating textual cues with visual features strengthens semantic grounding and discriminative capability. Pioneering studies established joint embedding spaces [28] and introduced pooled textual features as auxiliary supervision [25]. Subsequent methods enhanced cross-modal modeling using Fisher vectors [26], [27], self-attention [42], and dynamic MLPs [66]. More recently, prompt tuning has become a dominant paradigm for adapting frozen vision-language models such as CLIP [67]. Representative methods include CoOp [37], CoCoOp [36], MaPLe [38], and multi-granularity prompt learning [68]. Large-scale vision-language models, including BLIP-2 [30] and LLaVA [69], improve multimodal understanding but are not designed for fine-grained recognition. Recent studies have also explored hierarchical

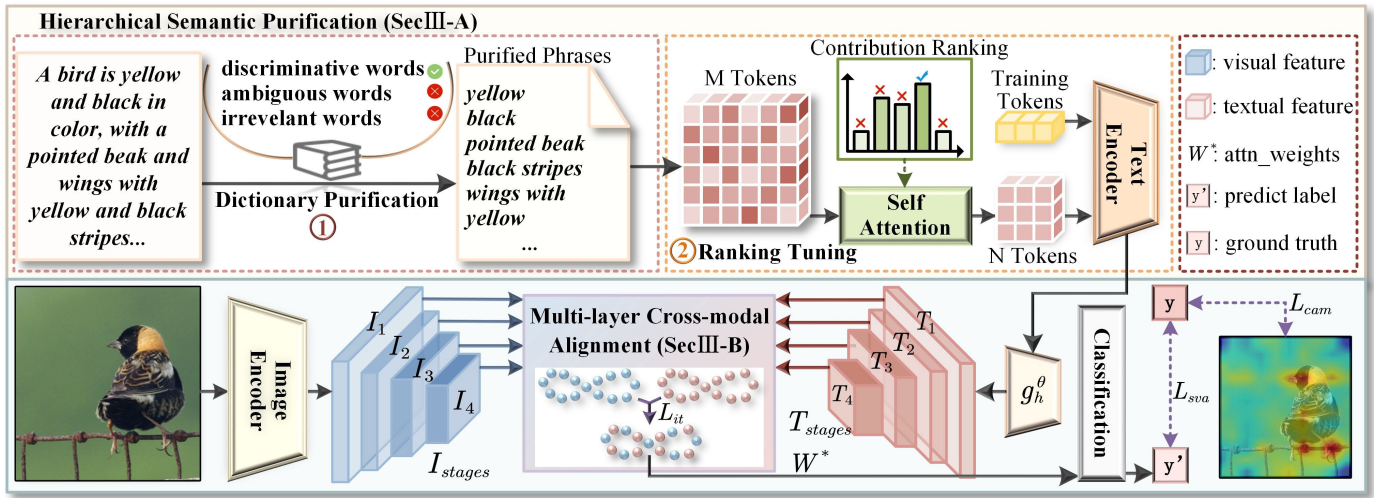


Fig. 2. Overview of the proposed multimodal FGVC framework. The upper section illustrates the HSP module, which refines and purifies the most relevant textual features through a 2-stages purification process. The lower section denotes the MCA module, which synergistically enhances multi-layer cross-modal feature interaction. The g_h^θ comprises a linear layer and four convolutional layers, transferring the single-level textual feature into a multi-layer textual representation.

architectures [70], domain adaptation [71], [72], and text-guided feature refinement [73] to enhance robustness and cross-modal interaction. Inspired by these advances, our work focuses on purifying textual descriptions before hierarchical cross-modal alignment, reducing the adverse effects of linguistic noise.

Consequently, despite the remarkable progress of multimodal FGVC methods, two key limitations remain: 1) **Redundant or ambiguous textual descriptions**: existing methods rely on unfiltered descriptions, introducing semantic noise; and 2) **Underutilization of hierarchical visual features**: most methods align single-layer visual features with semantic embeddings, underutilizing deeper representations. Our method instead focuses on semantic distillation by refining salient semantics while retaining informative signals. To address these challenges, we propose a unified framework comprising a Hierarchical Semantic Purification (HSP) module and a Multi-layer Cross-modal Alignment (MCA) module. HSP refines textual input through a multi-stage process, preserving the most discriminative tokens while filtering out noise. MCA performs semantic-level cross-attention between multi-layer textual and visual representations, enabling more precise multimodal fusion.

III. METHODOLOGY

As illustrated in Fig. 2, our proposed framework consists of two pivotal components: the **Hierarchical Semantic Purification (HSP)** module and the **Multi-layer Cross-modal Alignment (MCA)** module. HSP refines the initial textual descriptions by filtering out redundant or irrelevant information and selecting the most semantically discriminative words. MCA facilitates effective integration and alignment between hierarchical visual and textual features.

A. Hierarchical Semantic Purification

HSP adopts a hierarchical text purification framework to refine textual descriptions. Specifically, it consists of two

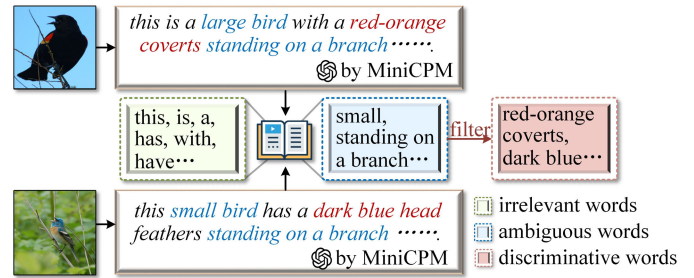


Fig. 3. Illustration of the Dictionary Purification.

stages: (1) dictionary-guided purification, which removes irrelevant and ambiguous words based on predefined linguistic priors; and (2) contribution-aware ranking, where tokens are scored and selected according to their contextual relevance. To further improve generalization, the purified tokens are jointly optimized with randomly initialized learnable tokens, acting as adaptive semantic complements. Through this multi-stage refinement process, HSP preserves discriminative textual cues while suppressing noisy information, facilitating fine-grained classification.

As depicted in Fig. 2, rule-based text filtering, such as stop-word and domain-specific removal, has been widely adopted in NLP to reduce semantic redundancy and improve downstream representations [74], [75]. Inspired by these practices, we construct a semantic dictionary $\text{Dict}(\cdot)$ that categorizes words into irrelevant words (e.g., “the”, “a”, “and”), semantically ambiguous words (e.g., “big”, “small”), and discriminative words. To ensure scalability and eliminate intensive manual engineering, $\text{Dict}(\cdot)$ is constructed via a hybrid approach combining data statistics and linguistic rules. Irrelevant and ambiguous words often provide limited discriminative information and may introduce noise during multimodal alignment. For example, the phrase “standing on a branch” frequently appears across descriptions of many bird species (Fig. 3), potentially misleading cross-modal correspondence. By filtering descriptions with a domain-tailored dictionary, HSP preserves visually

correlated attributes while suppressing redundant expressions, enhancing the robustness of subsequent cross-modal fusion. Notably, the textual descriptions generated by MiniCPM are objective visual observations and do not encode class labels. Consequently, the purified textual features serve as auxiliary semantic guidance for visual representation learning without leaking ground-truth category information.

During the Contribution Ranking stage, the purified description is tokenized into M discrete tokens via a tokenizer $Tk(\cdot)$. To identify the most discriminative semantic elements, we introduce a self-attention-based **Contribution Ranking** mechanism that captures contextual dependencies among tokens. Specifically, the token sequence is processed by a self-attention module $\text{Self}(\cdot)$ to compute pairwise token affinities, based on which a *Contribution Ranking Score* is derived for each token by aggregating its contextual connectivity (Fig. 2, top). Tokens are then ranked according to their contribution scores, and the top N tokens are selected to preserve the most informative semantic cues. To enhance adaptability while maintaining stable cross-modal alignment, these high-contribution tokens are concatenated with a set of randomly initialized learnable tokens t_k and jointly encoded by a frozen CLIP text encoder $E_t(\cdot)$ [76]. The resulting refined textual embeddings are subsequently employed for alignment.

HSP enhances semantic precision by systematically removing redundant or noisy elements and retaining only the most discriminative tokens. Through contribution-based ranking, the model emphasizes textual cues closely aligned with fine-grained visual semantics, thereby improving both textual representation quality and multimodal alignment. The overall operation can be formally expressed as:

$$F_T = E_t(\text{Self}(Tk(\text{Dict}(t_i))) + t_k) \in \mathbb{R}^{B \times L_s \times C_s}, \quad (1)$$

where B denotes the batch size, L_s corresponds to the spatial sequence length and C_s represents the feature channel dimension. The refined text feature F_T obtained through the HSP serves as the input to the MCA module. It is worth noting that while F_T is subsequently mapped to different dimensions to match multi-level visual features, these variants are scale-adapted and re-parameterized forms of the same unified, purified semantic representation, rather than implies that the text itself contains an intrinsic hierarchical structure.

Although the HSP module operates independently within each mini-batch, the use of a shared CLIP text encoder across all samples enables token representations to accumulate global semantic context throughout training. Consequently, while the Contribution Ranking mechanism selects top tokens locally within batches, the shared embedding space ensures that semantically salient tokens are consistently emphasized across batches. This design enables the model to learn global dependencies by aggregating semantic information, allowing it to establish relationships between discriminative tokens even if they never appear together within a single mini-batch.

B. Multi-Layer Cross-Modal Alignment

To strengthen the alignment between textual and visual modalities in multi-layer feature spaces, we propose **Multi-layer Cross-modal Alignment (MCA)** module. As illustrated

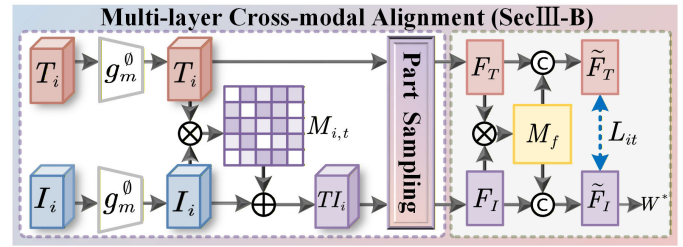


Fig. 4. Detailed process of the MCA module.

in Fig. 4, MCA establishes fine-grained associations between linguistic cues and visual features by integrating multi-layer representations from both modalities. This integration effectively bridges the semantic gap in multimodal learning by reinforcing cross-modal consistency and capturing complementary information across hierarchical levels.

Most existing approaches align a single-layer textual representation with visual features, which limits their ability to exploit the hierarchical and multi-scale structure of modern visual backbones. In addition, a global text embedding often lacks sufficient structural correspondence for effective multi-level fusion. To overcome these limitations, we project a shared purified text embedding into multiple spatial resolutions using lightweight linear projections followed by depth-aware convolutions. Notably, these multi-level textual features do not introduce new semantic content; rather, they are deterministic re-parameterizations of the same underlying semantics, adapted to match the spatial scales of corresponding visual layers. Since all textual representations are derived from a common embedding and optimized under unified supervision, semantic consistency across scales is naturally preserved, enabling structurally coherent and effective cross-modal fusion for fine-grained classification.

Formally, let $F_T \in \mathbb{R}^{B \times L_s \times C_s}$ denote the purified textual representation produced by HSP (see Eq. 1). Rather than assuming any inherent linguistic hierarchy, we introduce a semantic projection operator $g_h^0(\cdot)$ to perform task-oriented representational re-parameterization, adapting a single global textual embedding to the multi-stage structure of the visual encoder. Specifically, $g_h^0(\cdot)$ is implemented using lightweight linear projections followed by depth-aware convolutions, yielding a set of scale-aligned textual features:

$$T_{\text{stages}} = g_h^0(F_T) = \{T_1, T_2, T_3, T_4\}. \quad (2)$$

Here, each $T_i \in \mathbb{R}^{B \times H_i \times W_i \times C_i}$ is constructed to match the spatial resolution and channel dimension of the corresponding visual feature map I_i . The number of projection branches is determined by the depth of the visual backbone, rather than by any linguistic prior. This scale-consistent design facilitates structured cross-modal interaction across visual stages while introducing minimal inductive bias. Despite its simplicity, the proposed linear projection-based formulation proves effective in practice, as validated by the ablation studies in Section IV-D.4, which demonstrate consistent improvements in alignment quality and fine-grained classification accuracy.

Specifically, as illustrated in Fig. 2, MCA receives a set of hierarchical features from visual and textual modalities

as input. It is important to clarify that the MCA module is not intended to perform complex reasoning-oriented or generative multimodal fusion. Instead, its design goal is to enable alignment-driven semantic interaction between structurally corresponding visual and textual features. By operating under scale-consistent representations, MCA focuses on establishing reliable semantic correspondence with minimal architectural overhead, which is particularly suitable for fine-grained visual classification. The four-stage visual features $I_{\text{stages}} = \{I_1, I_2, I_3, I_4\}$ are extracted from the image encoder $E_i(\cdot)$, $1 \leq i \leq 4$, while the corresponding four-stage textual features $T_{\text{stages}} = \{T_1, T_2, T_3, T_4\}$ are obtained via semantic projection of the purified text embeddings, as introduced in the previous section.

To enable cross-modal interaction, visual and textual features at semantic level are projected into a shared latent space using a modality alignment operator $g_{\theta_m}(\cdot)$, formulated as:

$$\begin{aligned} \hat{I}_i &= g_{\theta_m}(I_i), \\ \hat{T}_i &= g_{\theta_m}(T_i), \quad i = 1, 2, 3, 4, \end{aligned} \quad (3)$$

where $\hat{I}_i$ and $\hat{T}_i$ denote the aligned visual and textual feature at the i -th level. This projection into a unified embedding space facilitates the learning of cross-modal representations, forming the foundation for subsequent cross-attention-based fusion.

Then, the cross-modal interaction is modeled using attention-based correlation matrices. We compute an attention matrix $M_{i,t} \in \mathbb{R}^{H_i \times W_i}$ at each semantic level i to measure the spatial relevance between $\hat{I}_i$ and the corresponding $\hat{T}_i$:

$$M_{i,t} = \hat{I}_i \otimes \hat{T}_i, \quad (4)$$

where $\otimes$ denotes matrix multiplication along the feature dimension. Such similarity-based alignment has been shown to be effective and interpretable for establishing cross-modal relationships without introducing excessive inductive bias.

Based on the computed attention matrix $M_{i,t}$, the visual features are semantically enhanced via a residual fusion mechanism, which can be formulated as:

$$TI_i = \hat{I}_i + M_{i,t}, \quad (5)$$

where the residual connection facilitates gradient flow and integrates semantic priors from the textual modality directly into the visual stream. This integration enriches the visual representation with language-grounded context, improving the model's capacity to distinguish fine-grained details.

Subsequently, the enhanced visual-semantic representations $\{TI_i\}_{i=1}^4$ and the corresponding textual features $\{T_i\}_{i=1}^4$ are fed into the Part Sampling (PS) [62] module. This module leverages adaptive attention mechanisms to selectively aggregate salient semantic components from both modalities, resulting in a compact and discriminative fusion feature F_I . Meanwhile, the four-stage textual representation T_{stages} is reverted to its original purified form F_T , ensuring semantic consistency during alignment. The overall operation of the PS module can be expressed as:

$$F_I, F_T, W^* = \Psi_{PS}(TI_{\text{stages}}, T_{\text{stages}}), \quad (6)$$

where $\Psi_{PS}(\cdot)$ denotes the adaptive feature aggregation process performed by the Part Sampling module, and W^* represents

the attention weights computed during sampling, which are further utilized in the alignment objective.

To establish robust cross-modal correspondence, we introduce a similarity-based alignment loss that enforces semantic consistency between fused visual features F_I and their corresponding textual embeddings F_T . Specifically, we compute a cross-modal similarity matrix $M_f \in \mathbb{R}^{B \times H_i \times W_i}$ as:

$$M_f = \sigma(F_I \times F_T), \quad (7)$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, and $\times$ represents the inner product between the two modalities. The similarity matrix M_f captures the degree of local semantic correspondence between visual and textual representations at each spatial location. It guides the alignment of the joint embedding space and promotes learning of modality-invariant features essential for fine-grained classification.

To further refine the multimodal representations, we perform a label-free cross-modal feature fusion by concatenating the similarity scores with the original modality features and projecting them into a shared embedding space:

$$\begin{aligned} \tilde{F}_I &= \phi(\text{cat}(M_f, F_I)), \\ \tilde{F}_T &= \phi(\text{cat}(M_f, F_T)), \end{aligned} \quad (8)$$

where $\text{cat}(\cdot)$ denotes concatenation along the feature dimension, and $\phi(\cdot)$ represents a learnable projection layer shared across modalities. Importantly, the similarity matrix M_f is computed solely from visual and textual features without any access to ground-truth labels or class-specific information. All fusion operations in Eq. 8 are performed prior to classification, and the ground-truth labels are only introduced afterward through the classification loss. The fused multimodal representation $\tilde{F}_I$ is then used as input to the classification head.

Finally, to explicitly align the projected visual and textual embeddings, we define an alignment loss $\mathcal{L}_{it}$ based on the mean squared error (MSE) between them. This loss encourages the fused visual features and their corresponding textual features to remain close within the shared embedding space, promoting better cross-modal consistency. It is formulated as:

$$\mathcal{L}_{it} = \frac{1}{B} \sum_{b=1}^B \left\| \tilde{F}_I^{(b)} - \tilde{F}_T^{(b)} \right\|_2^2, \quad (9)$$

where $\tilde{F}_I^{(b)}$ and $\tilde{F}_T^{(b)}$ represent the projected visual and textual embeddings of the b -th sample, respectively.

Although $\mathcal{L}_{it}$ is implemented as an MSE loss, strong semantic alignment is implicitly enforced by category-level supervision and hierarchical cross-modal interactions, rather than relying on the alignment loss alone.

C. Loss Function

To effectively supervise the multimodal fine-grained visual classification task, we adopt a composite loss function comprising three components: (1) the standard cross-entropy loss $\mathcal{L}_{sva}$, (2) the cross-modal alignment loss $\mathcal{L}_{it}$ introduced in Eq. 9, and (3) a class activation consistency loss $\mathcal{L}_{cam}$ based on Grad-CAM [77], [78]. These loss terms are jointly optimized to enhance classification accuracy, semantic alignment across modalities, and spatial interpretability.

1) *Cross-Entropy Loss*: The fused image representation $\tilde{F}_I$, which encodes cross-modal similarity, is fed into the classification head. It is first normalized by a layer normalization $\text{LN}(\cdot)$, followed by a fully connected layer $\text{FC}(\cdot)$. The final prediction is supervised by the standard cross-entropy loss:

$$\mathcal{L}_{sva} = - \sum_{c=1}^K y_c \log \left(\frac{\exp((\text{FC}(\text{LN}(\tilde{F}_I)))_c)}{\sum_{j=1}^K \exp((\text{FC}(\text{LN}(\tilde{F}_I)))_j)} \right), \quad (10)$$

where K is the number of classes and y_c is the ground-truth.

This objective preserves discriminative power and robustness in the fused representation for fine-grained classification.

2) *Class Activation Consistency Loss*: To encourage the model to attend to semantically meaningful regions, we introduce an auxiliary supervision mechanism based on Grad-CAM [77]. This consistency constraint aligns the model's attention with class-relevant visual evidence.

We begin by constructing a multi-scale fused feature map $F^* \in \mathbb{R}^{H \times W \times C}$ by concatenating the visual-textual interaction features $\{TI_i\}_{i=1}^4$ (see Eq. 5). A ReLU activation $\sigma(\cdot)$ is applied to introduce non-linearity:

$$F^* = \sigma(\text{Concat}(TI_1, TI_2, \dots, TI_s)), \quad s = 4. \quad (11)$$

Following the standard procedure, we calculate the importance weight α_k^c for the k -th channel with respect to the target class score y^c (prior to softmax) by averaging the gradients:

$$\alpha_k^c = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W \frac{\partial y^c}{\partial F_{ij}^k}, \quad (12)$$

where F_{ij}^k is the activation at spatial location (i, j) in channel k of the fused feature map F^* . This weighting reflects the relative influence of each channel on the prediction of class c .

Next, we compute the raw class activation map by linearly combining the feature channels using the weights α_k^c :

$$M_{\text{raw}}^c = \sum_{k=1}^C \alpha_k^c F^k, \quad (13)$$

where $F^k \in \mathbb{R}^{H \times W}$ represents the k -th channel slice of F^* . We then apply ReLU $\sigma(\cdot)$ to retain only features with a positive contribution to class c yielding a class-discriminative attention heatmap $W_{\text{CAM}}^c \in \mathbb{R}^{H \times W}$:

$$W_{\text{CAM}}^c = \sigma(M_{\text{raw}}^c). \quad (14)$$

To encourage spatial consistency, we compare W_{CAM}^c with the attention map $W^* \in \mathbb{R}^{H \times W}$ produced by the part sampling (PS) module. We define the class activation consistency loss as the normalized Euclidean distance between the two:

$$\mathcal{L}_{\text{cam}} = \left\| \frac{W_{\text{CAM}}^c}{\|W_{\text{CAM}}^c\|_2} - \frac{W^*}{\|W^*\|_2} \right\|_2^2, \quad (15)$$

which encourages our model to focus on spatial regions that are both visually and semantically aligned with the target class.

In practice, normalization promotes scale invariance and stabilizes training, especially when handling multi-level semantic features. This consistency regularization, when combined with classification and alignment losses, enhances both localization

TABLE I
DETAILS OF THE SEVEN DATASETS AND
CORRESPONDING LEARNING RATES

Dataset	Train	Test	Class	LR
NABirds [2]	23929	24633	555	5e-4
Food-101 [79]	75751	25249	101	2e-3
ISIC 2018 [80]	7007	3008	7	5e-3
CUB-200-2011 [1]	5994	5794	200	5e-4
FGVC-Aircraft [4]	6667	3333	100	1e-4
Stanford Cars [81]	8144	8041	196	1e-4
Stanford Dogs [3]	12000	8580	120	5e-5

precision and semantic coherence between visual attention and class-relevant textual cues—thereby facilitating more accurate fine-grained multimodal understanding.

3) *Overall Loss*: The overall training loss $\mathcal{L}_{\text{all}}$ integrates three loss components with a weighting hyperparameter λ :

$$\mathcal{L}_{\text{all}} = \lambda \mathcal{L}_{\text{it}} + \mathcal{L}_{\text{sva}} + \mathcal{L}_{\text{cam}}. \quad (16)$$

λ is a dataset-dependent scalar that balances the cross-modal alignment loss $\mathcal{L}_{\text{it}}$ against the classification loss $\mathcal{L}_{\text{sva}}$ and attention consistency loss $\mathcal{L}_{\text{cam}}$. This joint objective promotes accurate classification, semantic alignment between modalities, and interpretable attention maps (See Section IV-D.3).

Our network facilitates hierarchical and interpretable cross-modal interaction, effectively addressing the challenges of semantic redundancy and shallow alignment that are commonly encountered in multimodal fine-grained visual classification.

IV. EXPERIMENTS

A. Datasets and Implementation Details

1) *Datasets*: Comprehensive experiments are conducted on 7 widely used benchmark datasets covering both natural and biomedical domains, as illustrated in Table I.

CUB-200-2011 [1] comprises 11,788 images of birds from 200 species, with an official split of 5,994 training images and 5,794 test images. **Stanford Dogs** [3] includes 20,580 images across 120 dog breeds, with 12,000 images for training and 8,580 for testing. **NABirds** [2] consists of 48,562 images of North American birds from 555 subcategories, split into 23,929 training and 24,633 testing samples. **Food-101** [79] contains 101,000 images spanning 101 food categories, with 75,751 images used for training and 25,249 for testing. **FGVC-Aircraft** [4] comprises 10,000 images of 100 aircraft variants, partitioned into 6,667 training and 3,333 testing samples. **Stanford Cars** [81] includes 16,185 images from 196 types, with 8,144 images used for training and 8,041 for testing. **ISIC 2018** [80] comprises 10,015 dermoscopic images spanning 7 skin lesion categories.

To incorporate textual semantics into our multimodal framework, we pair each image with an auto-generated description using MiniCPM [82], a vision-language model that captures visual attributes such as object appearance, posture, and background. As shown in Table II, dataset-specific templates (e.g., “The bird in the photo...”) are filled with relevant details. However, these descriptions may contain ambiguous or redundant

TABLE II
ILLUSTRATIVE TEXTUAL DESCRIPTIONS AND
CATEGORY-SPECIFIC AMBIGUOUS WORDS

Dataset	Descriptions	Ambiguous Words
Stanford Dogs	“The dog in the photo....”	dog, big, small, pet
CUB-200-2011	“The bird in the photo....”	bird, big, small, branch
NABirds	“The bird in the photo....”	bird, big, small, branch
Food-101	“The food in the photo....”	food, dish, fork, plate
Stanford Cars	“The car in the photo....”	car, road, street
FGVC-Aircraft	“The aircraft in the photo....”	aircraft, sky, ground
ISIC 2018	“The skin lesion in the photo....”	skin, big, small

tokens (e.g., “big”, “on the branch”), which are filtered by the proposed HSP module to enhance discriminative power. In addition to such ambiguous terms, we also eliminate visually irrelevant words such as articles, prepositions, and auxiliary verbs (e.g., “the”, “a”, “in”, “is”) through a dictionary-based purification process applied consistently across all datasets. To validate the robustness and generalizability of our approach across different textual inputs, we further evaluate it with BLIP-2-generated [30] captions. Detailed results are reported in the supplementary material (See Section A.3).

2) *Implementation Details*: We adopt the Swin Transformer Base as the visual backbone and employ the CLIP text encoder to process purified textual descriptions. All images are resized to 384×384 pixels. During training, standard augmentations such as random horizontal flipping and Gaussian blur are applied, while center cropping is used during inference for consistency. The model is trained for 50 epochs using the SGD optimizer with a momentum of 0.9, a weight decay of $1e-8$, and a batch size of 16. Learning rates are tuned separately for each dataset based on validation performance, as detailed in Table I, and optimized using a cosine annealing scheduler with a 5-epoch linear warm-up phase. In the proposed HSP module, high-contribution tokens are empirically selected based on ablation results [Appendix (see supplementary material)]. All experiments are implemented in PyTorch and conducted on NVIDIA RTX 3090 GPUs.

B. Quantitative Comparison

1) *Comparison Results on the Stanford Dogs Dataset*: As presented in Table III, our method achieves a remarkable accuracy of 97.9% on Stanford Dogs, significantly outperforming all state-of-the-art approaches. Compared to CAMF [64], which is also based on the Swin Transformer, our method achieves a notable improvement of 5.1%. Other transformer-based methods such as IELT [23] (91.9%) and TransFG [59] (90.5%) remain far behind. Compared with MP-FGVC [19] (91.0%), which also incorporates textual features with minimal text fine-tuning, our method achieves a 6.9% gain. This improvement primarily stems from the proposed HSP. Unlike MP-FGVC’s prompt tuning, HSP adopts a staged filtering mechanism to discard irrelevant tokens and retain critical tokens. Although MPSA [62] achieves competitive results through multi-scale feature integration, its reliance on single-layer representations limits its ability to model complex patterns. In contrast, our MCA module jointly leverages hierarchical visual features and purified text

TABLE III
COMPARISONS WITH COMPETITIVE METHODS ON STANFORD DOGS. * INDICATES RESULTS REPRODUCED UNDER OUR IMPLEMENTATION USING THE SAME REPORTED SETTINGS

Method	Publication	Dogs (%)
CrossX [53]	CVPR 2019	88.9
FDL [55]	AAAI 2020	84.9
Swin-B [49]	ICCV 2021	92.0
CAMF [64]	SPL 2021	92.8
P2P-Net* [20]	CVPR 2022	86.2
PMG-V2 [58]	TPAMI 2022	90.7
ViT-Net [83]	ICML 2022	90.3
SFFF [56]	TCSVT 2023	88.0
TransFG [59]	AAAI 2023	90.5
IELT [23]	TMM 2023	91.9
ACC-ViT [19]	PR 2024	92.9
MPSA* [62]	TIP 2024	95.4
FAL-ViT [63]	TCSVT 2025	91.1
MP-FGVC [19]	AAAI 2024	91.0
Ours	-	97.9

TABLE IV
COMPARISONS WITH COMPETITIVE METHODS ON CUB-200-2011 AND NABIRDS. * INDICATES RESULTS REPRODUCED UNDER OUR IMPLEMENTATION USING THE SAME REPORTED SETTINGS

Method	Publication	CUB(%)	NABirds(%)
CrossX [53]	CVPR 2019	87.7	86.2
AP-CNN [17]	TIP 2021	88.4	-
Swin-B [49]	ICCV 2021	90.7	90.5
CAMF [64]	SPL 2021	91.2	-
PMG-V2 [58]	TPAMI 2022	90.0	88.4
P2P-Net* [20]	CVPR 2022	90.2	88.7
SFFF [56]	TCSVT 2023	85.4	-
TransFG [59]	AAAI 2023	91.7	90.8
IELT [23]	TMM 2023	91.8	90.8
TransIFC [61]	TIP 2024	91.0	90.9
MPSA* [62]	TIP 2024	92.5	92.2
FAL-ViT [63]	TCSVT 2025	91.7	91.1
AMRFF [57]	TCSVT 2025	87.5	-
MP-FGVC [19]	AAAI 2024	91.8	91.0
Ours	-	92.6	92.5

representations from the HSP module, facilitating multimodal integration and ultimately leading to superior recognition performance.

2) *Comparison Results on the CUB-200-2011 and NABirds Dataset Datasets*: As shown in Table IV, our proposed method achieves state-of-the-art accuracy on two widely used fine-grained benchmarks: 92.6% on CUB-200-2011 and 92.5% on NABirds. Compared to P2P-Net [20], which leverages curriculum learning and an auxiliary detector for part-level proposals, our method improves performance by 2.4% on CUB and 3.8% on NABirds. Although both bird and dog datasets fall under fine-grained animal classification, gains on bird benchmarks are smaller than on Stanford Dogs. This discrepancy may be attributed to the greater variation in birds’ size, pose, and orientation, leading to more ambiguous visual cues. Such cues are difficult to infer directly from visual data, yet they often play a critical role in distinguishing between visually

TABLE V

COMPARISONS WITH COMPETITIVE METHODS ON STANFORD CARS AND FGVC-AIRCRAFT. * INDICATES RESULTS REPRODUCED UNDER OUR IMPLEMENTATION USING THE SAME REPORTED SETTINGS

Method	Publication	Cars(%)	Aircraft(%)
CrossX [53]	CVPR 2019	94.6	92.6
FDL [55]	AAAI 2020	94.0	91.3
AP-CNN [17]	TIP 2021	95.4	94.1
Swin-B [49]	ICCV 2021	95.0	93.9
AKEN-B [84]	TCSVT 2021	93.9	94.0
CAMF [64]	SPL 2021	95.3	93.3
PMG-V2 [58]	TPAMI 2022	95.4	94.1
P2P-Net* [20]	CVPR 2022	95.4	92.4
ViT-Net [83]	ICML 2022	95.0	-
SFFF [56]	TCSVT 2023	94.4	93.1
TransFG [59]	AAAI 2023	94.8	-
IELT [23]	TMM 2023	-	86.8
TransIFC [61]	TIP 2024	94.7	-
MPSA* [62]	TIP 2024	95.2	92.6
AMRFF [57]	TCSVT 2025	94.3	93.5
CoOp [37]	IJCV 2022	68.1	28.8
CoCoOp [36]	CVPR 2022	72.0	28.8
MaPLe [38]	CVPR 2023	73.5	36.5
Ours	-	95.5	94.7

TABLE VI

COMPARISONS WITH COMPETITIVE METHODS ON FOOD-101. * INDICATES RESULTS REPRODUCED UNDER OUR IMPLEMENTATION USING THE SAME REPORTED SETTINGS

Method	Publication	Food(%)
DCL [54]	CVPR 2019	88.9
P2P-Net* [20]	CVPR 2022	89.9
TransFG [59]	AAAI 2023	92.1
IELT [23]	TMM 2023	91.4
DoD [60]	ACM 2023	94.9
MP-FGVC [19]	AAAI 2024	93.0
GL-Swin [85]	EAAI 2024	92.9
MPSA* [62]	TIP 2024	93.0
FAL-ViT [63]	TCSVT 2025	91.8
CoOp [37]	IJCV 2022	85.2
CoCoOp [36]	CVPR 2022	90.1
MaPLe [38]	CVPR 2023	91.4
Ours	-	95.7

similar bird species. Despite these challenges, our method achieves competitive performance compared with representative approaches, including MPSA [62] and TransFG [59], suggesting its effectiveness and good generalization ability for fine-grained visual classification tasks.

3) *Comparison Results on the Food-101 Dataset:* As shown in Table VI, our method achieves 95.7% accuracy on Food-101, demonstrating strong generalization beyond the domain of fine-grained animal images. This large-scale dataset (101,000 images across 101 categories) is challenging due to high intra-class variability, subtle inter-class differences, and complex backgrounds. Despite these challenges, our model significantly outperforms existing state-of-the-art methods. For instance, DoD [60], designed for fine-grained food classification, achieves 94.9% accuracy by leveraging LLMs to

TABLE VII

QUANTITATIVE COMPARISONS ON THE ISIC 2018 [80] DATASET. * INDICATES RESULTS REPRODUCED UNDER OUR IMPLEMENTATION USING THE SAME REPORTED SETTINGS

Method	Publication	ISIC(%)
Conformer-base-p16 [86]	ICCV 2021	82.7
ViT-B 32 [48]	ICLR 2021	77.9
Swin-B [49]	ICCV 2021	79.8
P2P-Net* [20]	CVPR 2022	74.3
ConvNeXt-B [87]	CVPR 2022	79.9
Focal-B [88]	NIPS 2022	79.6
UniFormer-B [89]	TPAMI 2023	82.4
BiFormer-B [90]	CVPR 2023	82.7
MPSA* [62]	TIP 2024	85.9
HiFuse [91]	BSPC 2024	86.0
Ours	-	88.5

construct category-specific subsets. In comparison, our method delivers a 0.8% improvement. Furthermore, compared to recent vision-language models such as CoOp [37], CoCoOp [36], and MaPLe [38], our model achieves improvements of 10.5%, 5.6%, and 4.3%, respectively. These results validate the effectiveness of HSP and MCA in capturing discriminative visual-textual cues, even in highly variable domains like food.

4) *Comparison Results on the Stanford Cars and FGVC-Aircraft Datasets:* As shown in Table V, our proposed method achieves accuracies of 95.5% on Stanford Cars and 94.7% on FGVC-Aircraft. Unlike previous benchmarks involving non-rigid objects such as birds, dogs, and food, these two datasets represent rigid categories with consistent structural patterns. Despite the shift from non-rigid to rigid object categories, our model maintains strong performance across object types. Compared with Swin transformer-based methods such as MPSA [62] (95.2% on Cars and 92.6% on Aircraft) and Swin [49] (95.0% on Cars and 93.9% on Aircraft), our method achieves higher accuracy on both datasets. Additionally, vision-language methods such as CoOp [37] and CoCoOp [36] exhibit lower performance, suggesting limited capacity to capture fine-grained structures. The superior performance of our network can be attributed to HSP for eliminating redundant textual tokens and MCA for hierarchical fusion of visual and purified textual features, enabling effective cross-modal representation learning for rigid objects.

5) *Comparison Results on the ISIC 2018 Dataset:* To further evaluate the generalization of our model in medical imaging, we conduct experiments on ISIC 2018. As shown in Table VII, our method achieves a top-1 accuracy of 88.5%, significantly outperforming all compared approaches. Specifically, it surpasses HiFuse [91], a leading medical classification model, by 2.5%. It also outperforms recent transformer-based models, achieving higher accuracy than BiFormer-B (82.7%) and Focal-B (79.6%). We reproduced MPSA [62] and P2P-Net [20] under identical settings for fair comparison, yielding accuracies of 85.9% and 74.3%, respectively. In addition, hybrid models such as Conformer [86] and UniFormer [89] achieve 82.7% and 82.4%, respectively. Their inferior performance suggests visual features alone are insufficient for fine-grained clinical semantics. To address this, our method incorporates

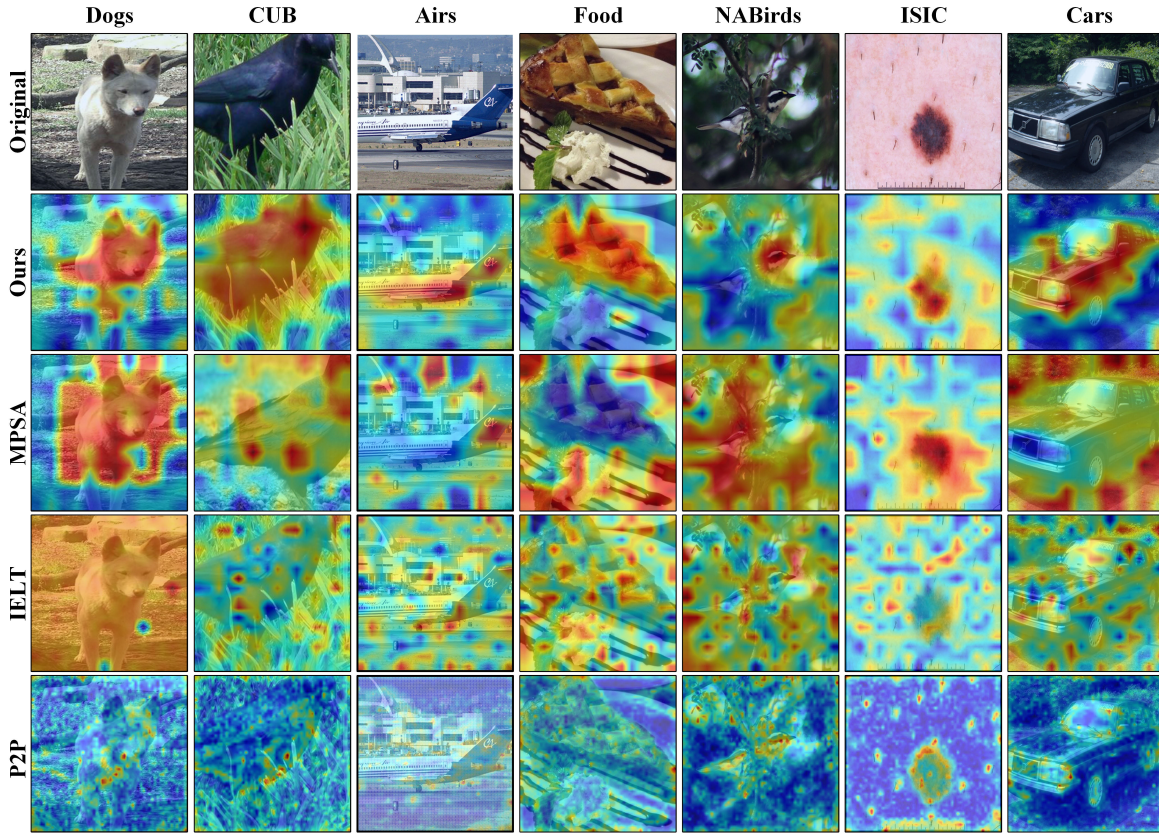


Fig. 5. Visualization of attention maps on 7 popular FGVC datasets.

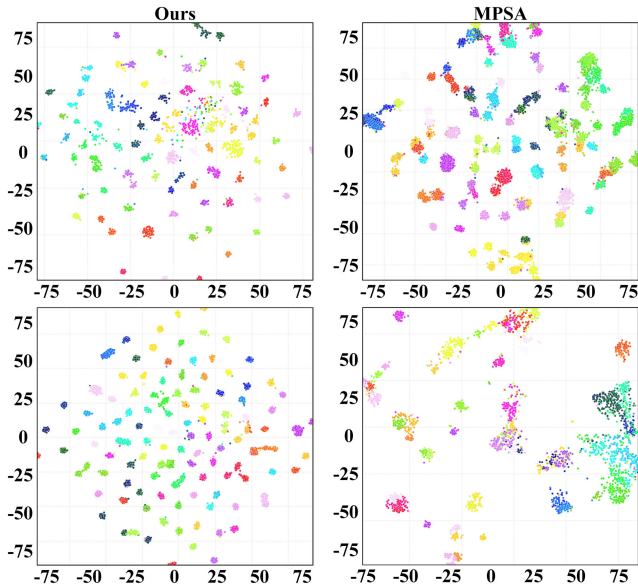


Fig. 6. The left column denotes the t-SNE visualization results of our method on Stanford Dogs (Left Top) and FGVC-Aircraft (Left Down), and the right column represents the t-SNE visualization results of MPSA on Stanford Dogs (Right Top) and FGVC-Aircraft (Right Down).

purified textual cues via HSP and performs hierarchical cross-modal fusion through MCA, yielding robust and generalizable representations. The results on ISIC validate the effectiveness and generalizability of our approach.

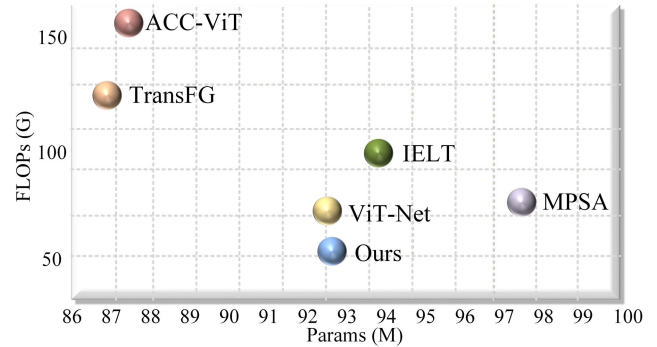


Fig. 7. Computational efficiency comparison across recent multimodal fine-grained classification methods. Our approach achieves competitive accuracy with significantly lower FLOPs, demonstrating superior efficiency.

6) Comparison With SOTA Methods: We compare our method with recent Transformer-based approaches under similar experimental settings. As illustrated in Fig. 7, our model achieves competitive performance with substantially reduced computational cost, requiring only 54.4G FLOPs and 92.4M parameters. In contrast, methods such as ACC-ViT and TransFG exhibit comparable parameter scales but incur significantly higher computational overhead, while MPSA, IELT, and ViT-Net involve either higher FLOPs or noticeably larger parameter counts than our approach. Detailed numerical comparisons are provided in Supplementary Material (see section D.2). These results demonstrate that, by incorporating lightweight cross-modal modules (HSP and MCA), our

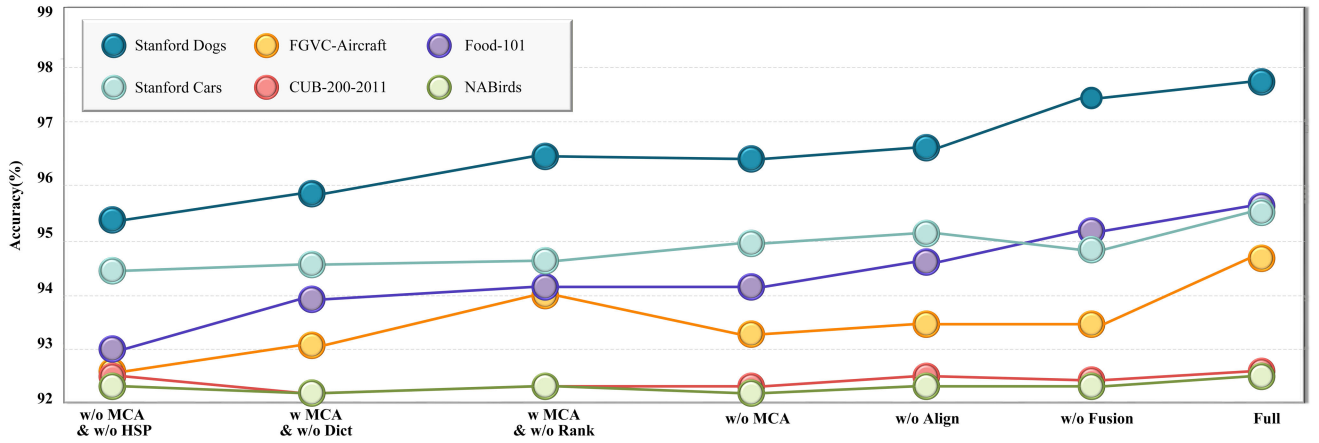


Fig. 8. Accuracy improvements from progressive integration of HSP and MCA components across six datasets. The full model achieves the best performance, confirming the effectiveness of the proposed architecture.

method attains a more favorable balance between accuracy and efficiency, making it particularly well suited for deployment in resource-constrained or real-time scenarios.

C. Visual Analysis

1) *Qualitative Comparison*: Following common practice in fine-grained visual classification, we employ Grad-CAM to visualize the regions attended by the model. Fig. 5 presents Grad-CAM [77] comparisons between our method and three baselines (MPSA, IELT, and P2P-Net) across seven datasets spanning natural and biomedical domains. MPSA and IELT often produce dispersed and imprecise activation maps. In particular, IELT tends to highlight peripheral or background regions, potentially due to weak semantic grounding. While P2P-Net improves localization through self-supervised pose alignment, its attention remains diffuse over non-discriminative areas. In contrast, our method generates compact and well-aligned attention maps, accurately highlighting class-specific regions such as bird beaks (CUB-200-2011), dog muzzles (Stanford Dogs), aircraft engines (FGVC-Aircraft), and lesion contours (ISIC 2018). These improvements stem from HSP, which removes noisy textual features, and MCA, which hierarchically aligns visual and textual representations. Consequently, our model focuses on more discriminative regions, improving fine-grained recognition. These results demonstrate the effectiveness of our multimodal fusion strategy in refining visual attention and enhancing interpretability across different domains. Additional Grad-CAM visualizations and layer-wise analyses are provided in the Appendix (see supplementary material).

2) *Further Comparison*: To evaluate the discriminative power of our learned representations, we visualize embedding space using t-SNE [92] on FGVC-Aircraft and Stanford Dogs datasets. As shown in Fig. 6, our method (left column) achieves clearer inter-class separation with compact, well-isolated clusters. In contrast, MPSA [62] (right column) produces scattered embeddings with noticeable class overlap and less structured distributions. On FGVC-Aircraft, our model distinguishes visually similar variants with minimal

overlap, while MPSA exhibits dispersed and overlapping features, hindering fine-grained differentiation. A similar pattern appears on Stanford Dogs, where our method yields tighter intra-class cohesion and clear decision boundaries, whereas MPSA shows scattered and confusing clusters. These results confirm that our HSP and MCA modules enhance discriminative capacity by structuring performance in quantitative evaluations.

D. Ablation Study

1) *Comprehensive Ablation Analysis of HSP and MCA*: To evaluate the individual contribution of each module, we conduct a joint ablation study on the four key components in our framework: dictionary purification (Dict), contribution ranking (Rank), multi-stage fusion (Fusion), and cross-modal alignment (Align), as summarized in Table VIII. HSP consists of Dict and Rank, while MCA includes Fusion and Align. The first row reports the baseline performance without HSP and MCA, whereas the remaining rows progressively enable each component to evaluate its contribution across seven FGVC datasets. Both Dict and Rank individually improve performance, with Dict showing a greater contribution to text purification. Likewise, Fusion and Align exhibit complementary effects across datasets: Align provides larger gains on Stanford Dogs, whereas Fusion is more beneficial on CUB. Removing both MCA components leads to noticeable performance degradation, confirming the importance of hierarchical cross-modal alignment. As shown in Fig. 8, progressively integrating these modules consistently improves performance, while additional Grad-CAM visualizations in the Appendix (see supplementary material) further illustrate their effectiveness. Overall, the results demonstrate the complementary roles of HSP and MCA and the benefit of their joint optimization.

2) *Alignment Strategy Analysis*: To investigate the optimal interaction mechanism between visual and textual features within the MCA module, we compare four alignment strategies on four datasets, as illustrated in Fig. 9. Specifically, strategy **a** performs direct concatenation ($F_I + F_T$), strategy **b** employs element-wise multiplication ($F_I \times F_T$), strategy **c** adds the fused features back to the textual representations

TABLE VIII

JOINT ABLATION STUDY OF THE PROPOSED HSP AND MCA MODULES ACROSS 7 FINE-GRAINED VISUAL CLASSIFICATION DATASETS. HSP IS A PREREQUISITE FOR MCA; REMOVING HSP INDICATES THAT NO TEXTUAL FEATURES ARE INTRODUCED, AND THUS MCA IS NOT APPLIED

	Module				Dogs(%)	CUB(%)	Aircraft(%)	NABirds(%)	Cars(%)	Food(%)	ISIC(%)
	Dict	Rank	Fusion	Align							
Baseline					95.4	92.5	92.6	92.2	94.4	93.0	85.9
w/o Dict		✓	✓	✓	95.9 +0.5	92.0 -0.5	93.1 +0.5	92.0 -0.2	94.5 +0.1	93.9 +0.9	86.3 +0.4
w/o Rank	✓		✓	✓	96.7 +1.3	92.2 -0.3	93.9 +1.3	92.2 +0.0	94.6 +0.2	94.1 +1.1	87.1 +1.2
w/o MCA	✓	✓			96.5 +1.1	92.2 -0.3	93.3 +0.7	92.0 -0.2	94.9 +0.5	94.1 +1.1	88.1 +2.2
w/o Align	✓	✓	✓		96.8 +1.4	92.5 +0.0	93.5 +0.9	92.2 +0.0	94.8 +1.8	94.5 +1.5	88.4 +2.5
w/o Fusion	✓	✓		✓	97.6 +2.2	92.3 -0.2	93.5 +0.9	92.2 +0.0	94.8 +0.4	95.1 +2.1	87.6 +1.7
Full	✓	✓	✓	✓	97.9 +2.5	92.6 +0.1	94.7 +2.1	92.5 +0.3	95.5 +1.1	95.7 +2.7	88.5 +2.6

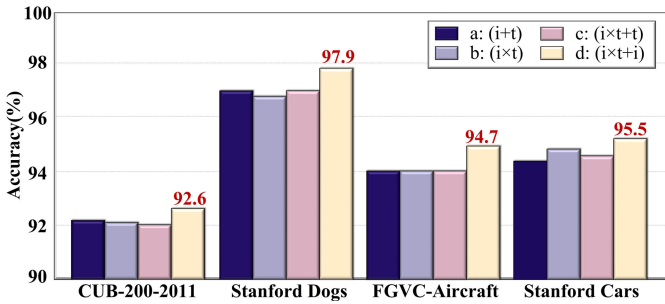


Fig. 9. Comparison of different alignment strategies for cross-modal fusion. The notations a–d correspond to: (a) $F_I + F_T$, (b) $F_I \times F_T$, (c) $F_I \times F_T + F_T$, and (d) $F_I \times F_T + F_I$. Strategy (d) yields the best results across all datasets.

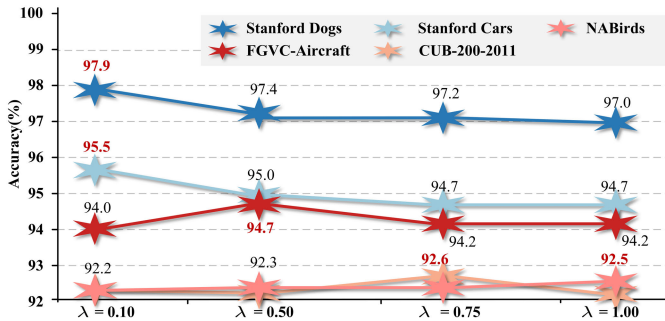


Fig. 10. Analysis of different parameters λ .

($F_I \times F_T + F_T$), and strategy **d** introduces a residual connection to the visual features ($F_I \times F_T + F_I$). Among them, strategy **d** consistently achieves the highest accuracy, reaching 97.9% on Stanford Dogs and 92.6% on CUB-200-2011. This indicates that reinforcing the visual modality through interaction-aware residual connections better preserves discriminative spatial information. In contrast, strategies that update textual features, such as strategy **c**, may be less effective due to the lower spatial resolution of text representations. Therefore, we adopt strategy **d** as the default alignment strategy.

3) *Parameter λ Analysis*: To assess the impact of the alignment loss weight λ on overall performance, we conduct a sensitivity analysis on five datasets. As summarized in Fig. 10, model performance varies with different λ values, indicating that the effect of the alignment loss is dataset-dependent. Stanford Dogs and Stanford Cars achieve their

highest accuracies at $\lambda = 0.10$, suggesting that a smaller contribution of the alignment loss L_{it} is more appropriate, possibly because the visual modality already provides highly discriminative features. In contrast, CUB-200-2011, FGVC-Aircraft, and NABirds achieve their best results with larger λ values (0.75, 0.50, and 1.00, respectively), indicating that stronger visual-textual alignment is more beneficial. These findings highlight the importance of selecting dataset-specific λ values, which are adopted in our final experiments.

4) *Textual Feature Hierarchy Analysis*: To evaluate the contribution of hierarchical textual features in MCA, we conduct an ablation study comparing the full model, which aligns multi-layer textual and visual features, with a simplified variant that performs alignment between the single text embedding F_T and the final-stage visual representation I_4 . The full model achieves 97.9% accuracy on the Stanford Dogs dataset, outperforming the single-layer variant at 94.8%. This performance gap highlights the importance of incorporating multi-level textual semantics, which offer richer contextual cues and more effectively complement multi-scale visual representations. These findings demonstrate that hierarchical cross-modal alignment plays a critical role in enhancing semantic fusion and improving fine-grained classification performance.

V. CONCLUSION

We propose a task-oriented multimodal framework for fine-grained visual classification (FGVC) that tackles two core challenges: redundant textual descriptions and the insufficient utilization of hierarchical semantic features. To this end, we introduce the Hierarchical Semantic Purification (HSP) module, which filters irrelevant or ambiguous tokens to generate concise and discriminative text features, and the Multi-layer Cross-modal Alignment (MCA) module, which facilitates hierarchical fusion between visual and textual modalities. Together, these modules establish robust cross-modal interactions. Extensive experiments on seven fine-grained datasets from natural and biomedical domains validate the effectiveness and generalizability of our approach. Ablation studies further demonstrate the complementary strengths of HSP and MCA. The method depends on the quality of generated descriptions, which may degrade under domain shifts or low-contrast conditions. Moreover, the current pipeline is sequential, with text generation and classification decoupled, limiting joint

optimization. Future work will explore end-to-end architectures that unify generation and alignment, open-vocabulary and few-shot FGVC, and deployment in challenging real-world scenarios. Despite its effectiveness, our method has limitations. Performance may degrade when text descriptions contain malicious misinformation that contradicts visual facts or when images suffer from severe occlusion. Future work will explore cross-modal conflict resolution to address these extreme cases.

REFERENCES

- [1] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.
- [2] G. Van Horn et al., "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2015, pp. 595–604.
- [3] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR Workshop Fine-Grained Vis. Categorization*, 2011, vol. 2, no. 1, pp. 1–2.
- [4] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," Jun. 2013, *arXiv:1306.5151*.
- [5] G. Van Horn et al., "The iNaturalist species classification and detection dataset," 2017, *arXiv:1707.06642*.
- [6] J. Gong et al., "Deep learning-based stage-wise risk stratification for early lung adenocarcinoma in CT images: A multi-center study," *Cancers*, vol. 13, no. 13, p. 3300, Jun. 2021.
- [7] X. Hu et al., "Expert knowledge-guided multi-granularity multi-scale fusion for weakly-supervised histological segmentation," *Inf. Fusion*, vol. 125, Jan. 2026, Art. no. 103432. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1566253525005056>
- [8] F. Zhang et al., "Federated cross-incremental self-supervised learning for medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 7, pp. 13498–13511, Jul. 2025.
- [9] F. Zhang et al., "Cross co-teaching for semi-supervised medical image segmentation," *Pattern Recognit.*, vol. 152, Aug. 2024, Art. no. 110426. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0031320324001778>
- [10] H. Naz, N. J. Ahuja, and R. Nijhawan, "Diabetic retinopathy detection using supervised and unsupervised deep learning: A review study," *Artif. Intell. Rev.*, vol. 57, no. 5, p. 131, May 2024.
- [11] Q. Cai et al., "Multimodal artificial intelligence for disease diagnosis: Advances, applications, and challenges," *Pattern Recognit.*, vol. 178, Oct. 2026, Art. no. 113456.
- [12] Y. Di, Z. Jiang, and H. Zhang, "A public dataset for fine-grained ship classification in optical remote sensing images," *Remote Sens.*, vol. 13, no. 4, p. 747, Feb. 2021, doi: [10.3390/rs13040747](https://doi.org/10.3390/rs13040747).
- [13] J. Wang, G. Xu, F. Yan, J. Wang, and Z. Wang, "Defect transformer: An efficient hybrid transformer architecture for surface defect detection," *Measurement*, vol. 211, Apr. 2023, Art. no. 112614.
- [14] X. Zheng, S. Zheng, Y. Kong, and J. Chen, "Recent advances in surface defect inspection of industrial products using deep learning techniques," *Int. J. Adv. Manuf. Technol.*, vol. 113, nos. 1–2, pp. 35–58, Mar. 2021.
- [15] X.-S. Wei et al., "Fine-grained image analysis with deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8927–8948, Dec. 2022.
- [16] J. Han, X. Yao, G. Cheng, X. Feng, and D. Xu, "P-CNN: Part-based convolutional neural networks for fine-grained visual categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 579–590, Feb. 2022, doi: [10.1109/TPAMI.2019.2933510](https://doi.org/10.1109/TPAMI.2019.2933510).
- [17] Y. Ding et al., "AP-CNN: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification," *IEEE Trans. Image Process.*, vol. 30, pp. 2826–2836, 2021, doi: [10.1109/TIP.2021.3055617](https://doi.org/10.1109/TIP.2021.3055617).
- [18] H. Zheng, J. Fu, Z.-J. Zha, J. Luo, and T. Mei, "Learning rich part hierarchies with progressive attention networks for fine-grained image recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 476–488, 2020, doi: [10.1109/TIP.2019.2921876](https://doi.org/10.1109/TIP.2019.2921876).
- [19] X. Jiang, H. Tang, J. Gao, X. Du, S. He, and Z. Li, "Delving into multimodal prompting for fine-grained visual classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 3, Mar. 2024, pp. 2570–2578.
- [20] X. Yang, Y. Wang, K. Chen, Y. Xu, and Y. Tian, "Fine-grained object classification via self-supervised pose alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7399–7408.
- [21] Y. Ding, Y. Zhou, Y. Zhu, Q. Ye, and J. Jiao, "Selective sparse sampling for fine-grained image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6599–6608, doi: [10.1109/ICCV.2019.00670](https://doi.org/10.1109/ICCV.2019.00670).
- [22] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. Comput. Vis. Pattern Recognit. (ICLR)*, 2021.
- [23] Q. Xu, J. Wang, B. Jiang, and B. Luo, "Fine-grained visual classification via internal ensemble learning transformer," *IEEE Trans. Multimedia*, vol. 25, pp. 9015–9028, 2023, doi: [10.1109/TMM.2023.3244340](https://doi.org/10.1109/TMM.2023.3244340).
- [24] J. Wang, X. Yu, and Y. Gao, "Feature fusion vision transformer for fine-grained visual categorization," 2021, *arXiv:2107.02341*.
- [25] X. He and Y. Peng, "Fine-grained image classification via combining vision and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5994–6002.
- [26] A. Mafra, S. Dey, A. F. Bitten, L. Gomez, and D. Karatzas, "Fine-grained image classification and retrieval by combining visual and locally pooled textual features," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2950–2959.
- [27] A. Mafra, S. Dey, A. F. Bitten, L. Gomez, and D. Karatzas, "Multi-modal reasoning graph for scene-text based fine-grained image classification and retrieval," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 4023–4033.
- [28] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 49–58, doi: [10.1109/CVPR.2016.13](https://doi.org/10.1109/CVPR.2016.13).
- [29] T. B. Brown et al., "Language models are few-shot learners," 2020, *arXiv:2005.14165*.
- [30] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 19730–19742.
- [31] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.
- [32] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," 2023, *arXiv:2304.10592*.
- [33] P. Wang et al., "OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 23318–23340.
- [34] X. Liu, G. Meng, and C. Pan, "Scene text detection and recognition with advances in deep learning: A survey," *Int. J. Document Anal. Recognit. (IJДАР)*, vol. 22, no. 2, pp. 143–162, Jun. 2019.
- [35] L. Zhu, T. Chen, J. Yin, S. See, and J. Liu, "Learning Gabor texture features for fine-grained recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 1621–1631.
- [36] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Sep. 2022, pp. 16816–16825.
- [37] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, Sep. 2022, doi: [10.1007/s11263-022-01653-1](https://doi.org/10.1007/s11263-022-01653-1).
- [38] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "MaPLE: Multi-modal prompt learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19113–19122.
- [39] P. Gao et al., "CLIP-adaptor: Better vision-language models with feature adapters," *Int. J. Comput. Vis.*, vol. 132, no. 2, pp. 581–595, Feb. 2024.
- [40] R. Zhang et al., "Tip-adaptor: Training-free adaption of clip for few-shot classification," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2022, pp. 493–510.
- [41] M. Xu, J. Wang, M. Tao, B.-K. Bao, and C. Xu, "CookGALIP: Recipe controllable generative adversarial CLIPs with sequential ingredient prompts for food image generation," *IEEE Trans. Multimedia*, vol. 27, pp. 2772–2782, 2025.
- [42] J. Xu et al., "Improving fine-grained image classification with multi-modal information," *IEEE Trans. Multimedia*, vol. 26, pp. 2082–2095, 2024.
- [43] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

- [45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
- [46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [47] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 4700–4708.
- [48] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [49] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
- [50] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.
- [51] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2020, pp. 213–229.
- [52] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," 2021, *arXiv:2106.08254*.
- [53] W. Luo et al., "Cross-x learning for fine-grained visual categorization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8242–8251.
- [54] Y. Chen, Y. Bai, W. Zhang, and T. Mei, "Destruction and construction learning for fine-grained image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5157–5166.
- [55] C. Liu, H. Xie, Z.-J. Zha, L. Ma, L. Yu, and Y. Zhang, "Filtration and distillation: Enhancing region attention for fine-grained visual categorization," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 11555–11562, Apr. 2020.
- [56] M. Wang, P. Zhao, X. Lu, F. Min, and X. Wang, "Fine-grained visual categorization: A spatial-frequency feature fusion perspective," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 6, pp. 2798–2812, Jun. 2023.
- [57] Y. Yang, D. Chang, R. Du, Y.-Z. Song, and Z. Ma, "Adaptive multi-resolution feature fusion for fine-grained visual classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 8, pp. 8252–8264, Aug. 2025.
- [58] R. Du, J. Xie, Z. Ma, D. Chang, Y.-Z. Song, and J. Guo, "Progressive learning of category-consistent multi-granularity features for fine-grained visual classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9521–9535, Dec. 2022.
- [59] J. He et al., "TransFG: A transformer architecture for fine-grained recognition," *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 1, pp. 852–860, Jun. 2022.
- [60] J. M. Rodríguez-de Vera et al., "Dining on details: Llm-guided expert networks for fine-grained food recognition," in *Proc. 8th Int. Workshop Multimedia Assist. Dietary Manage.*, 2023, pp. 43–52.
- [61] H. Liu, C. Zhang, Y. Deng, B. Xie, T. Liu, and Y.-F. Li, "TransIFC: Invariant cues-aware feature concentration learning for efficient fine-grained bird image classification," *IEEE Trans. Multimedia*, vol. 27, pp. 1677–1690, Jan. 2025.
- [62] J. Wang, Q. Xu, B. Jiang, B. Luo, and J. Tang, "Multi-granularity part sampling attention for fine-grained visual classification," *IEEE Trans. Image Process.*, vol. 33, pp. 4529–4542, 2024.
- [63] Y. Huang, Z. Hechen, M. Zhou, Z. Li, and S. Kwong, "An attention-locating algorithm for eliminating background effects in fine-grained visual classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 6, pp. 5993–6006, Jun. 2025.
- [64] Z. Miao, X. Zhao, J. Wang, Y. Li, and H. Li, "Complemental attention multi-feature fusion network for fine-grained classification," *IEEE Signal Process. Lett.*, vol. 28, pp. 1983–1987, 2021.
- [65] S. Li et al., "LearnMat: Semantic-aware self-supervision fine-grained visual recognition," *IEEE Trans. Image Process.*, vol. 35, pp. 2555–2569, 2026.
- [66] L. Yang et al., "Dynamic MLP for fine-grained image classification by leveraging geographical and temporal information," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2022, pp. 10945–10954.
- [67] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
- [68] X. Han, X. Zhang, G. Niu, M. Zhou, and Z. Pan, "Soft-label guided multi-granularity prompts learning for human-object interaction detection," *Appl. Soft Comput.*, vol. 192, Apr. 2026, Art. no. 114765. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1568494626002139>
- [69] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 34892–34916.
- [70] X. Zhang, X. Li, M. Zhou, M. Gan, and C. L. Philip Chen, "ASCFormer: An adaptive structure-aware cascaded transformer for 3D object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 36, no. 2, pp. 2218–2232, Feb. 2026.
- [71] F. Huang, S. Song, and L. Zhang, "Gradient harmonization in unsupervised domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10319–10336, Dec. 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:271600933>
- [72] F. Huang et al., "Unsupervised robust domain adaptation: Paradigm, theory and algorithm," *Int. J. Comput. Vis.*, vol. 134, no. 1, Jan. 2026, Art. no. 5, doi: 10.1007/s11263-025-02614-0.
- [73] F. Huang, L. Zhang, Y. Zhou, and X. Gao, "Adversarial and isotropic gradient augmentation for image retrieval with text feedback," *IEEE Trans. Multimedia*, vol. 25, pp. 7415–7427, 2023.
- [74] Y. Fan, C. Arora, and C. Treude, "Stop words for processing software engineering documents: Do they matter?" in *Proc. IEEE/ACM 2nd Int. Workshop Natural Language-Based Softw. Eng. (NLBSE)*, May 2023, pp. 40–47. [Online]. Available: <https://api.semanticscholar.org/CorpusID:257631838>
- [75] F. Alshanik, A. W. Apon, A. Herzog, I. Safro, and J. Sybrandt, "Accelerating text mining using domain-specific stop word lists," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Nov. 2020, pp. 2639–2648. [Online]. Available: <https://api.semanticscholar.org/CorpusID:227305771>
- [76] M. Jia et al., "Visual prompt tuning," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2022, pp. 709–727.
- [77] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
- [78] Y. Shu, B. Yu, H. Xu, and L. Liu, "Improving fine-grained visual recognition in low data regimes via self-boosting attention mechanism," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2022, pp. 449–465.
- [79] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101—mining discriminative components with random forests," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2014, pp. 446–461.
- [80] N. Codella et al., "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*.
- [81] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 554–561.
- [82] Y. Yao et al., "MiniCPM-V: A GPT-4 V level MLLM on your phone," 2024, *arXiv:2408.01800*.
- [83] S. Kim, J. Nam, and B. C. Ko, "Vit-Net: Interpretable vision transformers with neural tree decoder," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 11162–11172.
- [84] Y. Hu, Y. Yang, J. Zhang, X. Cao, and X. Zhen, "Attentional kernel encoding networks for fine-grained visual categorization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 301–314, Jan. 2021.
- [85] J.-H. Kim, N. Kim, and C. S. Won, "Global-local feature learning for fine-grained food classification based on Swin transformer," *Eng. Appl. Artif. Intell.*, vol. 133, Jul. 2024, Art. no. 108248.
- [86] Z. Peng et al., "Conformer: Local features coupling global representations for recognition and detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9454–9468, Aug. 2023.
- [87] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," 2022, *arXiv:2201.03545*.
- [88] J. Yang, C. Li, X. Dai, and J. Gao, "Focal modulation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 4203–4217.
- [89] K. Li et al., "UniFormer: Unifying convolution and self-attention for visual recognition," 2022, *arXiv:2201.09450*.
- [90] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. Lau, "BiFormer: Vision transformer with bi-level routing attention," 2023, *arXiv:2303.08810*.
- [91] X. Huo et al., "HiFuse: Hierarchical multi-scale feature fusion network for medical image classification," *Biomed. Signal Process. Control*, vol. 87, Jan. 2024, Art. no. 105534.
- [92] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.



Menghuan Zhang received the B.Eng. degree from the College of Software, Hebei Normal University, Shijiazhuang, China, in 2023. She is currently pursuing the M.Eng. degree with the Faculty of Information Science and Engineering, Ocean University of China, Qingdao, China, under the supervision of Prof. Qing Cai. Her current research focuses on fine-grained image classification based on vision-language multimodal alignment and fusion. Her research interests include fine-grained image classification and multimodal alignment and fusion.



Junyu Dong (Member, IEEE) received the B.Sc. and M.Sc. degrees from the Department of Applied Mathematics, Ocean University of China, Qingdao, China, in 1993 and 1999, respectively, and the Ph.D. degree in image processing from the Department of Computer Science, Heriot-Watt University, Edinburgh, U.K., in 2003. Currently, he is a Professor and the Vice-Dean of the College of Information Science and Engineering, Ocean University of China. His research interests include computer vision, underwater image processing, and machine learning.



Qing Cai (Member, IEEE) received the M.Sc. and Ph.D. degrees from the Department of Automation, Northwestern Polytechnical University, Xi'an, China, in 2016 and 2019, respectively. He is currently a Professor with the Innovation School of Artificial Intelligence, Hefei University of Technology. His main research interests include computer vision, medical image analysis, image segmentation, 3D reconstruction, deep learning, and machine learning.



David Zhang (Life Fellow, IEEE) received the degree in computer science from Peking University, the M.Sc. and Ph.D. degrees in computer science from Harbin Institute of Technology (HIT), in 1982 and 1985, respectively, and the second Ph.D. degree in electrical and computer engineering from the University of Waterloo, ON, Canada, in 1994. From 1986 to 1988, he was a Postdoctoral Fellow with Tsinghua University and then an Associate Professor with the Academia Sinica, Beijing. Currently, he is the Presidential Chair Professor with The Chinese University of Hong Kong, Shenzhen. So far, he has published over 20 monographs, more than 500 international journal articles, and more than 40 patents from USA/Japan/HK/China. He has been listed as a Global Highly Cited Researchers in Engineering by Clarivate Analytics from 2014 to 2021 (eight years). He is also ranked about 85 with an H-index of 123 at Top 1000 Scientists for international computer science and electronics. He is a Croucher Senior Research Fellow, a Distinguished Speaker of the IEEE Computer Society, and a fellow of the Royal Society of Canada and Canadian Academy of Engineering, and an IAPR/AAI Fellow.



Fan Zhang received the B.Sc. degree from the Department of Electronics and Control Engineering, Chang'an University, in 2018, and the Ph.D. degree from the School of Automation, Northwestern Polytechnical University, Xi'an, China, in 2024. He is currently an Assistant Professor (Research) with the Department of Health Technology and Informatics, The Hong Kong Polytechnic University. His research interests include computer vision, pattern recognition, deep learning, and intelligent systems.



Sheng Chen (Life Fellow, IEEE) received the B.Eng. degree in control engineering from the East China Petroleum Institute, Dongying, China, in January 1982, the Ph.D. degree in control engineering from City University, London, in September 1986, and the D.Sc. degree from the University of Southampton, Southampton, U.K., in August 2005. He is an Emeritus Professor with the University of Southampton, Southampton, U.K. From 1986 to 1999, he held research and academic appointments with the University of Sheffield, the University of Edinburgh, and the University of Portsmouth, all in U.K. In September 1999, he joined the School of Electronics and Computer Science, University of Southampton, and retired in June, 2026. Since then, he has been a University Visitor associated with the School of Electronics and Computer Science, University of Southampton. In 2023, he was appointed as a Distinguished Professor with the Ocean University of China. He has established an over 40-year distinguished academic career in the fields of adaptive signal processing, wireless communications, neural networks, and machine learning. He has authored/co-authored over 750 research papers. He has more than 23 000 Web of Science citations with an H-index of 67, and more than 44 000 Google Scholar citations with an H-index of 88. He was elected to a fellow of United Kingdom Royal Academy of Engineering in 2014. He is a fellow of Asia-Pacific Artificial Intelligence Association, the Industry Academy of the International Artificial Intelligence Industry Alliance, and IET; and one of the original 200 ISI Highly Cited Researcher in engineering (March 2004). He was a recipient of the IEEE ComSoc Signal Processing and Computing for Communications (SPCC) 2025 Technical Recognition Award.



Guanbin Li (Member, IEEE) received the Ph.D. degree from The University of Hong Kong in 2016. He is currently a Professor with the School of Computer Science and Engineering, Sun Yat-sen University. He has authored or co-authored more than 100 papers in top-tier academic journals and conferences. His current research interests include computer vision, image processing, and deep learning. He was a recipient of the ICCV 2019 Best Paper Nomination Award. He serves as the Area Chair for the conference of VISAPP. He has been serving as a reviewer for several academic journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IJCV, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CYBERNETICS, CVPR, ICCV, ECCV, and NeurIPS.