QUATERNION STRUCTURAL-SPECTRAL AGGREGATION FOR TRAINING-FREE FINE-GRAINED LEAF IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Training-free fine-grained leaf image retrieval is attractive for continually expanding cultivar collections because it avoids target-specific annotation and fine-tuning. Existing approaches mainly summarize frozen feature maps using real-valued statistics or local pattern transforms, leaving both intra-patch two-dimensional organization and contour-wise cyclic structure underexplored. We propose Quaternion Structural-Spectral Aggregation (QSSA), a training-free descriptor built on multi-scale boundary-aligned patches and frozen CNN features. For each feature channel, QSSA constructs a quadrant-layout quaternion and a symmetry-discrepancy quaternion to encode absolute spatial layout and paired-region relations. The resulting contour-ordered quaternion sequences are aggregated using a left-sided discrete quaternion Fourier transform (DQFT). Quaternion magnitudes provide invariance to the arbitrary contour starting point, while low-frequency truncation suppresses high-frequency variations associated with local boundary perturbations. Experiments on two publicly available fine-grained leaf image retrieval datasets, SoyCultivar200 and PeanCultivar120, demonstrate that QSSA achieves state-of-the-art or competitive performance without target-specific training.

1 INTRODUCTION

Fine-grained leaf image retrieval (FGLIR) aims to retrieve leaves belonging to the same cultivar within a plant species Chen et al. (2022). In contrast to conventional species-level leaf retrieval Ling & Jacobs (2007); Zhao et al. (2015); Wang et al. (2019); Wang & Gao (2014), FGLIR must distinguish leaves with highly similar global appearance, where the discriminative evidence is often limited to subtle differences in local venation, texture distribution, and regional morphology, as illustrated in Fig. 1. This makes FGLIR considerably more challenging than coarse-grained leaf retrieval.

Cultivar collections continually expand, while large-scale and consistently annotated retrieval datasets are difficult to maintain Liu et al. (2020). Repeatedly adapting task-specific models to newly introduced cultivars is therefore costly and often impractical, motivating *training-free* retrieval based on off-the-shelf visual models without target-specific fine-tuning or additional supervision.

Frozen deep features have shown promising performance for training-free FGLIR Chen et al. (2024). Conventional retrieval methods typically aggregate convolutional activations through pooling or weighting Gong et al. (2014); Pang et al. (2019); Kalantidis et al. (2016); Tolias et al. (2016), often treating them as unordered local descriptors and discarding their spatial organization. More recent approaches explicitly exploit structural cues. For example, SDePR Chen et al. (2024) compares symmetric deep patch pairs to capture spatial relations in intermediate activations.

However, existing training-free structural representations remain predominantly *real-valued* and typically encode selected local relations through pairwise differences, local pattern codes, or grouped statistics. They do not jointly preserve the two-dimensional organization within each deep feature channel and the ordered evolution of local structures along the leaf contour. An effective descriptor should therefore model both intra-patch spatial structure and contour-level cyclic regularity.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

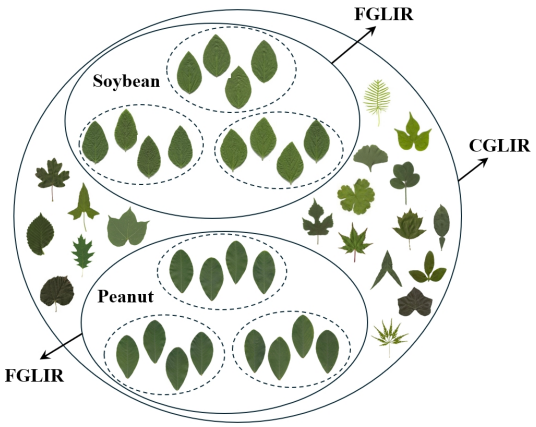


Figure 1: FGLIR distinguishes cultivars among visually similar leaves, whereas coarse-grained leaf image retrieval (CGLIR) distinguishes species.

To this end, we propose *Quaternion Structural-Spectral Aggregation* (QSSA), a fully training-free framework for boundary-aware fine-grained leaf retrieval. Unlike conventional quaternion image representations that mainly couple raw color channels Sangwine (1996); Bihan & Sangwine (2003); Angulo (2009); Li et al. (2015), QSSA uses quaternions to encode the spatial organization of frozen deep feature channels. Specifically, each boundary-aligned channel map is represented by two complementary quaternions: a quadrant-layout quaternion encoding absolute regional responses and a symmetry-discrepancy quaternion encoding horizontally and vertically paired relations.

The resulting quaternion descriptors are arranged in contour order and aggregated using a left-sided discrete quaternion Fourier transform (DQFT) Ell et al. (2014); Ell (2013). Channel-wise quaternion magnitudes remove phase variation caused by the arbitrary contour starting point, while low-frequency retention suppresses high-frequency variations associated with boundary perturbations and yields a compact representation. Thus, QSSA jointly captures intra-patch spatial organization and contour-level cyclic structure without target-specific training.

The main contributions of this work are summarized as follows:

- We propose **QSSA**, a training-free, boundary-aware framework that jointly models intra-patch structure and contour-level cyclic regularity for fine-grained leaf image retrieval.
- We introduce a **dual-branch quaternion encoding** of frozen CNN feature maps, combining quadrant-layout and symmetry-discrepancy representations.
- We develop a **left-sided DQFT aggregation** scheme for contour-ordered quaternion descriptors, providing starting-point invariance through quaternion magnitudes and compact spectral aggregation through low-frequency retention.

Experiments on multiple cultivar-level leaf retrieval benchmarks show that QSSA achieves state-of-the-art or competitive performance against handcrafted, training-free structural, and deep retrieval baselines.

2 Related Work

2.1 Training-Free Deep Image Retrieval

Training-free image retrieval constructs descriptors directly from pre-trained visual models without task-specific optimization Wan et al. (2014); Razavian et al. (2014); Babenko & Lempitsky (2015); Chen et al. (2023a). Existing methods typically aggregate convolutional activations through global pooling Babenko & Lempitsky (2015); Tolias et al. (2016), feature weighting Kalantidis et al. (2016); Pang et al. (2019), or orderless encoding Gong et al. (2014); Forcen et al. (2020), and have shown strong performance in generic image retrieval Wei et al. (2017); Yang et al. (2021); Lee et al. (2023); Shao et al. (2023). However, such aggregation often weakens spatial organization, which is

important for fine-grained retrieval where discriminative cues are localized and structurally arranged Wei et al. (2017); Lee et al. (2023); Wang et al. (2024). QSSA retains the training-free setting while explicitly preserving spatial structure and boundary-induced descriptor order.

2.2 Deep Structural Representations for FGLIR

Structural modeling is particularly relevant to FGLIR, where cultivar collections continue to expand while annotated data for repeated adaptation remain limited Korir et al. (2013); Liu et al. (2020). SDePR Chen et al. (2024) extracts structural cues from frozen intermediate features through symmetric patch comparison and multi-scale self-similarity. However, these representations remain primarily real-valued and do not jointly model the two-dimensional organization of individual feature channels and the cyclic evolution of local structures along the leaf boundary. QSSA addresses these two levels of structure within a unified quaternion structural-spectral representation.

2.3 Quaternion-Based Image Representation

Quaternion representations provide an algebraic framework for jointly modeling coupled image components and have been widely used in color image analysis, including quaternion Fourier transforms Sangwine (1996), quaternion PCA Bihan & Sangwine (2003), structure-tensor modeling Angulo (2009), and quaternion Fourier descriptors Li et al. (2015). These methods primarily construct quaternions from raw image or color components. In contrast, QSSA uses quaternions to couple spatial regions within individual frozen deep feature channels and applies a left-sided DQFT Ell et al. (2014); Ell (2013) to model their ordered evolution along the closed leaf contour.

3 Proposed Method

QSSA constructs a training-free descriptor through three stages: boundary-aligned patch sampling, dual-branch quaternion encoding, and contour-wise spectral aggregation. The backbone remains frozen throughout.

3.1 Boundary-Aligned Multi-Scale Patch Rings

Let I denote a leaf image and Ω its closed boundary. We uniformly sample N contour points in counterclockwise order and write them as $(Z_0, \dots, Z_{N-1})$. For each Z_i, a shape-dependent reference scale is obtained from its maximum Euclidean distance to the remaining contour points. At scale $k \in \{1, \dots, K_s\}$, the patch radius is

$$\rho_{i,k} = 2^{-k} \max_{j \neq i} \|Z_i - Z_j\|_2 . \tag{1}$$

The circle centered at Z_i with radius $\rho_{i,k}$ intersects Ω on both sides of the sampled point. The bisector of the nearest clockwise and counterclockwise intersection chords defines the local orientation $\theta_{i,k}$. We crop a square patch $P_{i,k}$ around Z_i, align it with $\theta_{i,k}$, and size it proportionally to $\rho_{i,k}$. For each scale, the patches form a cyclic ring $\mathcal{P}_k = (P_{0,k}, \dots, P_{N-1,k})$ in contour order. Repeating this construction over K_s scales captures local morphology at different geometric extents while preserving the sequence required for spectral aggregation.

3.2 Two-Branch Quaternion Structural Encoding

Each patch is resized to the backbone input resolution and passed through a frozen CNN. Let ϕ_ℓ denote feature extraction at layer ℓ. The resulting tensor is

$$F_{i,k}^{(\ell)} = \phi_\ell(P_{i,k}) \in \mathbb{R}^{H \times W \times C}, \tag{2}$$

where $X_{i,k,c}^{(\ell)} = F_{i,k}^{(\ell)}(:,:,c)$ denotes the c-th channel map. As illustrated in Fig. 3, each channel map is partitioned into four shared spatial quadrants and processed by two complementary branches. The quadrant-layout branch encodes absolute regional responses, while the symmetry-discrepancy branch captures aligned relations between horizontally and vertically paired quadrants. Applying both branches channel-wise produces two quaternion structural vectors. For readability, the indices (i, k, c, ℓ) are omitted in the following definitions.

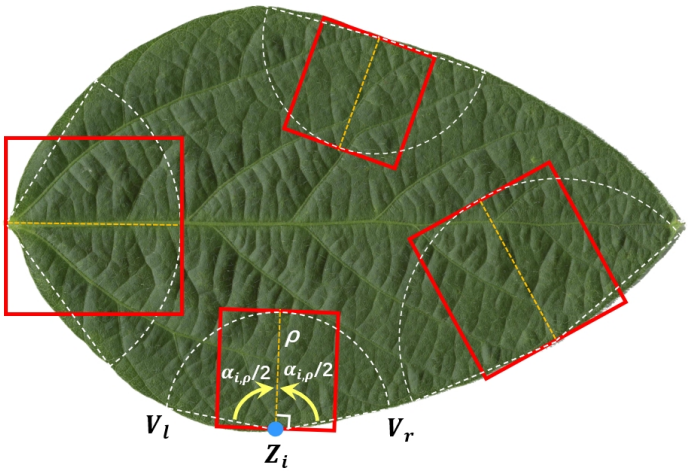


Figure 2: Boundary-aligned patch extraction at scale $k = 2$. Patches are centered at sampled contour points and aligned with their local boundary orientations.

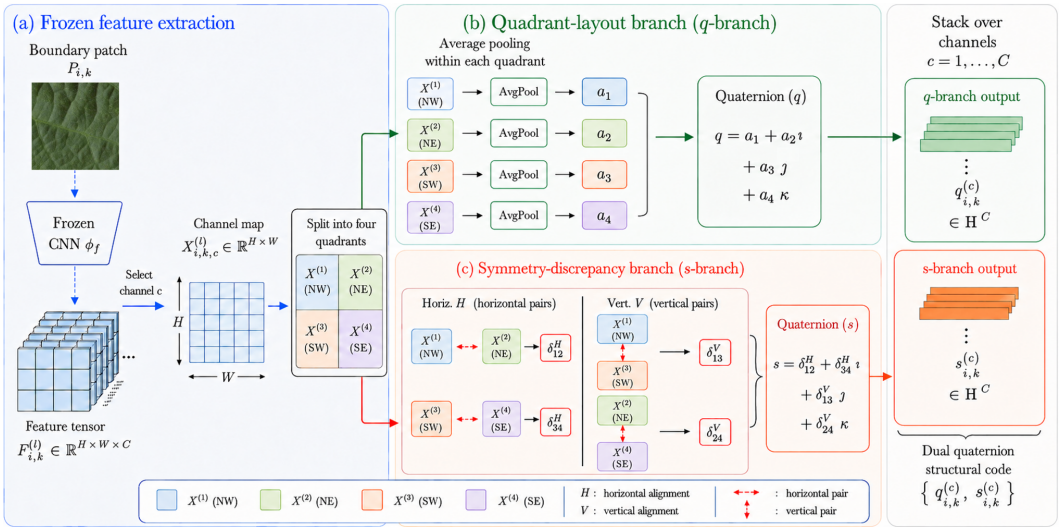


Figure 3: Dual-branch quaternion structural encoding. Each frozen CNN channel map is partitioned into four quadrants. The quadrant-layout branch encodes absolute regional responses, while the symmetry-discrepancy branch captures paired-region relations. Stacking the channel-wise quaternions yields two complementary structural vectors.

Quadrant-layout quaternion. We divide a channel map into four center-aligned quadrants $X^{(1)}, \ldots, X^{(4)}$, ordered as top-left, top-right, bottom-left, and bottom-right. With quaternion basis $(e_1, e_2, e_3, e_4) = (1, \imath, \jmath, \kappa)$, the pooled responses are encoded as

$$q = \sum_{t=1}^{4} \mathrm{Avg}\left(X^{(t)}\right) e_t. \tag{3}$$

The four components encode the coarse two-dimensional layout of one deep feature channel rather than color information. Restoring the indices and stacking all channels yields $\mathbf{q}^{(\ell)}_{i,k} \in \mathbb{H}^C$.

Symmetry-discrepancy quaternion. Quadrant averages do not directly describe local symmetry breaking. We compare pairs $(1, 2)$ and $(3, 4)$ after horizontal flipping, and pairs $(1, 3)$ and $(2, 4)$

after vertical flipping. Let $\mathcal{H}$ and $\mathcal{V}$ denote these operators. For aligned maps $A, B \in \mathbb{R}^{h\times w}$, define

$$\delta(A, B; \mathcal{T}) = \frac{1}{hw} \|A - \mathcal{T}(B)\|_F^2, \tag{4}$$

where $\mathcal{T} \in \{\mathcal{H}, \mathcal{V}\}$, with $\mathcal{H}$ and $\mathcal{V}$ denoting horizontal and vertical flipping, respectively; $\|\cdot\|_F$ is the Frobenius norm, and $h \times w$ is the quadrant size.

The four aligned discrepancies are organized as

$$\begin{aligned} s = \delta\Big(X^{(1)}, X^{(2)}; \mathcal{H}\Big) + \delta\Big(X^{(3)}, X^{(4)}; \mathcal{H}\Big)\, \boldsymbol{\imath} \\ + \delta\Big(X^{(1)}, X^{(3)}; \mathcal{V}\Big)\, \boldsymbol{\jmath} + \delta\Big(X^{(2)}, X^{(4)}; \mathcal{V}\Big)\, \boldsymbol{\kappa}. \end{aligned} \tag{5}$$

This branch captures relational structure complementary to the absolute quadrant layout. Restoring the indices and stacking the channels gives $\mathbf{s}_{i,k}^{(\ell)} \in \mathbb{H}^C$. For subsequent aggregation, we use $\mathbf{z}_{i,k}^{(\ell,b)}$ to denote either branch, where $b \in \{q, s\}$ and $\mathbf{z}_{i,k}^{(\ell,q)} = \mathbf{q}_{i,k}^{(\ell)}$, $\mathbf{z}_{i,k}^{(\ell,s)} = \mathbf{s}_{i,k}^{(\ell)}$. The two branches remain separate so that each retains its own contour-wise spectral distribution.

3.3 Boundary-Ring Quaternion Spectral Aggregation

At scale k, the descriptors of branch b follow the counterclockwise boundary order. We arrange them into $\mathbf{Z}_k^{(\ell,b)} \in \mathbb{H}^{N\times C}$, whose n-th row is $\mathbf{z}_{n,k}^{(\ell,b)\top}$. Each column is therefore a quaternion sequence describing the evolution of one feature channel along the boundary ring.

Let $\mu \in \mathbb{H}$ be a fixed unit pure quaternion satisfying $\mu^2 = -1$. The column-wise left-sided discrete quaternion Fourier transform is

$$\widehat{\mathbf{Z}}_k^{(\ell,b)}(u, :) = \sum_{n=0}^{N-1} e^{-\mu 2\pi u n/N} \mathbf{Z}_k^{(\ell,b)}(n, :), \tag{6}$$

for $u = 0, \ldots, N-1$. Left multiplication fixes the transform convention and preserves the coupling among the four components of each quaternion sample.

Changing the contour starting point by τ positions circularly shifts the descriptor sequence. We denote the shifted sequence by $\mathbf{Z}_{k,\tau}^{(\ell,b)}$, defined as

$$\mathbf{Z}_{k,\tau}^{(\ell,b)}(n, :) = \mathbf{Z}_k^{(\ell,b)}\big((n-\tau) \bmod N, :\big), \quad n = 0, \ldots, N-1, \tag{7}$$

where $\tau \in \{0, \ldots, N-1\}$ denotes the circular shift induced by selecting a different contour starting point.

We use $\widehat{\mathbf{Z}}_{k,\tau}^{(\ell,b)}$ to denote its left-sided DQFT. Let $\mathrm{qmag}(\cdot)$ compute quaternion magnitudes component-wise across channels; for $x = a + b\boldsymbol{\imath} + c\boldsymbol{\jmath} + d\boldsymbol{\kappa}$, $\mathrm{qmag}(x) = \sqrt{a^2+b^2+c^2+d^2}$. A circular shift introduces only a unit-quaternion phase factor, and therefore

$$\mathrm{qmag}\Big(\widehat{\mathbf{Z}}_{k,\tau}^{(\ell,b)}(u, :)\Big) = \mathrm{qmag}\Big(\widehat{\mathbf{Z}}_k^{(\ell,b)}(u, :)\Big). \tag{8}$$

Equation (8) shows that the channel-wise quaternion magnitude spectrum is invariant to the contour starting point. A complete derivation of the starting-point invariance is provided in Appendix A.1.

High-frequency coefficients are more sensitive to local boundary jitter and patch misalignment. For each retained frequency $u = 0, \ldots, K_f - 1$, we first obtain its real-valued channel magnitude vector and then normalize it:

$$\begin{aligned} \mathbf{r}_{k,u}^{(\ell,b)} &= \mathrm{qmag}\Big(\widehat{\mathbf{Z}}_k^{(\ell,b)}(u, :)\Big), \\ \mathbf{m}_{k,u}^{(\ell,b)} &= \frac{\mathbf{r}_{k,u}^{(\ell,b)}}{\Big\|\mathbf{r}_{k,u}^{(\ell,b)}\Big\|_2}. \end{aligned} \tag{9}$$

Here, $\mathbf{r}_{k,u}^{(\ell,b)} \in \mathbb{R}_+^C$ contains one quaternion magnitude per channel, where $\mathbb{R}_+ = [0, \infty)$. The operator $\|\cdot\|_2$ denotes the Euclidean norm of this real-valued vector. Retaining only low frequencies suppresses unstable fine-scale variation.

The layer-specific image descriptor concatenates the retained frequencies from both branches and all scales:

$$\mathbf{d}^{(\ell)}(I) = concat_{k,b,u}\mathbf{m}_{k,u}^{(\ell,b)}, \tag{10}$$

where $k = 1, \ldots, K_s$, $b \in \{q, s\}$, and $u = 0, \ldots, K_f - 1$, so the descriptor dimensionality is $2K_sK_fC$. Images are finally ranked using the fused ℓ_1 dissimilarity.

4 Experiments

4.1 Datasets and Evaluation Protocol

We conduct experiments on two cultivar-level leaf retrieval benchmarks characterized by high visual similarity between classes and subtle local structural differences.

SoyCultivar200. SoyCultivar200 Wang et al. (2020) contains leaf images from 200 soybean cultivars. Following the standard protocol, we evaluate QSSA on three single-leaf subsets, denoted by *Soy-Up*, *Soy-Mid*, and *Soy-Low*, corresponding to leaves collected from upper, middle, and lower plant positions, respectively. We additionally report results on *Soy-Joint*, which groups leaves from different plant positions according to cultivar.

PeanCultivar120. PeanCultivar120 Chen et al. (2023b) is a cultivar-level peanut leaf retrieval dataset. Its distinct plant species and acquisition conditions provide an additional benchmark for evaluating transfer across cultivar-level leaf retrieval settings.

Evaluation metrics. Following prior fine-grained leaf retrieval studies Chen et al. (2024; 2023b), we adopt the Bulls-eye score Ling & Jacobs (2007); Wang & Gao (2014) as the primary metric. For each query, the top-$2C_q$ retrieved images are evaluated, where C_q denotes the number of relevant database images for that query. The Bulls-eye score measures the percentage of relevant samples retrieved within this range. We additionally report precision–recall curves to evaluate ranking quality across different recall levels.

4.2 Implementation Details

Training-free protocol. All experiments follow a strictly training-free protocol. The visual backbone remains frozen, and no target-dataset images or labels are used for fine-tuning, representation learning, or hyperparameter selection. Unless otherwise stated, the default configuration is fixed a priori as $N = 64$, $K_s = 4$, $K_f = 4$, with VGG-16 as the frozen backbone. Retrieval is performed by nearest-neighbor matching in the resulting descriptor space.

Boundary patch construction. For each leaf image, we extract its contour, uniformly sample $N = 64$ boundary points, and construct boundary-aligned patches at $K_s = 4$ geometric scales. Each patch is resized to the input resolution required by the visual backbone.

Frozen feature extraction. Unless otherwise stated, VGG-16 is used as the frozen backbone. Feature maps are extracted from *relu4_3* and *relu5_3*. To reduce descriptor dimensionality and computational cost, every four adjacent channels are averaged, resulting in $C = 128$ channels. An ℓ_1 distance is computed independently for each selected layer, and the resulting layer-wise distances are fused using KNN-HDFF Chen et al. (2022).

Quaternion spectral aggregation. For each scale and quaternion branch, the contour-ordered patch descriptors are transformed using the left-sided DQFT with the fixed transform axis $\mu = \boldsymbol{\imath}$. We retain the first $K_f = 4$ low-frequency components, compute their channel-wise quaternion magnitudes, and apply ℓ_2 normalization to each channel-magnitude vector.

Table 1: Bulls-eye scores (%) of the competing methods on the four SoyCultivar200 subsets. The best result in each column is shown in bold, and the second-best result is underlined.

Method	Soy-Up	Soy-Mid	Soy-Low	Soy-Joint
SBT Chen et al. (2022)	47.57	48.40	47.92	81.69
FB-BDP Chen et al. (2023b)	49.12	51.76	50.59	84.20
HeW Pang et al. (2019)	48.07	49.13	45.23	80.42
DOLG Yang et al. (2021)	56.61	55.90	56.14	88.39
SENet Lee et al. (2023)	61.04	58.19	53.94	90.66
CVNet Lee et al. (2022)	56.96	55.51	52.11	85.31
SuperGlobal Shao et al. (2023)	54.43	55.99	53.74	89.49
Whiten-MTD Ma et al. (2024)	52.68	52.28	52.08	87.61
SDePR Chen et al. (2024)	<u>65.59</u>	<u>64.28</u>	<u>59.91</u>	**92.64**
dino.txt Jose et al. (2025)	42.60	43.78	44.26	78.08
FoL-global Wang et al. (2025)	48.96	50.31	47.24	81.31
Talk2DINO Barsellotti et al. (2025)	55.92	56.02	53.19	86.34
Proposed QSSA	**68.23**	**66.26**	**61.15**	<u>92.52</u>

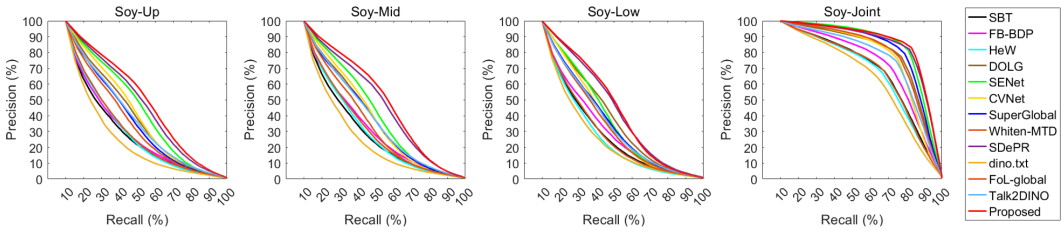


Figure 4: Precision–recall curves of the competing methods on the Soy-Up, Soy-Mid, Soy-Low, and Soy-Joint subsets of SoyCultivar200.

Hardware. Experiments are conducted on an Intel Core i9-13900HX CPU and an NVIDIA GeForce RTX 4060 GPU. Detailed runtime and storage analyses are provided in Appendix A.4.

4.3 Comparison with State-of-the-Art Methods

We compare QSSA with 12 representative methods. SBT Chen et al. (2022) and FB-BDP Chen et al. (2023b) are handcrafted fine-grained leaf retrieval descriptors. HeW Pang et al. (2019), DOLG Yang et al. (2021), SENet Lee et al. (2023), CVNet Lee et al. (2022), SuperGlobal Shao et al. (2023), and Whiten-MTD Ma et al. (2024) are deep image retrieval methods. SDePR Chen et al. (2024) is a training-free structural method designed for fine-grained leaf retrieval. We further include dino.txt Jose et al. (2025), FoL-global Wang et al. (2025), and Talk2DINO Barsellotti et al. (2025), which construct retrieval representations from pre-trained visual or vision-language models.

Results on SoyCultivar200. As shown in Table 1 and Fig. 4, QSSA achieves Bulls-eye scores of 68.23%, 66.26%, 61.15%, and 92.52% on Soy-Up, Soy-Mid, Soy-Low, and Soy-Joint, respectively. It ranks first on all three single-leaf subsets, outperforming SDePR by 2.64, 1.98, and 1.24 percentage points, while remaining comparable on Soy-Joint (92.52% vs. 92.64%). Averaged across the four subsets, QSSA reaches 72.04%, exceeding SDePR by 1.44 points. These results, together with the precision–recall curves, show that jointly modeling intra-patch spatial organization and contour-wise structural order yields a more discriminative representation for subtle cultivar-level morphology.

Results on PeanCultivar120. As shown in Table 2 and Fig. 5, QSSA achieves a Bulls-eye score of 66.70%, outperforming SDePR by 6.77 percentage points and CVNet by 11.47 points. The precision–recall curves further confirm its stronger ranking quality across recall levels. These re-

Table 2: Bulls-eye scores (%) of the competing methods on PeanCultivar120. The best result is shown in bold, and the second-best result is underlined.

Method	Score (%)	Method	Score (%)
SBT Chen et al. (2022)	46.83	DOLG Yang et al. (2021)	53.77
SuperGlobal Shao et al. (2023)	49.67	dino.txt Jose et al. (2025)	47.87
FB-BDP Chen et al. (2023b)	51.80	SENet Lee et al. (2023)	54.53
Whiten-MTD Ma et al. (2024)	51.17	FoL-global Wang et al. (2025)	49.30
HeW Pang et al. (2019)	49.27	CVNet Lee et al. (2022)	55.23
SDePR Chen et al. (2024)	<u>59.93</u>	Talk2DINO Barsellotti et al. (2025)	52.47
Proposed QSSA	**66.70**		

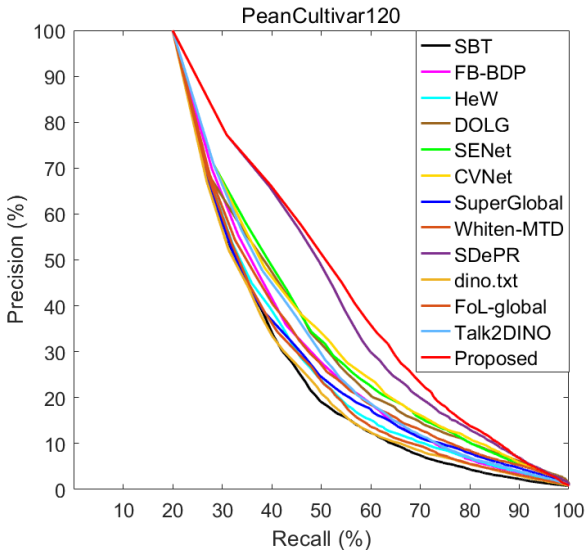


Figure 5: Precision–recall curves of the competing methods on PeanCultivar120.

sults show that the proposed structural-spectral representation transfers effectively from soybean to the evaluated peanut benchmark and remains effective across different cultivar-level leaf retrieval benchmarks.

4.4 Ablation Studies

Unless otherwise stated, all ablations are conducted on PeanCultivar120 using $N = 64$, $K_s = 4$, $K_f = 4$, and VGG-16. We evaluate parameter sensitivity, backbone robustness, structural encoding, and contour-sequence aggregation. Additional analyses of contour localization robustness and computational efficiency are provided in Appendix A.4.

Parameter sensitivity. Table 3 summarizes the effects of contour sampling density N, the number of geometric scales K_s, and the number of retained frequencies K_f. Increasing N from 16 to 64 improves the Bulls-eye score from 65.90% to 66.70%, while $N = 128$ provides no further gain. Increasing K_s from 1 to 4 steadily improves performance from 60.47% to 66.70%, indicating that multi-scale boundary patches provide complementary structural cues. For K_f, performance peaks at 67.97% with $K_f = 8$; however, we use $K_f = 4$ to reduce descriptor dimensionality with only a modest performance loss. These results support the default configuration $N = 64$, $K_s = 4$, and $K_f = 4$ used throughout the experiments.

Frozen visual backbones. We evaluate QSSA with VGG-16, ResNet-50 He et al. (2016), DenseNet-201 Huang et al. (2017), ShuffleNet Zhang et al. (2018), and CLIP-RN50 Radford et al.

Table 3: Sensitivity to N, K_s, and K_f on PeanCultivar120.

Parameter	Setting / Bulls-eye score (%)				
N	16 / 65.90	32 / 66.40	64 / **66.70**	128 / 66.50	–
K_s	1 / 60.47	2 / 64.53	3 / 66.03	4 / **66.70**	–
K_f	1 / 63.40	2 / 65.03	4 / 66.70	8 / **67.97**	16 / 67.63

Table 4: Backbone comparison on PeanCultivar120.

Backbone	VGG-16	ResNet-50	DenseNet-201	ShuffleNet	CLIP-RN50
Score (%)	**66.70**	65.40	65.47	64.90	65.90

Table 5: Component ablations on PeanCultivar120.

Component	Variant / Score (%)				
Spatial encoding	1-ary / 58.40	H-2ary / 62.10	V-2ary / 63.33	Real-4ary / 63.11	Quat.-4ary / **66.70**
Quaternion branches	q / 66.12	s / 65.10	$q + s$ / **66.70**		
Contour aggregation	Q-Avg / 62.53	Q-Max / 62.19	DCT / 47.56	DQFT / **66.70**	

(2021). For each backbone, features are taken from the last two convolutional stages or their closest counterparts under the same training-free protocol. As shown in Table 4, QSSA achieves Bulls-eye scores between 64.90% and 66.70%, with a maximum variation of 1.80 points. The default VGG-16 backbone achieves the highest score among the evaluated architectures, while the relatively small variation indicates limited sensitivity to backbone choice. Detailed feature-layer and channel-reduction settings are provided in Appendix A.2.

Structural encoding and branch complementarity. Table 5 summarizes the main component ablations. Increasing the spatial arity from 1-ary pooling to Quat.-4ary encoding improves the score from 58.40% to 66.70%. The matched Real-4ary variant achieves 63.11%, while Quat.-4ary improves this result by 3.59 points, supporting the benefit of joint quaternion-domain modeling over independent real-component processing. The q- and s-branches achieve 66.12% and 65.10% individually, while their combination reaches 66.70%, indicating complementary information from absolute layout and symmetry discrepancy. Detailed definitions and controlled settings of the ablation variants are provided in Appendix A.3.

Contour-sequence aggregation. DQFT is compared with quaternion average pooling (Q-Avg), quaternion max pooling (Q-Max), and component-wise DCT under identical settings. As shown in Table 5, DQFT achieves 66.70%, outperforming Q-Avg, Q-Max, and DCT by 4.17, 4.51, and 19.14 percentage points, respectively. These results support the effectiveness of quaternion-domain spectral aggregation for modeling cyclic contour sequences.

5 Conclusion

We presented Quaternion Structural-Spectral Aggregation (QSSA), a training-free framework for fine-grained leaf image retrieval. QSSA combines boundary-aligned multi-scale sampling, dual-branch quaternion structural encoding, and low-frequency left-sided DQFT aggregation to capture intra-patch spatial structure and contour-level cyclic regularity. Quaternion magnitudes provide invariance to the arbitrary contour starting point, while low-frequency retention suppresses high-frequency variations associated with local boundary perturbations. Experiments on SoyCultivar200 and PeanCultivar120, together with extensive ablations, demonstrate the effectiveness of the proposed representation. These results suggest that quaternion structural-spectral modeling provides a promising training-free alternative for contour-dominant fine-grained retrieval.

REPRODUCIBILITY STATEMENT

The complete QSSA construction and experimental protocol are described in Sections 3–4. Appendix A.1 provides the derivation of contour starting-point invariance, Appendices A.2–A.3 give detailed backbone configurations and controlled ablation settings, and Appendix A.4 reports robustness and computational analyses.

AI USE STATEMENT

ChatGPT (GPT-5.6) was used for language polishing to improve the readability of the manuscript. We have reviewed all AI-assisted work and take responsibility for the final content of this work.

REFERENCES

Jesús Angulo. Structure tensor of colour quaternion image representations for invariant feature extraction. In *Computational Color Imaging*, pp. 91–100, 2009.

Artem Babenko and Victor Lempitsky. Aggregating deep convolutional features for image retrieval. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1269–1277, 2015.

Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, Fabio Carrara, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, and Rita Cucchiara. Talking to DINO: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22025–22035, 2025.

Nicolas Le Bihan and Stephen J. Sangwine. Quaternion principal component analysis of color images. In *Proceedings of the International Conference on Image Processing (ICIP)*, pp. 809–812, 2003.

Wei Chen, Yu Liu, Weiping Wang, Erwin M. Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and S. Lew Michael. Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7270–7292, 2023a.

Xin Chen, Bin Wang, and Yongsheng Gao. Symmetric binary tree based co-occurrence texture pattern mining for fine-grained plant leaf image retrieval. *Pattern Recognition*, 129(108769), 2022.

Xin Chen, Bin Wang, and Yongsheng Gao. Fan-Beam Binarization Difference Projection (FB-BDP): A Novel Local Object Descriptor for Fine-Grained Leaf Image Retrieval. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11102–11111, 2023b.

Xin Chen, Bin Wang, Jinzheng Jiang, Kunkun Zhang, and Yongsheng Gao. SDePR: Fine-Grained Leaf Image Retrieval with Structural Deep Patch Representation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 2497–2505, 2024.

Todd A. Ell. Quaternion fourier transform: Re-tooling image and signal processing analysis. In *Quaternion and Clifford Fourier Transforms and Wavelets*, pp. 3–14. Birkhäuser / Springer, 2013. doi: 10.1007/978-3-0348-0603-9_1.

Todd A. Ell, Stephen J. Sangwine, and Simon J. Moxey. *Quaternion Fourier Transforms for Signal and Image Processing*. Wiley, 2014. doi: 10.1002/9781118930908.

J.I. Forcen, Miguel Pagol, Edurne Barrenechea, and Humberto Bustince. Co-occurrence of deep convolutional features for image search. *Image and Vision Computing*, 97(103909), 2020.

Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *European Conference on Computer Vision (ECCV)*, pp. 392–407, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, Oriane Siméoni, Huy V. Vo, Patrick Labatut, and Piotr Bojanowski. Dinov2 meets text: A unified framework for image- and pixel-level vision-language alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24905–24916, 2025.

Yannis Kalantidis, Clayton Mellina, and Simon Osindero. Cross-dimensional weighting for aggregated deep convolutional features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 685–701, 2016.

Nicholas Kibet Korir, Jian Han, Lingfei Shangguan, Chen Wang, Emrul Kayesh, Yanyi Zhang, and Jinggui Fang. Plant variety and cultivar identification: advances and prospects. *Critical Reviews in Biotechnology*, 33(2):111–125, 2013.

Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. Correlation verification for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5374–5384, 2022.

Seongwon Lee, Suhyeon Lee, Hongje Seong, and Euntai Kim. Revisiting self-similarity: Structural embedding for image retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23412–23421, 2023.

Heng Li, Zhiwen Liu, Yali Huang, and Yonggang Shi. Quaternion generic fourier descriptor for color object recognition. *Pattern Recognition*, 48(12):3895–3903, 2015.

Haibin Ling and David W. Jacobs. Shape classification using the inner-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):286–299, 2007.

Xueqin Liu, Jianbo He, Yufeng Wang, Guangnan Xing, Yan Li, Shouping Yang, Tuanjie Zhao, and Junyi Gai. Geographic differentiation and phylogeographic relationships among world soybean populations. *The Crop Journal*, 8(2):260–272, 2020.

Zhe Ma, Jianfeng Dong, Shouling Ji, Zhenguang Liu, Xuhong Zhang, Zonghui Wang, Sifeng He, Feng Qian, Xiaobo Zhang, and Lei Yang. Let all be whitened: Multi-teacher distillation for efficient visual retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4126–4135, 2024.

Shanmin Pang, Jin Ma, Jianru Xue, Jihua Zhu, and Vicente Ordonez. Deep feature aggregation and image re-ranking with heat diffusion for image retrieval. *IEEE Transactions on Multimedia*, 21 (6):1513–1523, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, pp. 157–166, 2014.

Stephen J. Sangwine. Fourier transforms of colour images using quaternion or hypercomplex numbers. *Electronics Letters*, 32(21):1979–1980, 1996.

Shihao Shao, Kaifeng Chen, Arjun Karpur, Qinghua Cui, André Araujo, and Cao Bingyi. Global features are all you need for image retrieval and reranking. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11036–11046, 2023.

Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of CNN activations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 1–12, 2016.

Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 157–166, 2014.

Bin Wang and Yongsheng Gao. Hierarchical string cuts: a translation, rotation, scale and mirror invariant descriptor for fast shape retrieval. *IEEE Transactions on Image Processing*, 23(9):4101–4111, 2014.

Bin Wang, Yongsheng Gao, Changming Sun, Michael Blumenstein, and John La Salle. Chord bunch walks for recognizing naturally self-overlapped and compound leaves. *IEEE Transactions on Image Processing*, 28(12):5963–5976, 2019.

Bin Wang, Yongsheng Gao, Xiaohui Yuan, Shengwu Xiong, and Xianzhong Feng. From species to cultivar: Soybean cultivar recognition using joint leaf image patterns by multiscale sliding chord matching. *Biosystems Engineering*, 194:99–111, 2020.

Changwei Wang, Shunpeng Chen, Yukun Song, Rongtao Xu, Zherui Zhang, Jiguang Zhang, Haoran Yang, Yu Zhang, Kexue Fu, Shide Du, Zhiwei Xu, Longxiang Gao, Li Guo, and Shibiao Xu. Focus on local: Finding reliable discriminative regions for visual place recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7536–7544, 2025.

Shijie Wang, Jianlong Chang, Zhihui Wang, Haojie Li, Wanli Ouyang, and Qi Tian. Content-aware rectified activation for zero-shot fine-grained image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4366–4380, 2024.

Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing*, 26(6): 2868–2881, 2017.

Min Yang, Dongliang He, Miao Fan, Baorong Shi, Xuetong Xue, Fu Li, Errui Ding, and Jizhou Huang. Dolg: Single-stage image retrieval with deep orthogonal fusion of local global features. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11772–11781, 2021.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6848–6856, 2018.

Cong Zhao, Sharon S.F. Chan, Wai-Kuen Cham, and L.M. Chu. Plant identification using leaf shapes—a pattern counting approach. *Pattern Recognition*, 48(10):3203–3215, 2015.

A APPENDIX

This appendix provides a complete derivation of contour starting-point invariance, backbone-specific configurations, controlled ablation definitions, robustness analysis, and computational analyses. Figures, tables, and equations use independent appendix numbering.

A.1 COMPLETE DERIVATION OF CONTOUR STARTING-POINT INVARIANCE

This section provides a complete derivation of the contour starting-point invariance used in Sec. 3.3. We first model a change of contour starting point as a circular shift, then derive the corresponding shift property of the left-sided DQFT, and finally show that taking quaternion magnitudes removes the induced phase factor.

A.1.1 CONTOUR SHIFT MODEL

For a fixed layer ℓ, scale k, branch $b \in \{q, s\}$, and feature channel, let

$$z[n] \in \mathbb{H}, \qquad n = 0, \dots, N-1, \tag{A1}$$

denote the quaternion descriptor sequence arranged in counterclockwise contour order.

Choosing another valid contour starting point changes only the sequence index. A circular shift by $\tau \in \{0, \dots, N-1\}$ is therefore defined as

$$z_\tau[n] = z[(n-\tau) \bmod N]\,. \tag{A2}$$

The contour direction is fixed to counterclockwise order throughout QSSA. Hence, the starting-point ambiguity considered here corresponds to a circular shift rather than a reversal of the sequence.

A.1.2 SHIFT THEOREM FOR THE LEFT-SIDED DQFT

Let $\mu \in \mathbb{H}$ be a fixed unit pure quaternion satisfying

$$\mu^2 = -1. \tag{A3}$$

The left-sided discrete quaternion Fourier transform (DQFT) of $z[n]$ is

$$\widehat{z}[u] = \sum_{n=0}^{N-1} e^{-\mu 2\pi u n/N}\, z[n], \qquad u = 0, \dots, N-1. \tag{A4}$$

Applying the same transform to the shifted sequence $z_\tau[n]$ gives

$$\widehat{z}_\tau[u] = \sum_{n=0}^{N-1} e^{-\mu 2\pi u n/N} z[(n-\tau) \bmod N] \tag{A5}$$

$$= \sum_{m=0}^{N-1} e^{-\mu 2\pi u (m+\tau)/N} z[m] \tag{A6}$$

$$= e^{-\mu 2\pi u \tau/N} \sum_{m=0}^{N-1} e^{-\mu 2\pi u m/N} z[m] \tag{A7}$$

$$= p_{u,\tau}\, \widehat{z}[u], \tag{A8}$$

where

$$p_{u,\tau} = e^{-\mu 2\pi u \tau/N}. \tag{A9}$$

To justify the circular reindexing, let

$$m = (n-\tau) \bmod N.$$

Then there exists $r \in \mathbb{Z}$ such that $n = m + \tau + rN$. Since $u, r \in \mathbb{Z}$,

$$e^{-\mu 2\pi u r} = \cos(2\pi u r) - \mu \sin(2\pi u r) = 1. \tag{A10}$$

Therefore,

$$e^{-\mu 2\pi un/N} = e^{-\mu 2\pi u(m+\tau)/N}. \tag{A11}$$

Moreover, all exponential factors belong to the commutative subalgebra

$$\mathbb{C}_\mu = \{a + b\mu \mid a, b \in \mathbb{R}\}, \tag{A12}$$

so that

$$e^{-\mu 2\pi u(m+\tau)/N} = e^{-\mu 2\pi u\tau/N} e^{-\mu 2\pi um/N}. \tag{A13}$$

Thus, the common phase factor is extracted from the *left* of each summand in Eq. A8, without being commuted through $z[m]$. This is consistent with the left-sided DQFT convention.

A.1.3 Quaternion-Magnitude Invariance

For any quaternions $a, b \in \mathbb{H}$, the quaternion norm is multiplicative:

$$|ab| = |a|\,|b|. \tag{A14}$$

Because $p_{u,\tau}$ in Eq. A9 is a unit quaternion,

$$|p_{u,\tau}| = 1. \tag{A15}$$

Using Eq. A8, we therefore obtain

$$\begin{aligned} |\widehat{z}_\tau[u]| &= |p_{u,\tau}\widehat{z}[u]| && \text{(A16)}\\ &= |p_{u,\tau}|\,|\widehat{z}[u]| && \text{(A17)}\\ &= |\widehat{z}[u]|. && \text{(A18)} \end{aligned}$$

For a quaternion-valued channel vector

$$\mathbf{h} = [h_1, \ldots, h_C] \in \mathbb{H}^C, \tag{A19}$$

define the channel-wise quaternion-magnitude operator as

$$\mathrm{qmag}(\mathbf{h}) = [|h_1|, \ldots, |h_C|] \in \mathbb{R}_+^C. \tag{A20}$$

Applying Eq. A18 independently to all C feature channels yields

$$\mathrm{qmag}\Big(\widehat{\mathbf{Z}}_{k,\tau}^{(\ell,b)}(u,:)\Big) = \mathrm{qmag}\Big(\widehat{\mathbf{Z}}_{k}^{(\ell,b)}(u,:)\Big). \tag{A21}$$

Therefore, changing the contour starting point does not affect the channel-wise quaternion magnitude spectrum. Concatenating these magnitude vectors across retained frequencies, scales, and quaternion branches preserves this equality for each layer descriptor. Subsequent ℓ_2 normalization also preserves the equality. Consequently, the layer-wise ℓ_1 dissimilarities remain unchanged under any valid choice of contour starting point.

Since the layer-wise distances are fused deterministically using KNN-HDFF, the final retrieval ranking is invariant to the contour starting point up to numerical floating-point precision.

A.2 Frozen-Backbone Configurations

All evaluated visual backbones are externally pretrained and remain frozen. To complement the backbone comparison reported in the main manuscript, Table A1 provides the feature-tensor selections and channel-reduction settings used in the corresponding experiments. The default VGG-16 configuration is included for completeness. Apart from the backbone-specific feature tensors and channel-reduction factors, all QSSA construction and retrieval settings remain unchanged.

Table A1: Feature-tensor selections and channel-reduction configurations for the evaluated frozen visual backbones. Output channels are listed separately for the two selected feature tensors.

Backbone	Selected feature tensors	Native channels	Channel reduction	Output channels
VGG-16	`relu4_3`, `relu5_3`	512, 512	Average every four adjacent channels in both tensors	128, 128
ResNet-50	`activation_40_relu`, `activation_49_relu`	1024, 2048	Average every eight and sixteen adjacent channels, respectively	128, 128
DenseNet-201	`pool4_pool`, `bn`	896, 1920	Average every eight and sixteen adjacent channels, respectively	112, 120
ShuffleNet	`node_153`, `node_199`	272, 544	Average every two and four adjacent channels, respectively	136, 136
CLIP-RN50	`relu4_3`, `relu5_3`	1024, 2048	Average every eight and sixteen adjacent channels, respectively	128, 128

A.3 Controlled Ablation Definitions and Analysis

All component ablations are conducted on PeanCultivar120 using $N = 64$, $K_s = 4$, and a frozen VGG-16 backbone. Transform-based variants use $K_f = 4$. To ensure controlled comparisons, the boundary patches, selected feature layers, channel-reduction strategy, normalization, distance computation, and layer-wise fusion are kept unchanged whenever applicable.

Table A2 shows that preserving regional spatial layout substantially improves retrieval performance over global pooling. Global 1-ary pooling obtains 58.40%, while the left–right and top–bottom two-region variants reach 62.10% and 63.33%, respectively. Coupling four spatial components in the quaternion domain further improves the score to 66.70%. Compared with Real-4ary, Quat.-4ary provides a gain of 3.59 percentage points, indicating that joint quaternion-domain processing is more effective than treating the four components independently.

The two quaternion branches also encode complementary structural information. The quadrant-layout branch and the symmetry-discrepancy branch achieve 66.12% and 65.10% when used separately, whereas combining their independently normalized spectral descriptors yields 66.70%. This improvement suggests that absolute regional responses and symmetry-aware discrepancies capture different aspects of fine-grained leaf structure.

The choice of contour aggregation has a substantial effect on performance. DQFT achieves 66.70%, exceeding Q-Avg and Q-Max by 4.17 and 4.51 percentage points, respectively. It also considerably outperforms the component-wise DCT under the same retained-frequency budget. These results support the use of quaternion-domain spectral aggregation for preserving the ordered and coupled structure of the closed contour sequence.

A.4 Robustness and Efficiency Analysis

A.4.1 Robustness to Contour Localization Noise

To simulate contour localization errors, we apply additive white Gaussian noise to the extracted contour-coordinate matrix using MATLAB's `awgn` function with the `'measured'` option. We evaluate SNRs of 40, 35, and 30 dB.

For each SNR level, the complete retrieval evaluation is repeated three times using independently generated contour perturbations, and the mean Bulls-eye score over the three trials is reported.

The perturbed contours are then resampled into $N = 64$ points, with local patch orientations recomputed accordingly. All subsequent feature extraction, descriptor construction, and retrieval settings

Table A2: Controlled definitions and Bulls-eye scores (%) of the component ablations on PeanCultivar120. All variants use $N = 64$, $K_s = 4$, and a frozen VGG-16 backbone. Transform-based variants use $K_f = 4$. All other applicable settings are kept unchanged across variants.

Component	Variant	Controlled modification	Score (%)
Spatial encoding	1-ary	Apply global average pooling to each channel map without spatial partitioning.	58.40
	H-2ary	Partition each channel map into left and right regions and retain their pooled responses separately.	62.10
	V-2ary	Partition each channel map into top and bottom regions and retain their pooled responses separately.	63.33
	Real-4ary	Replace each quaternion sequence with four real-valued component sequences, apply a real-valued DFT to each component independently, and use the same magnitude-based low-frequency construction.	63.11
	Quat.-4ary	Couple the four components as a quaternion and aggregate the contour sequence using the left-sided DQFT.	**66.70**
Quaternion branches	q-branch only	Retain only the quadrant-layout quaternion sequence.	66.12
	s-branch only	Retain only the symmetry-discrepancy quaternion sequence.	65.10
	$q + s$ branches	Concatenate the independently normalized spectral descriptors from both quaternion branches.	**66.70**
Contour aggregation	Q-Avg	Replace the DQFT with quaternion average pooling over the contour sequence.	62.53
	Q-Max	For each channel, select the contour sample with the largest quaternion magnitude and use the corresponding quaternion as the pooled representation.	62.19
	DCT	Apply a component-wise real DCT to the four quaternion components while retaining the same low-frequency budget.	47.56
	DQFT	Apply the proposed left-sided DQFT followed by channel-wise quaternion magnitude.	**66.70**

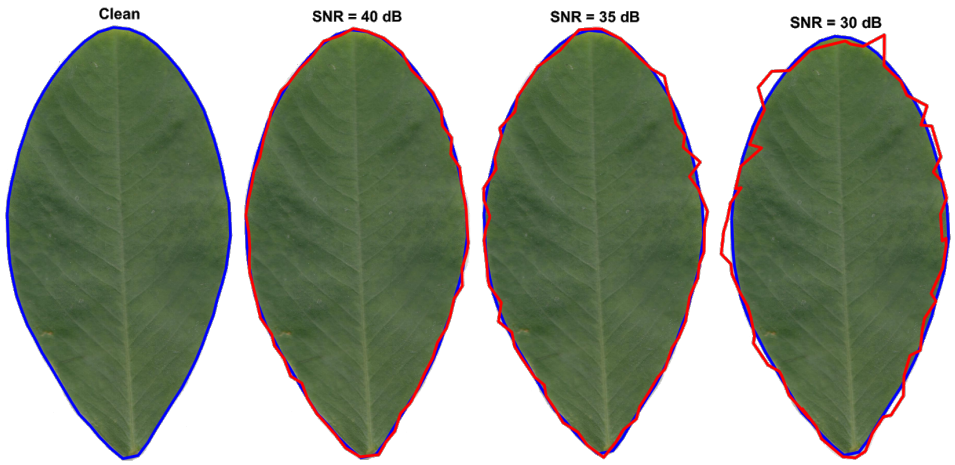


Figure A1: Representative visualizations of contour localization perturbations on PeanCultivar120. From left to right: the clean contour and representative perturbations at SNRs of 40, 35, and 30 dB. Blue curves denote the clean contour, while red curves denote the perturbed contours.

remain unchanged. Figure A1 shows representative contour perturbations, while Table A3 reports the corresponding mean retrieval performance.

As shown in Table A3, the mean retrieval performance decreases gradually as the noise level increases. Even at 30 dB, QSSA achieves an average Bulls-eye score of 65.20%, only 1.50 percentage points below the clean-contour result.

Table A3: Robustness to contour localization noise on PeanCultivar120. For each noisy condition, the score is averaged over three retrieval trials with independently generated contour perturbations; the clean score is obtained without perturbation.

Contour condition	Clean	40 dB	35 dB	30 dB
Bulls-eye score (%)	66.70	66.57	65.92	65.20

A.4.2 Descriptor Size and Storage

For each selected feature layer, QSSA concatenates the channel-magnitude vectors over K_s scales, two quaternion branches, and K_f retained frequencies. The resulting descriptor dimension is

$$D_{\text{layer}} = 2K_sK_fC. \tag{A22}$$

With the default $K_s = 4$ and $C = 128$, this reduces to $D_{\text{layer}} = 1024K_f$.

Table A4 summarizes the resulting descriptor size and storage for different K_f. QSSA uses two VGG-16 feature layers whose descriptors are stored separately, with their ℓ_1 distances subsequently fused using KNN-HDFF. Under the default $K_f = 4$, each layer contains 4,096 values, yielding 8,192 float32 values, or 32 KiB per image, across the two layers.

Table A4: Descriptor size and storage for different K_f. Storage uses float32 values and includes both VGG-16 layers.

K_f	Dim./layer	Total values	Storage/image
1	1,024	2,048	8 KiB
2	2,048	4,096	16 KiB
4	4,096	8,192	32 KiB
8	8,192	16,384	64 KiB
16	16,384	32,768	128 KiB

A.4.3 Runtime and Computational Complexity

With $N = 64$ contour samples and $K_s = 4$ geometric scales, QSSA constructs $NK_s = 64 \times 4 = 256$ boundary patches per image. All patches are stacked into a single batch and processed jointly by the frozen CNN. Structural encoding and spectral aggregation are then applied to the extracted feature maps.

The current implementation computes the full N-frequency DQFT before retaining the first K_f coefficients. Direct evaluation has complexity $\mathcal{O}(N^2C)$ per scale, branch, and feature layer. Since $N = 64$, this cost is small relative to batched CNN feature extraction. Increasing K_f mainly affects descriptor storage and matching cost, because frequency truncation is performed after the DQFT.

On an Intel Core i9-13900HX CPU and an NVIDIA GeForce RTX 4060 GPU, the complete descriptor-extraction pipeline takes approximately 0.31 s per image on average, corresponding to about 3.23 images/s.

For a database of M images, computing the layer-wise ℓ_1 distances requires $\mathcal{O}(MD_{\text{total}})$ operations per query, where D_{total} is the total number of stored descriptor values across the selected feature layers. The resulting layer-wise distances are subsequently fused using KNN-HDFF. Therefore, reducing K_f decreases both descriptor storage and distance-computation cost linearly.