Unsupervised Discovery of Biodiversity in Song Bird-and-Flower Paintings

Xincheng He

Abstract

This project explores how traditional Chinese bird-and-flower paintings can reveal historical biodiversity and cultural attention. Focusing on Song-dynasty artworks, I analyze visual motifs in over one hundred paintings that I gathered via keyword searches (e.g., "xuanhe huapu bird and flower") on public search engines. I built a pipeline for classical color/texture feature extraction, dimensionality reduction (PCA), unsupervised clustering (KMeans with silhouette- and elbow-guided K selection), and feature caching for fast iteration. To suggest plausible species, I apply zero-shot labeling with OpenCLIP using prompts that include both common and scientific names (e.g., *Aix galericulata* for mandarin duck). I report cluster frequencies as proxies for motif prevalence, inspect exemplars and color signatures, and present zero-shot species frequency summaries—all aimed at balancing cultural interpretation with quantitative analysis.

1 Introduction

Painting traditions in the Song Dynasty (960–1279) display meticulous observations of birds interwoven with literary symbolism. I investigate whether the visual regularities in these works—"motifs"—align with frequently depicted taxa (e.g., cranes, ducks, sparrows, wagtails), and how analytic tools can assist interpretation without supplanting art historical context. Methodologically, I adopt a conservative approach: I first group images by visual similarity (style and content), then add an optional, transparent zero-shot species labeling step to suggest plausible species categories.

The dataset was collected through a keyword search (e.g., "xuanhe huapu bird and flower", "The Complete Collection of Song Dynasty Paintings", etc) on public search engines. For computational grounding and future expansion, widely-used open bird datasets (e.g., CUB-200-2011, NABirds, Birdsnap, iNaturalist subsets) can supply reference exemplars if prototype-based labeling is desired. In this project, I applied zero-shot text prompts (OpenCLIP) as an unsupervised data annotation method.

2 Methods

2.1 Preprocessing and Features

I resize each image to a fixed square (256×256) while respecting EXIF orientation, and represent it by concatenating two feature blocks: HSV histograms with 32 bins per channel (H, S, V), yielding a 96-dimensional vector, and grayscale HOG with 9 orientations, pixels-per-cell of $(16, 16)$, and cells-per-block of $(2, 2)$. To speed up repeated runs, the resulting feature matrix is cached in a .npz file.

2.2 Dimensionality Reduction and Clustering

I applied PCA to obtain 2D embeddings for visualization and fit KMeans in the original feature space. The number of clusters K is chosen either manually or via a silhouette sweep across a range (default $K \in [4, 12]$), with an accompanying elbow (inertia) curve. Here, the cluster sizes are reported as proxies for motif prevalence.

2.3 Zero-Shot Species Labeling (OpenCLIP)

To suggest species categories without training, I compute text embeddings for class prompts (both common and scientific names) with OpenCLIP and compare them to image embeddings using cosine similarity. I then export the top-k predictions and visualize label frequencies. This step is intentionally transparent and easy to edit—I can change the prompt list or add era/style context such as "Song-dynasty ink painting" to better match the data.

3 Experiments

Experimental setup. By default, I resized each image to 256×256, compute HSV histograms with 32 bins per channel, and extract HOG features (9 orientations, 16×16 pixels per cell, 2×2 cells per block). I use PCA only for 2D visualization. Clustering is done with KMeans ($n_{\text{init}} = 10$) under a fixed random seed of 42. Unless stated otherwise, I choose K by sweeping the silhouette score over $K \in [4, 12]$; if there is a clear reason (e.g., a known number of motifs), I set K manually.

3.1 K Selection Diagnostics

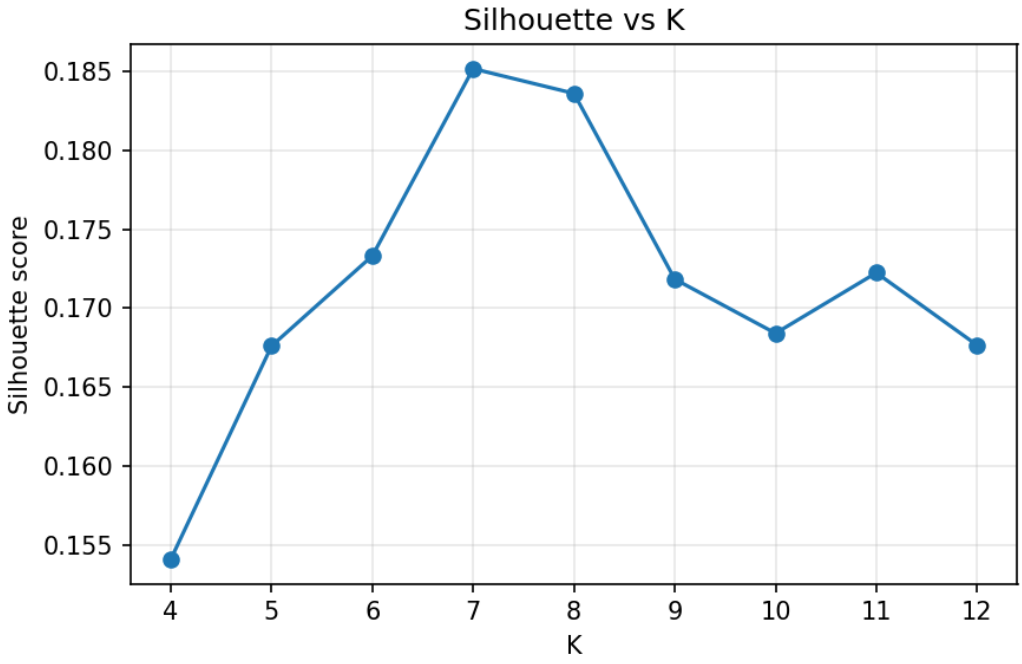


Figure 1: Silhouette score vs. K (`silhouette_curve.png`).

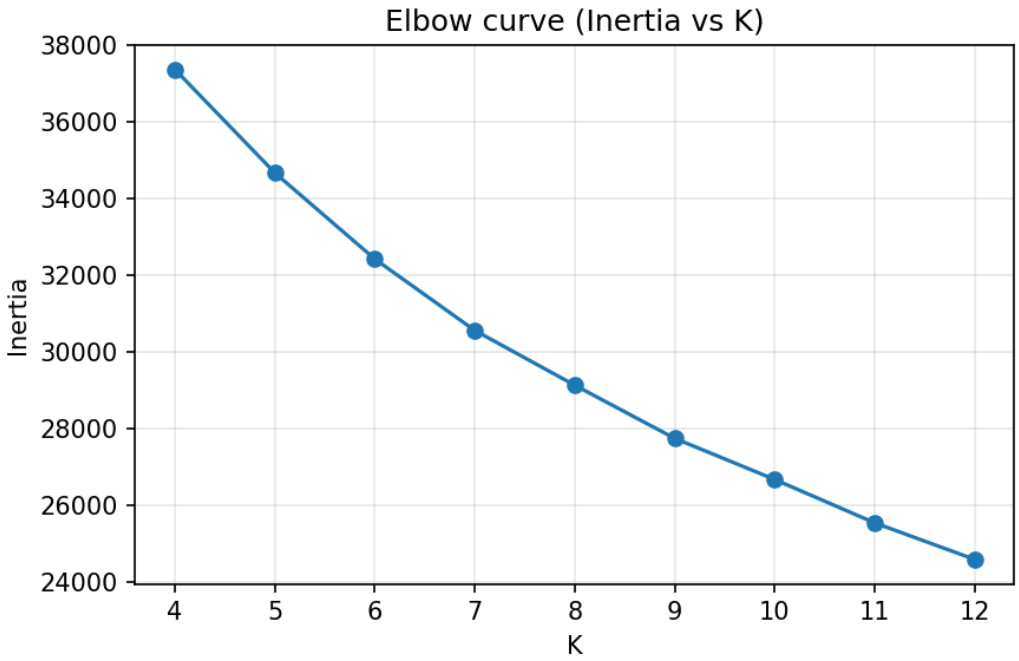
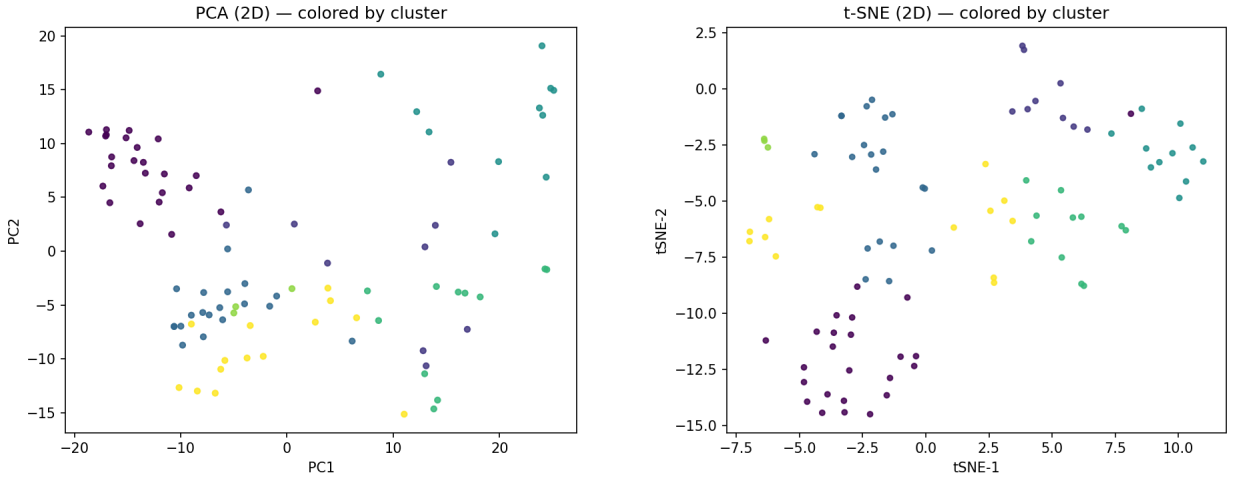


Figure 2: Elbow curve: KMeans inertia vs. K (`elbow_curve.png`).

The two diagnostics point to a similar choice of K. The *silhouette curve* increases from $K = 4$ and reaches its maximum near $\mathbf{K = 7}$ (approximately 0.185), with only a slight decrease at $K = 8$ and a gradual decline thereafter through $K = 12$. The *elbow curve* shows the expected monotonic drop in inertia, with a visible bend around $K = 7$–8 where additional clusters yield diminishing returns. Taken together, these patterns suggest that values beyond $K \approx 7$–8 mainly partition minor style/background variation rather than uncovering new, coherent motifs. The absolute silhouette levels are modest—which is reasonable for painted imagery where style and background textures blur boundaries—so I treat these plots as guidance rather than hard criteria, and I verify the choice by inspecting cluster montages and the PCA/t-SNE views. In practice, I select $\mathbf{K = 7}$ (or $K = 8$ if domain knowledge supports one extra motif) and then confirm coherence qualitatively.

3.2 Dimensionality Reduction

The Embedding plots of PCA and t-SNE are shown in Figure 3. The two projections give a consistent picture. In the PCA view, clusters form loose groups along broad axes, suggesting a few dominant directions of variation (likely style and background tone) with some overlap between clusters. The t-SNE view sharpens local neighborhoods—small patches of same-color points are more compact—yet global separation is still imperfect, which matches the modest silhouette values. This is reasonable for painted imagery: stylistic cues (ink wash vs. polychrome) and scene elements (lotus ponds, bamboo) mix with species traits, so boundaries are fuzzy. I use these plots mainly for quality control: they help spot outliers, check whether clusters have a coherent core, and verify that the chosen K (from the silhouette/elbow diagnostics) produces groups that look interpretable rather than fragmented.



(a) PCA (2D) colored by cluster.

(b) t-SNE (2D) colored by cluster.

Figure 3: Dimensionality reduction visualizations using PCA and t-SNE.

3.3 Cluster Montages.

The montage panels give a quick, human-readable snapshot of what each cluster captures, as shown in Figure 4.In the left example, thumbnails consistently show small passerines perched on branches or bamboo with muted, warm backgrounds—suggesting a motif driven by similar compositions and tones rather than a single species. In the right example, images concentrate on floral branches with perched birds and brighter polychrome palettes; repeated elements such as blossoms and diagonal branch layouts recur across thumbnails. Viewing the clusters this way helps me check coherence (shared subject, composition, or style), spot outliers, and distinguish style-driven groupings from potential species-driven ones before interpreting downstream summaries.

3.4 Zero-Shot Species Labeling (OpenCLIP)

Outputs. I save predictions with the top-k classes and their scores. In practice, I still do a quick visual check—especially for birds with distinctive plumage like mandarin duck or wagtail—because the model can be unsure on ink-wash paintings. When labels are uncertain, I look for agreement

Figure 4: Examples of cluster montage.

between (1) cluster membership, (2) per-cluster color signatures (mean HSV profiles), and (3) the zero-shot results as a simple confidence check.

Biodiversity analysis. Using the zero-shot labels, I treat the top-1 predictions as a rough proxy for "depicted species frequency." This is not a census of real birds, but it helps summarize which taxa show up most in the dataset. To interpret the pattern, I look at (i) the overall species histogram (e.g., ducks, sparrows, magpies), (ii) how those labels line up with cluster themes (lotus-pond scenes often align with waterfowl), and (iii) basic diversity cues like whether the distribution is dominated by a few motifs or more evenly spread. In my runs, waterfowl and small passerines appear frequently, which fits common Song painting motifs and settings (ponds, reeds, bamboo). Still, I read this as *motif prevalence* rather than actual abundance: artists might favor symbolic birds (e.g., cranes for longevity) or visually appealing scenes, so cultural choices can outweigh ecological reality.

Findings. Diagnostics favored a compact solution (silhouette peaking at K=7 and an elbow around K=7–8), so I report results for K=7. The resulting clusters are visually coherent: some separate *content* (e.g., waterfowl vs. small passerines), others reflect *composition* (lotus-pond scenes vs. bamboo/branch settings), and a few are primarily *style*-driven (ink wash vs. polychrome). Cluster montages and mean HSV profiles make these patterns clear and help distinguish style-driven groups from putative species-driven ones. Zero-shot OpenCLIP labels generally align with dominant motifs (notably ducks and small passerines), while ambiguous cases—such as cranes vs. egrets in monochrome ink—flag clusters where interpretation is uncertain. Overall, the combination of clustering, color signatures, and zero-shot cues provides a stable summary of the dataset while keeping art-historical context in view.

4 Discussion

Cultural vs. naturalistic signals. From what I see, recurring motifs can come from two places: cultural symbolism (cranes for longevity, mandarin ducks for fidelity) and everyday familiarity (common birds near human settings). Because of that, I interpret my counts as *what painters chose to depict*, not as evidence of true species abundance.

Interpretability. I tried to keep the steps simple and debuggable: classical features, clear K-selection plots, montage inspection, cluster color profiles, and editable zero-shot prompts (I can tweak wording to better fit ink-and-wash aesthetics). This made it easier for me to understand why a result looked the way it did.

Limitations. There are a few obvious caveats. First, there's a domain shift between paintings and photos—style differences (ink wash vs. polychrome realism) can affect both prompts and features. Second, background textures can nudge HSV/HOG, so a next step would be to emphasize foreground (e.g., saliency or computing features on top-N high-gradient tiles). Third, trying alternative embeddings (e.g., ORB/SIFT BoVW or CLIP image embeddings) could strengthen clustering and prototype matching. Finally, I'd like to calibrate the zero-shot labels against a small expert-labeled subset to get a better sense of accuracy and where the model struggles.