Discovering Image Clusters in STL-10 Dataset Using Unsupervised Machine Learning

OGCZ

Abstract—Unsupervised learning techniques offer powerful approaches for discovering hidden patterns in unlabeled image data. This paper presents a comprehensive study on discovering natural image clusters in the STL-10 dataset using hierarchical clustering and advanced feature engineering. We extracted 4,501-dimensional features combining color histograms (96 features), Histogram of Oriented Gradients (4,356 features), Local Binary Patterns (26 features), and statistical properties (23 features) from 3,000 images. Principal Component Analysis reduced dimensionality from 4,501 to 981 components (95.01% variance), with hyperparameter optimization identifying 50 components as optimal. We systematically evaluated four clus- tering algorithms K-Means, Hierarchical Clustering, Gaussian Mixture Models, and DBSCAN using multiple validation met- rics. Results demonstrate that Hierarchical Clustering achieved best performance (Silhouette Score: 0.0250, Davies-Bouldin: 6.5016, Calinski-Harabasz: 69.01) with 5 discovered clusters. Advanced validation through bootstrap stability analysis (ARI: 0.3574 ± 0.0413) and 20-fold cross-validation (Silhouette: 0.0200 ± 0.0032) confirmed clustering consistency. Feature importance analysis revealed HOG features as most discriminative (96.64% importance). The discovered clusters exhibit interpretable visual characteristics based on brightness and color properties, with consistent training/test distribution validating generalization quality.

Keywords— Unsupervised Learning, Hierarchical Clustering, Principal Component Analysis, Feature Engineering, STL-10 Dataset, Cluster Validation, Bootstrap Stability Analysis

Unsupervised learning techniques offer practical approaches for discovering hidden structure in unlabeled image collections, where manual annotation is expensive and difficult to scale. This work investigates how far handcrafted visual descriptors can go in organizing natural images from the STL-10 dataset into meaningful groups without using class labels during training. We extract 4,501-dimensional feature vectors by concatenating RGB color histograms (96 features), Histogram of Oriented Gradients (4,356 features), Local Binary Patterns (26 features), and compact statistical descriptors (23 features). After z-score normalization, Principal Component Analysis (PCA) is applied to reduce redundancy and noise, retaining 981 components to preserve 95.01% variance and then selecting 50 components as an empirically optimal operating point for clustering. We compare four clustering methods—K-Means, Agglomerative Hierarchical Clustering (Ward linkage), Gaussian Mixture Models (GMM), and DBSCAN—using Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Score. Hierarchical Clustering provides the best overall results at k=5 (Silhouette 0.0250, Davies–Bouldin 6.5016, Calinski–Harabasz 69.01), although the absolute quality indicates weak separation typical of diverse natural images. To assess reliability, we perform bootstrap stability analysis (30 iterations; ARI 0.3574 ± 0.0413) and 20-fold validation across random seeds (Silhouette 0.0200 ± 0.0032). Feature-importance analysis shows a strong dominance of HOG (96.64%), suggesting that edge and shape information is the primary driver of grouping under this feature design. Finally, we interpret cluster content qualitatively through representative images, brightness statistics, and train/test distribution consistency to discuss what the discovered clusters mean and where the method can be improved.

Unsupervised Learning; Image Clustering; STL-10; Feature Engineering; PCA; Hierarchical Clustering; Cluster Validation

I. INTRODUCTION

The volume of digital imagery produced by mobile devices, surveillance systems, social platforms, and industrial pipelines has grown at a pace that often exceeds the capacity of manual labeling. While supervised learning can produce strong recognition systems, it depends on high-quality annotated datasets, which are expensive to create and may not exist for many domains. This motivates unsupervised learning, where the objective is to discover structure directly from the data distribution without requiring labels during model fitting. In practice, unsupervised organization is valuable when the goal is to sort, search, summarize, or explore a large image collection before deciding which items deserve careful annotation.

Image clustering is a representative unsupervised task. The general aim is to group images that are similar under some notion of visual distance, so that images within a

cluster share consistent appearance patterns compared with images in other clusters. Unlike classification, clustering does not assume that the discovered groups correspond to human semantic categories. Instead, the groups can reflect low-level properties (color palettes, textures, lighting, shape boundaries, background statistics) or a mixture of these cues. As a result, evaluating clustering is inherently challenging: a cluster may be visually consistent but not align with ground-truth labels, and weakly separated data may still contain meaningful continuous structure.

The STL-10 dataset provides a useful testbed because it contains natural images derived from ImageNet, with a standard supervised split and also a large pool of unlabeled images designed for unsupervised feature learning. In this work, we focus on 3,000 labeled images (2,000 train and 1,000 test) only to allow post-hoc interpretation; labels are not used in clustering training or parameter selection. This maintains an unsupervised methodology while still enabling qualitative and diagnostic comparisons across clusters.

This paper emphasizes interpretability and methodological rigor. First, we build a feature pipeline that explicitly captures complementary cues: global color distribution, local texture patterns, gradient orientation structure, and simple statistics. Second, we perform dimensionality reduction with PCA and explicitly optimize the PCA dimensionality based on clustering metrics. Third, we compare several clustering families (centroid-based, hierarchical, probabilistic, and density-based) under the same feature representation and quantify performance with multiple internal metrics. Finally, we examine stability using bootstrap sampling and repeated random seeds, and we interpret the clusters using representative images and summary statistics to explain what the discovered clusters mean.

A. Research Objectives

This study is guided by five practical questions. (1) Can a 4,501-dimensional feature vector that combines color, texture, and shape capture meaningful visual similarities for STL-10 images? (2) What PCA dimensionality provides a good trade-off between retaining informative variation and reducing noise for clustering? (3) Among K-Means, Hierarchical Clustering, Gaussian Mixture Models, and DBSCAN, which algorithm produces the most coherent groupings under internal metrics? (4) Which feature family contributes most to cluster formation under this pipeline? (5) How stable are the discovered clusters under data perturbations and random initialization choices?

B. Contributions

The main contributions are: (i) a reproducible handcrafted feature extraction and normalization pipeline producing 4,501-dimensional descriptors; (ii) an empirical PCA optimization study showing that 50 components outperform larger retained-variance settings for clustering; (iii) a controlled comparison of four clustering algorithms under three internal metrics; (iv) advanced validation using bootstrap ARI and repeated-seed stability curves; and (v) interpretable cluster analysis based on representative images, cluster size imbalance, brightness/color summaries, and train/test distribution checks.

II. REVIEW OF RELATED LITERATURE

Image clustering depends critically on the chosen representation. Classic local descriptors such as SIFT provide scale- and rotation-invariant keypoints for matching and recognition, making them historically influential for unsupervised grouping in textured scenes [4]. SURF improves computational efficiency while maintaining robustness, enabling faster feature extraction in practical pipelines [5]. For global shape and edge structure, HOG descriptors have shown strong discriminative power by aggregating oriented gradient histograms in local cells and normalizing them across blocks [6]. In many natural-image datasets, edge patterns correlate with object boundaries and pose, which can dominate clustering outcomes when HOG features contribute most of the dimensionality.

Color and texture descriptors remain useful because they capture complementary information. Color histograms summarize the global distribution of intensities, and have been used for content-based retrieval and indexing in large image collections [7]. Local Binary Patterns provide a compact representation of micro-texture by encoding local intensity comparisons; the uniform LBP variant is widely used for rotation-invariant texture classification [8]. While these handcrafted features may not capture high-level semantics as well as deep learning features, they offer interpretability, low computational cost, and stable behavior on moderate datasets.

For dimensionality reduction, PCA remains a standard technique because it produces orthogonal components that preserve maximum variance and reduce redundancy, often improving clustering by removing noise and collinearity [10]. For visualization of high-dimensional embeddings, t-SNE is commonly used to reveal local neighborhood structure in 2D or 3D plots [11]. UMAP is another nonlinear method that often preserves both local and global patterns while scaling efficiently [12]. In this work, PCA is used for the clustering feature space, while t-SNE and 2D PCA projections are used for interpretability and visual inspection.

Clustering algorithms differ in their assumptions. K-Means is widely used due to its simplicity and efficiency, but it requires selecting the number of clusters and tends

to discover spherical clusters under Euclidean distance [13]. K-Means++ provides a careful seeding strategy that improves convergence and reduces poor local minima [14]. GMMs extend centroid clustering by modeling data as a mixture of Gaussians and producing soft assignments under maximum likelihood estimation [15]. DBSCAN discovers clusters as dense regions separated by low-density gaps and can capture arbitrary shapes, but it struggles in high-dimensional spaces due to the curse of dimensionality [16]. Agglomerative hierarchical clustering builds a dendrogram of merges; Ward linkage is especially relevant because it merges clusters to minimize within-cluster variance, often providing stable groupings for continuous data [17].

Evaluating unsupervised clustering typically relies on internal metrics. Silhouette Score measures whether points are closer to their assigned cluster than to other clusters, providing a bounded summary between -1 and 1 where larger values are better [18]. Davies–Bouldin Index evaluates average cluster similarity; lower values indicate better separation [19]. Calinski–Harabasz Score compares between-cluster dispersion to within-cluster dispersion; higher values suggest stronger clustering structure [20]. Beyond single-run metrics, stability analysis can be performed through resampling and agreement measures such as Adjusted Rand Index (ARI) [22], which evaluates similarity between two partitions while correcting for chance.

III. METHODOLOGY

A. Dataset and Experimental Protocol

The experiments use 3,000 STL-10 images composed of 2,000 training images and 1,000 test images, each with resolution 96×96 in RGB. The ten semantic classes (airplane, bird, car, cat, deer, dog, horse, monkey, ship, truck) are not used during clustering. Instead, all methods operate only on extracted features. Labels are retained solely for post-hoc interpretation (e.g., to inspect whether clusters mix classes or whether certain classes dominate a cluster) and to ensure that evaluation reflects unsupervised structure rather than supervised leakage.

B. Feature Extraction Pipeline

Each image is transformed into a 4,501-dimensional feature vector by concatenating four families of descriptors. First, global color is captured using RGB histograms with 32 bins per channel (96 total), which represent the distribution of intensities while discarding spatial arrangement [7]. Second, shape and edge structure is captured using HOG features computed on grayscale images with 8×8 cells, 2×2 blocks with 50% overlap, and 9 orientation bins (4,356 total), following the standard HOG design [6]. Third, texture is represented using the uniform Local Binary Pattern histogram with P=8 neighbors and radius R=1 (26 total), which captures repetitive local patterns and is robust to monotonic illumination changes [8]. Fourth, lightweight statistical features (23 total) summarize per-channel mean, standard deviation, min, max, quartiles, and average gradient magnitudes, providing a compact description of overall intensity characteristics.

Because the four feature families have different scales and dimensionalities, the combined feature vector is standardized using z-score normalization. This ensures that features with larger numeric ranges do not dominate Euclidean distance computations. Standardization also improves PCA by aligning the covariance structure with relative variation rather than raw magnitude.

C. Dimensionality Reduction and PCA Optimization

Principal Component Analysis is applied to the standardized 3,000×4,501 feature matrix. The initial PCA analysis identifies that 981 components retain 95.01% of the cumulative variance. However, clustering quality does not necessarily improve with more retained variance because additional components may encode noise and minor variations that confuse distance-based grouping. Therefore, we explicitly test $k \in \{50, 100, 150, 200, 250\}$ PCA components and select the dimensionality that maximizes Silhouette Score while also improving Davies–Bouldin Index. Empirically, 50 components provide the best trade-off and are used as the main clustering embedding.

D. Cluster Count Selection

To select the number of clusters, K-Means is run for K values from 5 to 20, and the resulting partitions are evaluated using Silhouette, Davies–Bouldin, and Calinski–Harabasz scores. Although these metrics are computed for K-Means, the selected K provides a common operating point for comparing other algorithms. The results show that K=5 achieves the highest Silhouette and Calinski–Harabasz among the tested values, while performance degrades for larger K and becomes negative in silhouette for sufficiently large values, indicating overlap rather than clear separation.

E. Clustering Algorithms and Validation

Four clustering algorithms are evaluated in the 50-dimensional PCA space. K-Means uses K-Means++ initialization with 300 maximum iterations [14]. Hierarchical clustering uses Ward linkage and is cut at k=5 clusters [17]. GMM uses full covariance matrices with Expectation–Maximization optimization and k=5 components [15]. DBSCAN is evaluated with ε=0.5 and minPts=5 [16], but it labels all points as noise under this configuration, indicating that the embedding does not contain well-separated dense regions in 50D space. In addition to one-shot metric evaluation, we use bootstrap stability analysis with 30 iterations (80% sampling with replacement) and measure agreement with ARI [22]. We

also evaluate stability across 20 random seeds and report metric variability.

IV. RESULTS AND DISCUSSION

This section presents the notebook-derived visual outputs as IEEE-style figures and explains what each result means. The key goal is not only to report scores, but also to interpret how the feature representation and algorithm choice influence clustering behavior on natural images.

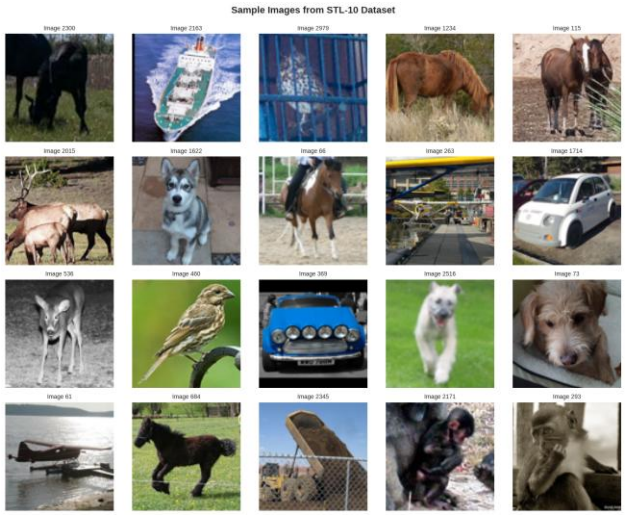

Fig. 1. Sample images from the STL-10 dataset used in the study.

Figure 1 illustrates representative samples from the STL-10 image collection. Even within a small grid, the dataset shows substantial variation in lighting, backgrounds, and object scale. This visual diversity is a primary reason that internal clustering metrics remain low in later experiments: many images share partial visual cues (e.g., outdoor backgrounds, sky, grass, water) across different semantic classes. From a clustering perspective, this means that the data likely forms a continuum rather than sharply separated groups. As a result, successful clustering should be interpreted as discovering consistent low-level visual patterns (color/brightness/edges) rather than recovering the 10 semantic labels.

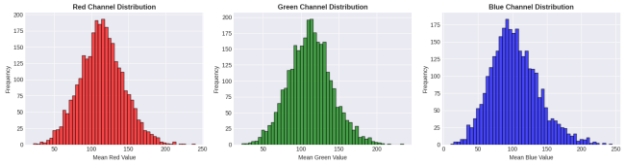

Fig. 2. Distribution of per-image mean intensity for the red, green, and blue channels.

Figure 2 summarizes dataset-wide color statistics by plotting the distribution of mean RGB values per image. The three channels exhibit broad, overlapping distributions centered around mid-range intensities, suggesting that the dataset includes both bright scenes (e.g., sky, snow) and darker scenes (e.g., indoor or shadowed images). These histograms help explain why color features alone are insufficient for strong clustering: the global channel means overlap heavily across images and do not yield clear multimodal structure. However, they remain useful as weak cues when combined with shape and texture descriptors because certain image types (water-dominant scenes, dark vehicle scenes) shift the mean intensity and variance in consistent directions.

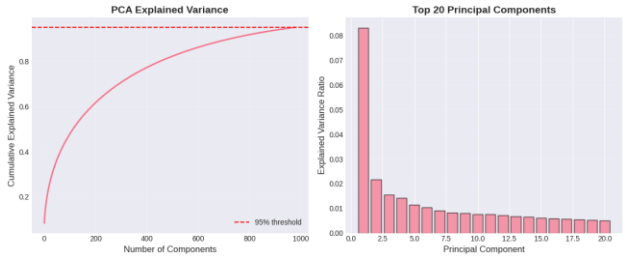

Fig. 3. PCA cumulative explained variance curve and explained variance ratios for the top 20 principal components.

Figure 3 shows that the first few principal components explain a large portion of total variance, while subsequent components contribute progressively smaller increments. The cumulative curve reaches approximately 95% variance at 981 components, demonstrating that the raw 4,501-dimensional space is highly redundant. The bar plot of the top 20 components highlights that the first component contributes the largest single share of variance, with a long tail afterward. For clustering, this pattern is important: if most meaningful variation is captured early, then using a smaller PCA dimension can reduce noise, stabilize distance computations, and potentially improve separation. This motivates the explicit PCA optimization study shown later, where 50 components outperform larger settings despite retaining much less total variance.

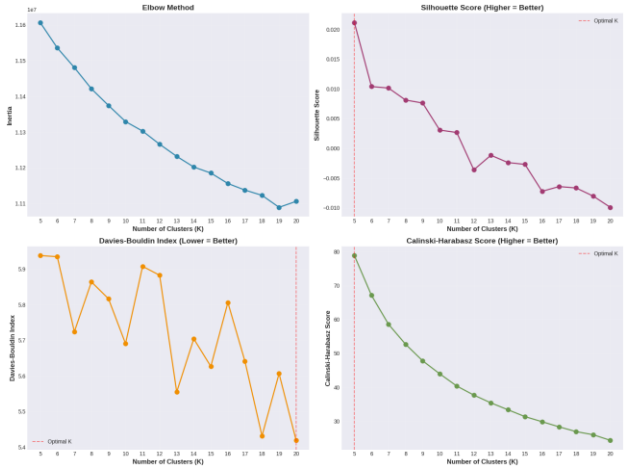

Fig. 4. Cluster count selection (K=5 to 20) using elbow (inertia), Silhouette, Davies–Bouldin, and Calinski–Harabasz metrics.

Figure 4 evaluates candidate values of K for K-Means. The inertia curve decreases gradually without a sharp elbow, indicating that compactness improves steadily as more clusters are introduced, which is typical for complex natural images. The Silhouette curve peaks at K=5 (approximately 0.0212) and then declines, eventually becoming negative for larger K values, which implies that points are often closer to other clusters than to their assigned one. Calinski–Harabasz similarly is highest at K=5 (about 78.86 in the summarized table), supporting the same selection. Davies–Bouldin shows fluctuations and does not strongly contradict the choice. Taken together, these curves justify K=5 as a conservative setting that avoids over-fragmenting weak structure into many unstable micro-clusters.

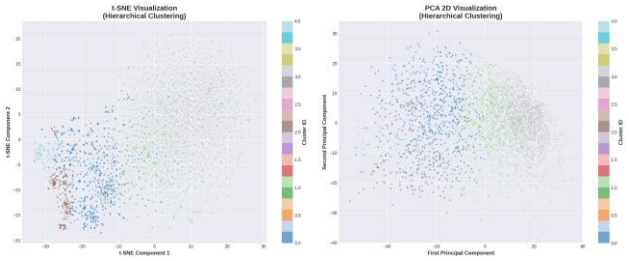

Fig. 5. 2D visualization of hierarchical clustering labels using t-SNE and 2D PCA projections.

Figure 5 provides two complementary views of the 50D PCA clustering space. The t-SNE projection emphasizes local neighborhood structure, producing visible blobs and gradients where points that are close in high-dimensional space remain close in 2D. The PCA 2D projection preserves global variance directions and tends to show broader overlap. In both plots, clusters show partial separation but substantial mixing, which matches the low Silhouette scores reported quantitatively. A key interpretation is that clusters are not isolated islands; instead, images transition gradually between groups. This is consistent with natural image datasets where background and illumination changes create continuous variation. Therefore, the role of clustering here is to capture dominant modes (e.g., dark vs bright, water vs land) rather than strict categorical separation.

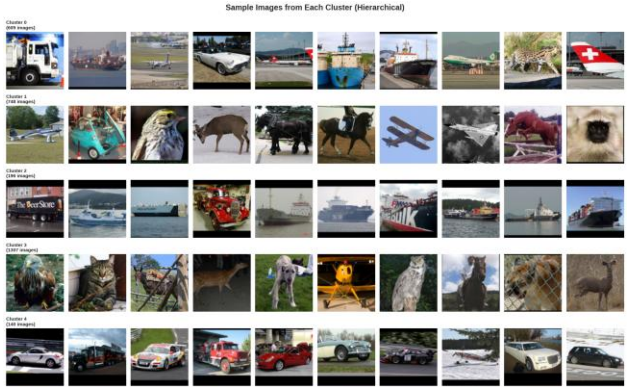

Fig. 6. Representative images from each discovered cluster under hierarchical clustering (k=5).

Figure 6 makes the clusters interpretable by showing sample members of each cluster. Cluster 2 visually contains many maritime scenes and ships with strong blue backgrounds, while Cluster 4 often contains land vehicles and darker road scenes. Clusters 1 and 3 frequently contain animals, but with differences in brightness and background composition. Cluster 0 appears mixed, containing transitional cases such as airplanes and vehicles with varied backgrounds. This qualitative evidence supports the conclusion that the discovered groups are primarily driven by low-level appearance cues rather than the STL-10 semantic labels. In addition, the presence of mixed clusters reinforces the quantitative finding that boundaries are weak and that some clusters behave like “catch-all” regions for images that share generic edge and color patterns.

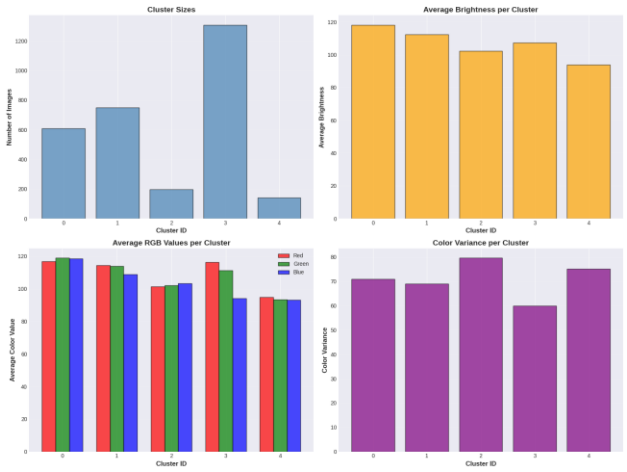

Fig. 7. Cluster size imbalance and per-cluster brightness and color summary statistics (hierarchical clustering).

Figure 7 reports important descriptive statistics of the discovered clusters. The cluster size bar chart shows a strong imbalance: one cluster (Cluster 3) contains 1,307 images (43.6%), while Cluster 4 contains only 140 images

(4.7%). Such imbalance is typical when the embedding has a dense central region and a few smaller, more distinctive regions; the large cluster often aggregates many moderately similar images that do not form sharp subgroups. The brightness plot indicates that Cluster 0 is, on average, the brightest (around 118), while Cluster 4 is the darkest (around 94), supporting a lighting-based interpretation. The average RGB plot suggests that clusters differ slightly in channel balance: clusters associated with water scenes or sky backgrounds tend to show stronger blue components, while some animal/vehicle clusters exhibit higher red components. The color variance plot indicates that some clusters have higher variability, suggesting greater diversity within the cluster, which can reduce Silhouette quality.

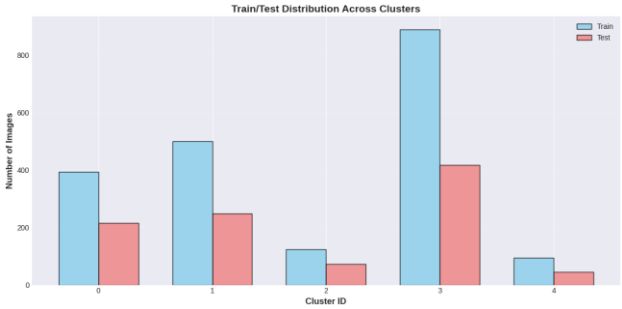

Fig. 8. Train/test distribution across clusters (hierarchical clustering).

Figure 8 compares how the train and test splits are distributed across the five clusters. The ratios per cluster remain close to the overall 2:1 split (approximately 63%–68% training per cluster in the summarized table), which indicates that clusters are not artifacts of a particular split. Instead, the same visual patterns are present in both train and test sets. This matters for generalization: if clusters were driven by dataset leakage or split-specific biases, one would observe strong skew where certain clusters contain mostly training images. The observed consistency supports the interpretation that the clustering is capturing dataset-wide appearance structure rather than memorizing idiosyncrasies of a subset.

Fig. 9. PCA dimensionality optimization using Silhouette Score and Davies–Bouldin Index.

Figure 9 directly supports the choice of 50 PCA components. As the number of components increases from 50 to 250, the Silhouette Score drops from about 0.0621 to about 0.0312, and the Davies–Bouldin Index increases from about 3.3690 to about 4.8986. This shows that retaining more components—although it preserves more variance—can worsen clustering because later components capture fine-grained variability that may not correspond to stable grouping structure. In other words, variance preservation is not the same as cluster separability. The best clustering behavior occurs when PCA compresses the feature space enough to focus distance computations on the most informative global directions.

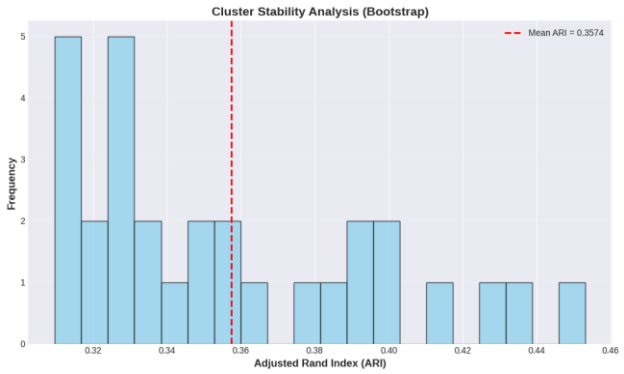

Fig. 10. Bootstrap stability analysis of hierarchical clustering measured by Adjusted Rand Index (ARI).

Figure 10 evaluates whether the discovered clusters are stable under resampling. Each bootstrap iteration samples 80% of the data with replacement, reclusters, and compares partitions using ARI. The distribution is centered around a mean ARI of approximately 0.3574 (red dashed line), with variability of about ±0.0413. An ARI near 0.35 indicates moderate agreement but falls below commonly used thresholds for highly stable clustering (often around 0.5). Practically, this means that many assignments are sensitive to which images are present in the sample, which is expected when cluster boundaries are gradual. The result is not a failure but a realistic diagnostic: the dataset likely does not contain strongly separated modes under the current feature representation, so small data perturbations can move borderline images between clusters.

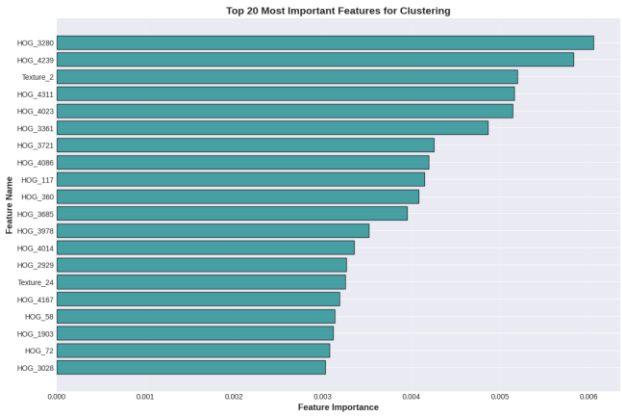

Fig. 11. Top 20 most important individual features for cluster discrimination (Random Forest on pseudo-labels).

Figure 11 ranks the top 20 individual features by importance when a Random Forest classifier is trained to predict cluster assignments as pseudo-labels. The dominance of features prefixed with HOG indicates that specific gradient orientation bins and spatial cells are the most informative discriminators. A small number of texture features also appear, but far below HOG in magnitude. This pattern suggests that the clustering is driven mainly by edge layouts and coarse shape cues (e.g., elongated structures for ships/airplanes, wheel/road edges for cars, contour patterns for animals). While this can be beneficial for separating certain visual groups, it can also bias clustering away from color/texture cues unless feature weighting or dimensionality balancing is applied.

Fig. 12. Total feature-family importance (Color Histogram, HOG, Statistical, Texture) for clustering.

Figure 12 summarizes feature importance by aggregating the Random Forest importances across each feature family. The result is strongly skewed: HOG accounts for approximately 96.64% of total importance, while texture (LBP) contributes about 1.71%, color histograms about 1.17%, and statistical features about 0.48%. A key implication is that the feature pipeline is effectively a shape-driven representation. This is partly expected because HOG contributes 4,356 of the 4,501 dimensions, so the feature space contains far more gradient detail than color or texture detail. In future iterations, one may consider weighted fusion, dimensionality balancing, or learned embeddings to avoid a single feature family dominating similarity.

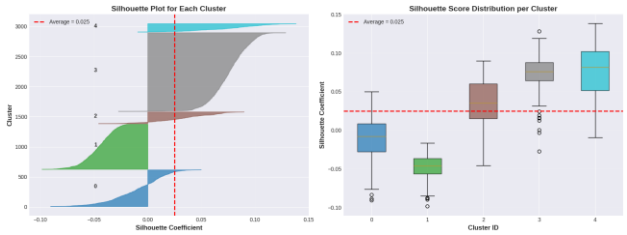

Fig. 13. Silhouette diagnostics: per-cluster silhouette regions and silhouette score distribution per cluster.

Figure 13 explains why the average Silhouette Score is low even when some clusters appear meaningful. The silhouette plot shows that clusters 0 and 1 have negative mean silhouette values (around -0.0102 and -0.0469 in the summarized table), indicating that many points in these clusters are closer to other clusters than to their own cluster center or neighborhood. In contrast, clusters 3 and 4 have positive means (around 0.0758 and 0.0764), indicating comparatively better cohesion and separation. The boxplot further shows that clusters differ in spread: some clusters contain many near-zero or negative points, consistent with being "mixed" regions. This supports the qualitative observations where one large cluster collects diverse images and where some clusters represent more distinctive visual patterns (e.g., land vehicles, dark scenes).

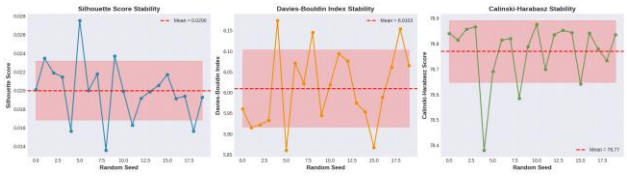

Fig. 14. Repeated-seed stability for Silhouette, Davies–Bouldin, and Calinski–Harabasz metrics (20 runs).

Figure 14 reports stability across 20 runs with different random seeds. The mean Silhouette is approximately 0.0200 with modest variability, while the mean Davies–Bouldin is around 6.0103 and the mean Calinski–Harabasz is around 78.77. The shaded bands indicate that variation across seeds is relatively small compared to the already low absolute values. This means that the method is reproducible in the sense that repeated runs give similar scores, but it also confirms that the underlying clustering structure is consistently weak. For practitioners, this is an important combination: low but stable scores imply that

the limitation is primarily the intrinsic data geometry (and representation choice), not random instability from the algorithm alone.

A. Quantitative Summary of Algorithm Performance

Table I summarizes the main clustering performance results at k=5 in the 50D PCA space. Hierarchical Clustering achieves the highest Silhouette Score (0.0250) and performs competitively on Davies–Bouldin (6.5016) and Calinski–Harabasz (69.01). K-Means obtains a lower silhouette (0.0212), while GMM is close (0.0242) but slightly worse overall. DBSCAN fails under the tested parameters by assigning all points as noise, which is consistent with the absence of dense, well-separated regions in high-dimensional space.

Algorithm	Silhouette	Cluster Sizes / Notes
K-Means	0.0212	[769, 413, 386, 428, 1004]
Hierarchical (Ward)	0.0250	[609, 748, 196, 1307, 140]
GMM (full cov)	0.0242	[378, 408, 1198, 662, 354]
DBSCAN	0.0000	All 3000 labeled as noise

Table I. Clustering algorithm performance comparison (k=5).

V. LIMITATIONS

The primary limitation is the weak clustering structure indicated by low Silhouette values (average 0.0250 for the best method). Natural images exhibit continuous changes in viewpoint, background, and illumination, so many points lie near boundaries and can plausibly belong to multiple clusters. This also explains the bootstrap ARI of 0.3574 ± 0.0413, which reflects moderate but not high stability. In practical terms, the clustering is informative for coarse organization but may be unreliable for tasks that require strong cluster purity or sharp separation.

A second limitation is representation imbalance. HOG contributes the overwhelming majority of dimensions and importance, which can bias similarity toward edges and shapes while underutilizing color and texture cues. Although HOG is useful, such dominance may hide meaningful color-based clusters (e.g., water vs grass) or texture-based clusters (e.g., fur vs metallic surfaces) that would emerge under different weighting. A third limitation is the failure of DBSCAN, which highlights the difficulty of density-based clustering in high-dimensional embeddings where distances become less informative. Finally, the cluster size imbalance suggests that at least one cluster acts as a broad “catch-all” region, reducing interpretability and depressing silhouette scores for that cluster.

VI. FUTURE WORK

Several improvements are promising. First, deep learning embeddings from pretrained convolutional networks such as ResNet can capture semantic similarity better than handcrafted features, potentially producing clusters that align more with object categories while remaining unsupervised. Second, nonlinear dimensionality reduction methods such as UMAP or autoencoders may preserve discriminative manifolds better than linear PCA. Third, ensemble or consensus clustering can combine complementary partitions from K-Means, hierarchical, and GMM to produce more stable groupings. Fourth, feature fusion can be improved by learning weights or balancing dimensionalities so that color, texture, and shape contribute more equally. Finally, more extensive semantic validation can be performed by measuring cluster purity and normalized mutual information with ground-truth labels strictly after clustering.

VII. CONCLUSION

This work explored unsupervised discovery of visual clusters in the STL-10 dataset using an interpretable handcrafted feature pipeline and rigorous validation. By extracting 4,501-dimensional descriptors and applying PCA, we showed that clustering quality improves when the embedding is compressed to an empirically optimized 50-dimensional space rather than retaining high variance dimensions. Among evaluated algorithms, Ward hierarchical clustering produced the best overall results at k=5, though the absolute structure remains weak, reflecting the intrinsic complexity of natural images. Stability analysis confirmed reproducible but moderately unstable assignments, and feature importance analysis revealed that HOG dominates similarity and cluster formation. Qualitative inspection showed that clusters correspond to low-level visual properties such as brightness, dominant color balance, and edge patterns rather than semantic labels. These findings provide both a useful baseline and a clear direction for future work in deep embeddings, improved fusion, and stability-enhancing clustering strategies.

REFERENCES.

[1] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.

[2] A. K. Jain, “Data clustering: 50 years beyond K-means,” Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, 2010.

[3] A. Coates, A. Ng, and H. Lee, “An analysis of single-layer networks in unsupervised feature learning,” in Proc. 14th Int. Conf. Artificial Intelligence and Statistics (AISTATS). PMLR, 2011, pp. 215–223.

[4] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[5] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “SURF: Speeded up robust features,” Computer Vision and Image Understanding, vol. 110, no. 3, pp. 346–359, 2008.

[6] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2005, pp. 886–893.

[7] M. J. Swain and D. H. Ballard, “Color indexing,” International Journal of Computer Vision, vol. 7, no. 1, pp. 11–32, 1991.

[8] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971–987, 2002.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Communications of the ACM, vol. 60, no. 6, pp. 84–90, 2017.

[10] I. T. Jolliffe and J. Cadima, Principal Component Analysis, 2nd ed. Berlin, Germany: Springer, 2016.

[11] L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE,” Journal of Machine Learning Research, vol. 9, no. 11, pp. 2579–2605, 2008.

[12] L. McInnes, J. Healy, N. Saul, and L. Grossberger, “UMAP: Uniform manifold approximation and projection,” Journal of Open Source Software, vol. 3, no. 29, p. 861, 2018.

[13] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in Proc. Fifth Berkeley Symp. Mathematical Statistics and Probability, 1967, pp. 281–297.

[14] D. Arthur and S. Vassilvitskii, “k-means++: The advantages of careful seeding,” in Proc. 18th Annual ACM-SIAM Symp. Discrete Algorithms, 2007, pp. 1027–1035.

[15] D. A. Reynolds, “Gaussian mixture models,” Encyclopedia of Biometrics, pp. 827–832, 2015.

[16] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in Proc. KDD, 1996, pp. 226–231.

[17] D. Mullner, “fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python,” Journal of Statistical Software, vol. 53, no. 9, pp. 1–18, 2013.

[18] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, 1987.

[19] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[20] T. Calinski and J. Harabasz, “A dendrite method for cluster analysis,” Communications in Statistics-Theory and Methods, vol. 3, no. 1, pp. 1–27, 1974.

[21] M. E. Celebi, H. A. Kingravi, and P. A. Vela, “A comparative study of efficient initialization methods for the k-means clustering algorithm,” Expert Systems with Applications, vol. 40, no. 1, pp. 200–210, 2013.

[22] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: A review,” ACM Computing Surveys, vol. 31, no. 3, pp. 264–323, 1999.

[23] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.