MAS480 Introduction to TDA 2023FALL Computational Project

Youngmin Ryou 20200406

November 2023

1 Figures

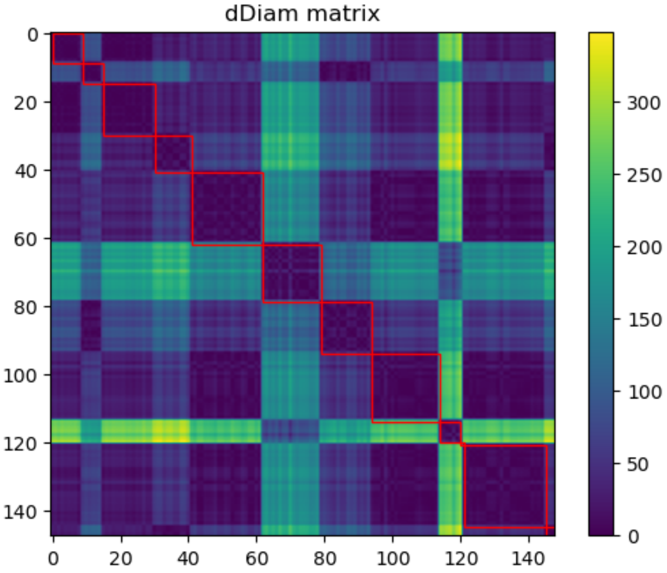

Figure 1: The heatmap of dDiam

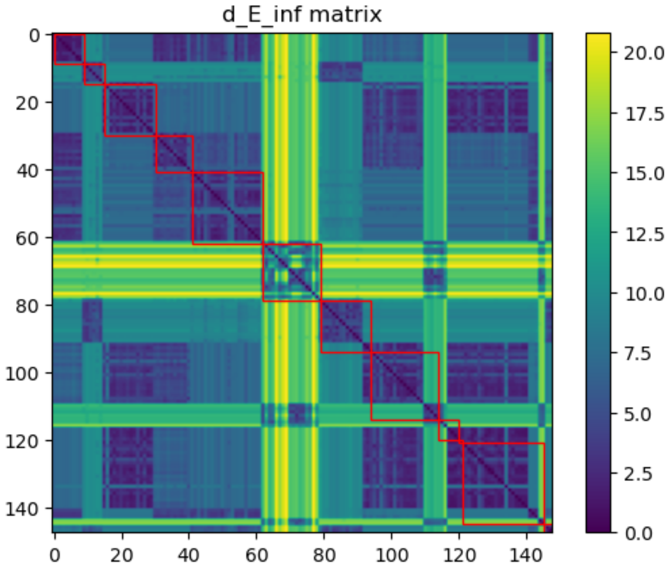

Figure 2: The heatmap of $d_{E,\infty}$

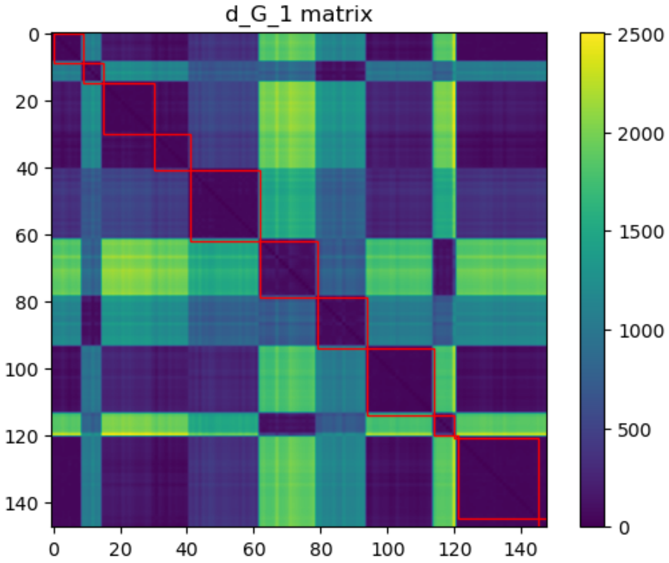

Figure 3: The heatmap of $d_{G,1}$

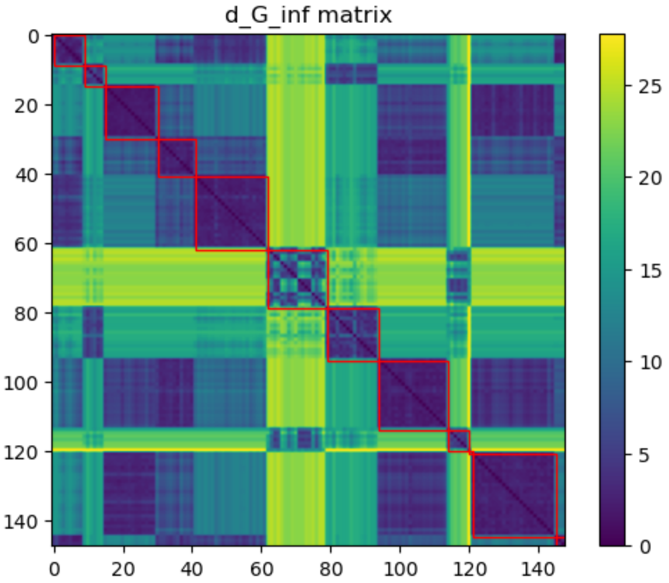

Figure 4: The heatmap of $d_{G,\infty}$

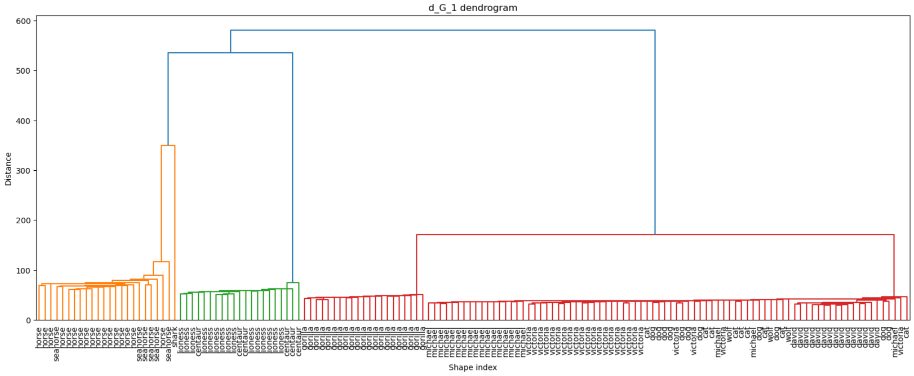

Figure 5: The Single Linkage Hierarchical Clustering Dendrogram of $d_{G,1}$

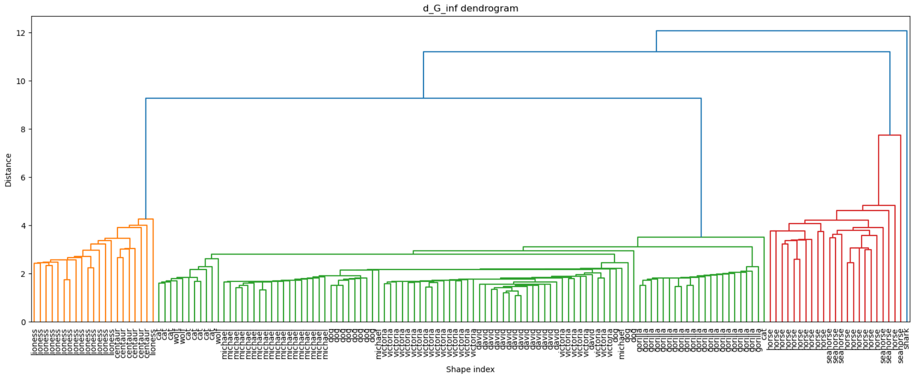

Figure 6: The Single Linkage Hierarchical Clustering Dendrogram of $d_{G,\infty}$

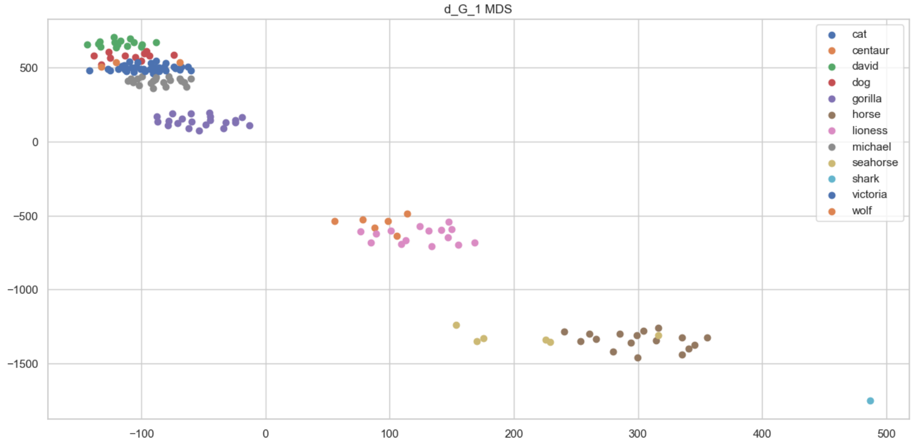

Figure 7: The MDS plot of $d_{G,1}$

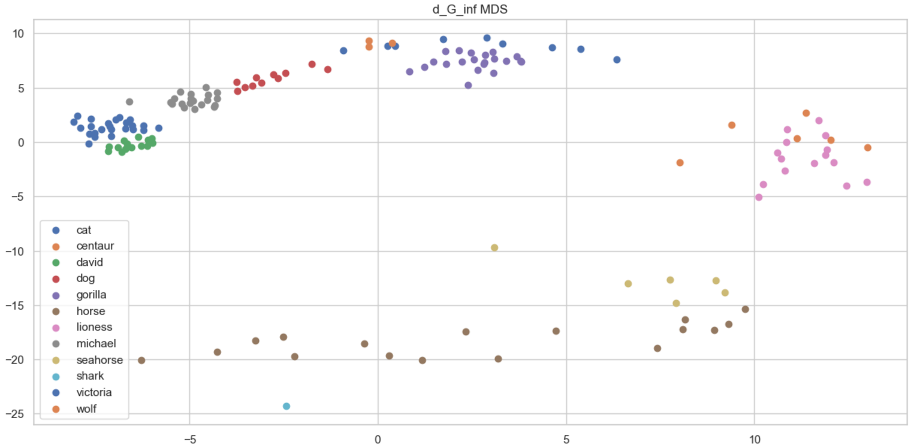

Figure 8: The MDS plot of $d_{G,\infty}$

2 Is dDiam an effective dissimilarity measure for distinguishing between different shape classes?

To assess the effectiveness of dDiam as a dissimilarity measure in distinguishing between different shape classes, we must consider its variability both within and across these classes. Ideally, dDiam should exhibit low variability within a class (low within-cluster scatter) and high variability between different classes (high between-cluster scatter).

Before our data analysis, the shapes were categorized into classes. If dDiam is an effective metric, we would expect low dDiam values for shapes within the same class and higher values for shapes across different classes. This expectation leads us to analyze the dDiam heatmap, where blocks consist of shapes of the same classes would have low values, indicating proximity within the same class. Conversely, areas outside these blocks should display greater distances.

2.1 Observations from the dDiam Matrix Heatmap

Observations of Figure 1 reveal two primary patterns:

1. **Within-Class Consistency:** The areas marked with red boxes, representing shapes of the same class, predominantly show colors close to zero. This indicates small within-cluster distances, suggesting that shapes of the same type maintain a consistent diameter. It implies that elements of the same class would cluster well together when grouped based on dDiam.

2. **Between-Class Overlap:** Even outside the red boxes, there are several blocks with colors near zero, indicating close distances between different classes. This suggests that not all different classes have significantly different diameters, posing a challenge for accurate class differentiation using dDiam alone.

2.1.1 Further Analysis on Specific Classes

In-depth analysis was conducted focusing on specific classes, notably:

- **Classes at Indices 60-80 and Around 120:** Identified as 'horse' and 'seahorse' classes, these groups displayed notably large distances when compared to shapes from nearly all other classes. Initially, it was hypothesized that the substantial difference in their diameters – the horse having a large diameter and the seahorse a small one – could be the primary factor contributing to this distinctness.

To validate these observations, an analysis involving the calculation of the mean diameter for each class was performed and I plotted the result in the Figure 9. Contrary to expectations, the seahorse class exhibited a larger mean diameter than anticipated. Despite this surprising finding, the mean diameters for both the seahorse and horse classes were significantly different from those of

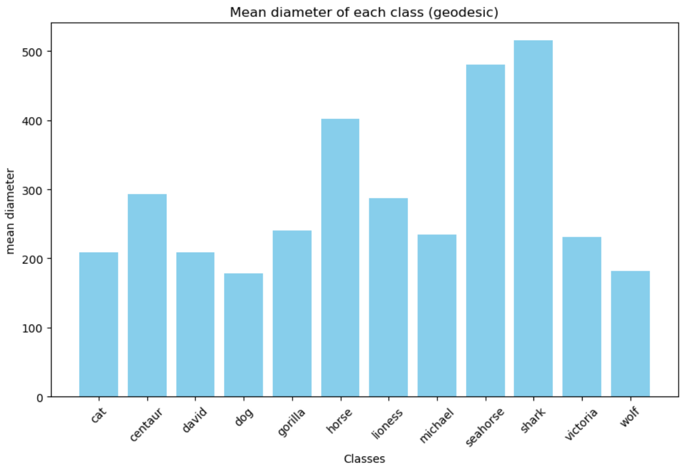

Figure 9: The mean diameter of each class, using geodesic distance.

other classes. This observation reinforces the original hypothesis: the marked disparity in the mean diameters of the horse and seahorse classes is a key factor in the heightened dDiam distances observed.

A detailed analysis was carried out using a bar plot to represent the within-cluster and between-cluster distances for each class. This analysis provides insights into the clustering performance of different metrics.

To calculate the within-cluster and between-cluster distances, we define the following:

Let C be the set of classes and D the dataset, with $I \in C$. Denote the set of datapoints belonging to class I as $D_I := \{(x, y) \in D \mid y = I\}$. Using a distance metric d, the within-cluster distance (WCD) for class I is computed as follows:

$$wcd_d(I) := \frac{\sum_{(x,y),(x',y') \in D_I} d(x, x')}{N} \tag{1}$$

where N is the number of summands.

Similarly, the between-cluster distance (BCD) for class I is calculated as:

$$bcd_d(I) := \frac{\sum_{(x,y) \in D_I,(x',y') \notin D_I} d(x, x')}{N} \tag{2}$$

Here, N is again the number of summands. Essentially, *wcd* and *bcd* represent the mean distance within and between clusters, respectively, for the class I.

Utilizing the dDiam distance metric, *wcd* and *bcd* were computed for each class and presented in Figure 10. The analysis revealed that within-cluster distances are significantly smaller than between-cluster distances. This finding suggests that dDiam is effective for clustering shapes within the same class.

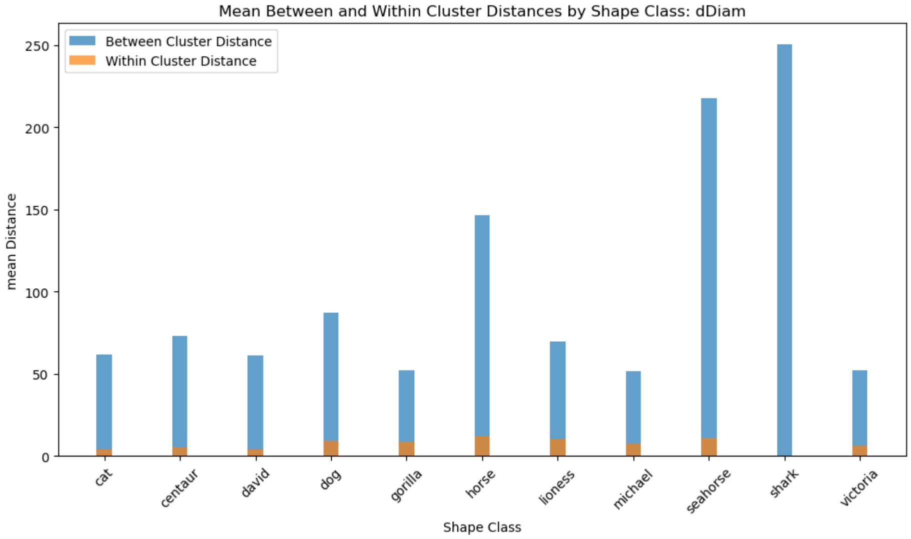

Figure 10: Mean Within-cluster-distance and Mean Between-Cluster-Distance for each class using dDiam as distance.

However, the observed lack of systematic variation in the between-cluster distances indicates a limitation in using dDiam for distinguishing between different classes. Despite this, dDiam's computational efficiency makes it a valuable tool for rapidly filtering objects based on diameter characteristics.

3 Can you envision a clustering scenario where shape diameters are valuable?

In envisioning scenarios where shape diameters, as measured by dDiam, are valuable for clustering, two distinct applications come to mind:

3.1 Animal Size Classification with Depth Cameras

In scenarios where the lengths between object classes vary significantly, such as in animal size classification, depth cameras can be particularly useful. For instance, categorizing animals into large, medium, and small sizes can be efficiently achieved with 3D scan data. This approach offers robust classification across different angles, contrasting with the extensive data collection from multiple angles required for visible light cameras. Utilizing dDiam in this context streamlines the model-building process by providing a direct, size-based classification metric.

3.2 Gesture Recognition in 3D Scans

Another application is in recognizing the state of objects with changing connections, like hand gestures. Consider a 3D scan of a hand making a gesture, with keypoints at the thumb and index finger tips. When forming an 'O' gesture by bringing these fingers together, the geodesic distance between the keypoints decreases significantly. This change in dDiam provides a quantitative measure of the gesture, indicating its closeness to an omnidirectional state.

3.3 Hierarchical Approach

Moreover, dDiam's consistency across different poses within the same object class, as evidenced by its small within-class variation, makes it a valuable tool for initial clustering stages. Coupled with its computational efficiency, dDiam can be effectively used in a hierarchical clustering approach. For instance, a tree-structured classifier could first utilize dDiam to categorize objects by size. Subsequent, more refined classifications could then focus on topological features within these size-based subsets.

This approach addresses scenarios where different objects, despite having varying sizes or approximate shapes, share similar topological features. Therefore, a hierarchical classification, starting with size differentiation using dDiam and then delving into finer topological details, enhances both accuracy and computational efficiency.

4 Compare the heat maps of $d_{E,\infty}$ and $d_{G,\infty}$. Which one shows a more favorable clustering outcome?

Upon examining the heatmaps, a good distance measure for clustering is indicated by colors close to zero within the red boxes (same class) and higher values in other areas.

In the $D_{E,\infty}$ heatmap of Figure 2, even within the same class (red boxes), there are noticeable variations in distance, suggesting less effective clustering.

Conversely, the $D_{G,\infty}$ heatmap in Figure 4 shows more consistency within classes, with less problematic variations in distance.

4.1 Comparative Analysis of Within and Between Class Distances

Comparing the mean within-cluster distance and mean between-cluster distance in $D_{G,\infty}$ reveals consistently smaller within-cluster distances. In $D_{E,\infty}$, however, both within-cluster and between-cluster distances are generally reduced. Notably, the mean between-cluster distances for certain classes (e.g., seahorse and michael) significantly increased in $D_{E,\infty}$. This is evident in the heatmaps

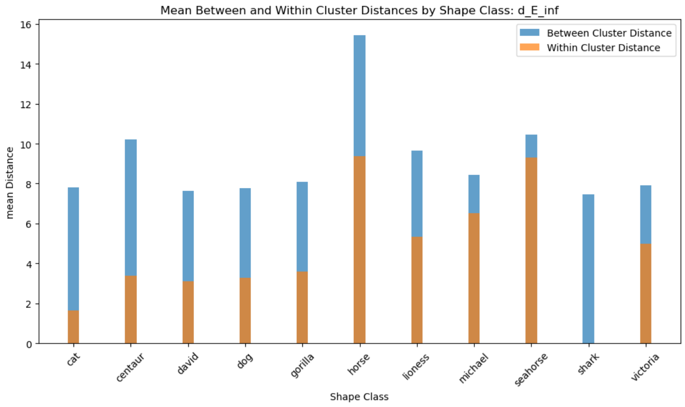

Figure 11: The within-cluster distance and between-cluster distance using $d_{E,\infty}$

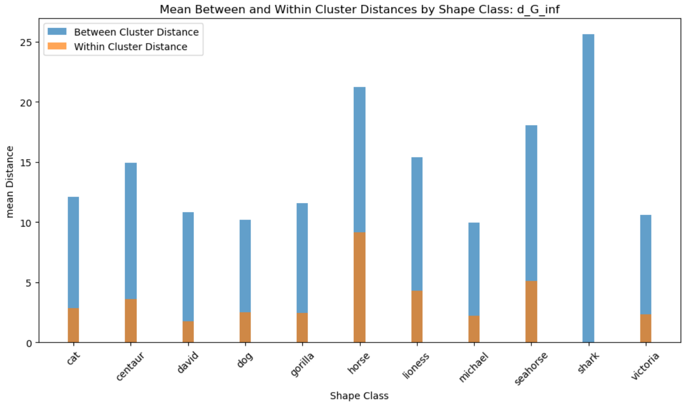

Figure 12: The within-cluster distance and between-cluster distance using $d_{G,\infty}$

Figure 2, Figure 4, and further supported by the following bar plots of Figure 11 and Figure 12.

4.2 Complete Linkage Hierarchical Clustering

I computed Complete Linkage Hierarchical Clustering and plotted the results on Figure 6 and Figure 13. Applying complete linkage hierarchical clustering to both $D_{E,\infty}$ and $D_{G,\infty}$ yields distinct results:

1. $D_{E,\infty}$ struggles to differentiate shapes of similar sizes, resulting in premature clustering of distinct classes in the dendrogram.
2. $D_{G,\infty}$, on the other hand, more effectively clusters objects of the same class at lower levels, as evident in the dendrogram.

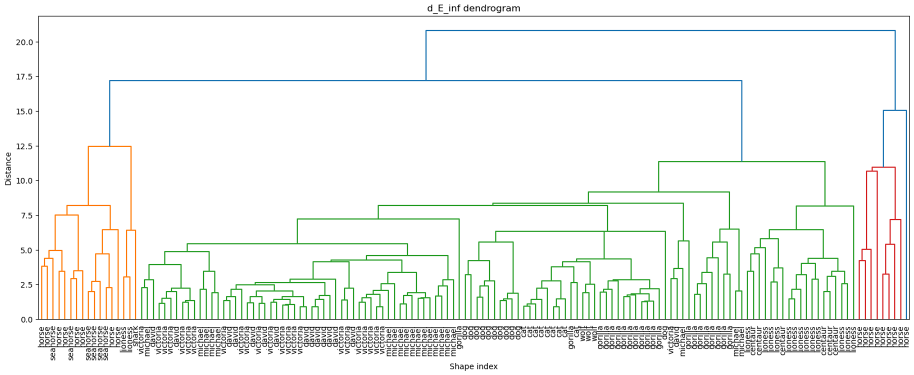

Figure 13: The complete linkage hierarchical clustering dendrogram using $d_{E,\infty}$

4.3 Quantitative Analysis Using Clustering Metrics

Next, we will employ widely used clustering evaluation metrics, such as the silhouette score and the Adjusted Rand Index, to quantitatively assess the clustering results. Before the analysis, explain the evaluation metrics I used.

4.3.1 Silhouette Score

In cluster analysis, the Silhouette Score is an established metric for evaluating the effectiveness of clustering algorithms. It measures the degree of similarity of an individual data point to its cluster compared to other clusters. For a given data point i, the Silhouette Score $s(i)$ is calculated as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3}$$

where $a(i)$ is the average distance from the i-th data point to the other points in the same cluster, and $b(i)$ is the smallest average distance from the i-th data point to points in a different cluster, minimized over all clusters. The Silhouette Score ranges from -1 to 1, with values close to +1 indicating a strong match to the assigned cluster and distinct separation from other clusters.

4.3.2 Adjusted Rand Index (ARI)

The Adjusted Rand Index (ARI) is a refinement of the Rand Index, a measure of the similarity between two clusterings. ARI adjusts for the chance grouping of elements, making it a reliable metric for comparing a clustering result against a ground truth classification. The ARI is defined as:

$$\text{ARI} = \frac{\text{Index} - \text{Expected Index}}{\text{Max Index} - \text{Expected Index}}$$

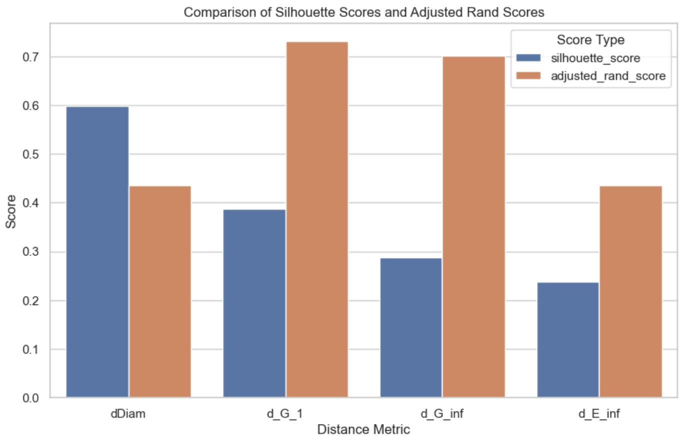

Figure 14: The silhouette score and Adjusted rand index with k-nn with different distances.

Here, the Index is the sum of agreements between two clusterings (the number of pairs of elements in the same or different groups in both clusterings), and the Expected Index is the sum of agreements expected by chance. The ARI value ranges from -1 to 1, where values closer to 1 indicate a high similarity between the clustering result and the ground truth.

4.3.3 Results and Analysis

In our analysis, we focus on a dataset comprising a total of 12 distinct classes, thus establishing the ground truth number of clusters as 12. To evaluate the performance of each distance metric, we conduct K-nearest neighbors (KNN) clustering for each metric, setting $K = 12$. The effectiveness of these metrics is then quantitatively assessed using two key measures: the Silhouette score and the Adjusted Rand Index (ARI). The results were plotted in Figure 14 and reported on Table 2.

Both the Silhouette score and ARI are crucial in this analysis, with higher values indicating superior clustering performance. Although a statistical significance test was not conducted, observations from this dataset suggest that geodesic distance metrics outperform the Euclidean distance in both the Silhouette score and ARI.

Based on the obtained results, which are systematically presented in the accompanying table and figures, we conclude that $d_{G,\infty}$ demonstrates better clustering performance compared to $d_{E,\infty}$. This conclusion is drawn from the comparative analysis of the two metrics, which consistently show higher values for $d_{G,\infty}$ in the context of our dataset.

Shape Class	$d_{G,1}$			$d_{G,\infty}$		
	E[BCD]	**E[WCD]**	**WCD / BCD**	**E[BCD]**	**E[WCD]**	**WCD / BCD**
cat	566.75	45.37	0.08	12.11	2.86	0.236940
centaur	850.31	70.63	0.08	14.95	3.61	0.241649
david	679.34	37.21	0.05	10.84	1.75	0.161473
dog	610.73	45.01	0.07	10.18	2.49	0.244886
gorilla	577.77	56.64	0.09	11.59	2.45	0.211509
horse	1398.46	95.77	0.06	21.19	9.17	0.432796
lioness	896.10	75.96	0.084	15.41	4.3	0.279962
michael	546.15	42.85	0.078	9.94	2.21	0.222669
seahorse	1366.71	91.11	0.066	18.03	5.12	0.284269
shark	1820.12	0.00	0.00	25.64	0.0	0.000000
victoria	565.78	42.49	0.075	10.57	2.34	0.222296

Table 1: average within-cluster distance, average between-cluster distance and ratio between them, of $d_{G,1}$ and $d_{G,\infty}$

5 Among the clustering results presented in Sections 3.2, 3.3, and 3.4, which one produced the most favorable clustering outcome?

To compare the clustering results, we employed four distinct approaches:

5.1 Qualitative Analysis of Heatmap Images

A suitable distance measure for clustering should show values close to zero within red boxes (indicating within-class proximity) and higher values in other areas. From this perspective, $d_{G,1}$ in Figure 3 consistently demonstrates values near zero within the red boxes, indicating small within-cluster scatter. However, there are regions outside the red boxes with very small pairwise distances, suggesting potential challenges in differentiating classes. On the other hand, $d_{G,\infty}$ in Figure 4 exhibits a slightly higher within-cluster distance but lower between-cluster scatter, suggesting better class separation compared to $d_{G,1}$. $d_{E,\infty}$ in Figure 2 has higher within-cluster distances and lower between-cluster distances compared to $d_{G,\infty}$, leading to the conclusion that $d_{G,\infty}$ is superior to $d_{E,\infty}$. Therefore, the critical comparison is between $d_{G,1}$ and $d_{G,\infty}$.

5.2 Quantitative Analysis via WCD/BCD Ratios

Analyzing the ratio of Within-Cluster Distance to Between-Cluster Distance (WCD/BCD) of Table 1, we find that $d_{G,1}$ has a more favorable ratio. This assessment is further supported by MDS plot analysis.

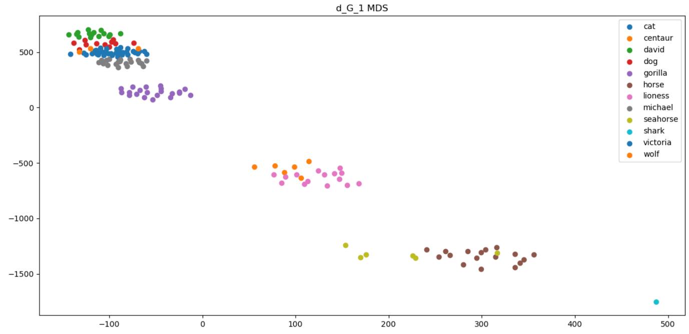

Figure 15: The MDS plot of $d_{G,1}$

Measure	**Silhouette Score**	**Adjusted Rand Score**
dDiam	**0.5989794079656797**	0.43632156260542454
d_G_1	0.3867879046416054	**0.7313297097587836**
d_G_inf	0.28840358589163356	0.7021907305204439
d_E_inf	0.23860844748613408	0.43603386019569174

Table 2: Comparison of Silhouette and Adjusted Rand Scores for Different Measures

5.3 MDS Plot Observations

The MDS plot Figure 7 compared to the Figure 8 reveals that $d_{G,1}$ has a smaller within-class scatter compared to $d_{G,\infty}$. Although the distances between different class clusters in $d_{G,1}$ appear close, the clusters do not overlap spatially. Contrastingly, $d_{G,\infty}$ shows closely intertwined different classes, such as lioness-wolf and michael-victoria-gorilla, complicating their separation. Thus, the MDS plot analysis favors $d_{G,1}$ for effective clustering.

5.4 Silhouette Scores and Adjusted Rand Index Comparison

We applied the K-Nearest Neighbors (K-NN) algorithm with $k = 12$ using both $d_{G,1}$ and $d_{G,\infty}$ as distance metrics. Subsequently, we computed the Silhouette score and the Adjusted Rand Index (ARI) for each clustering, using the provided Ground Truth labels for each shape. The results are presented in Table 2. Analysis of both the Silhouette score and ARI indicates that $d_{G,1}$ achieves superior clustering performance compared to $d_{G,\infty}$.

Based on the analyses conducted across multiple dimensions, we conclude that $d_{G,1}$ is the most effective distance measure for clustering in the context of our study.

6 Strategies for Improving Clustering Algorithms

6.0.1 Refined Clustering Algorithms Beyond K-NN

Exploring advanced clustering techniques like DBSCAN (Density-Based Spatial Clustering of Applications with Noise) could offer more nuanced outcomes. DBSCAN's ability to identify clusters of varying shapes and densities presents a versatile alternative to K-NN.

6.0.2 Dimensionality Reduction Techniques

Implementing dimensionality reduction methods, such as Uniform Manifold Approximation, can preserve topological and geometric features of the data. This is particularly beneficial when clustering labels are predetermined and can aid in supervised clustering or classification. However, this approach may entail information loss that could negatively impact the test accuracy of clustering. Moreover, preserving key topological and geometric features often requires significant computational resources.

6.0.3 Ensemble Clustering Approaches

Ensemble clustering can enhance the generalizability and robustness of clustering results. By combining multiple models or runs, ensemble methods aim to provide more reliable outcomes, albeit at higher computational and memory costs.

While these strategies promise improvements in clustering, they are not without challenges. Manual fine-tuning of algorithms can be labor-intensive and complex, especially with high-dimensional data. Advanced algorithms, while more capable, also demand more processing time. Thus, a balance must be struck to achieve effective and efficient clustering results.