Unsupervised Hierarchical Visual Taxonomy of Marble Natural Stone Using Cluster-Aware Self-Supervised Vision Transformers

Margarida Tânger de Oliveira Figueiredo *, Carlos M. A. Diogo, Gustavo Paneiro, Pedro Amaral, António Alves de Campos

Posted Date: 17 March 2026

doi: 10.20944/preprints202603.1344.v1

Keywords: self-supervised learning; vision transformer; DINO; deep clustering; hierarchical clustering; marble classification; unsupervised visual taxonomy

 doi:10.20944/preprints202603.1344.v1

Article

Unsupervised Hierarchical Visual Taxonomy of Marble Natural Stone Using Cluster-Aware Self-Supervised Vision Transformers

Margarida Tânger de Oliveira Figueiredo [1,*], Carlos M. A. Diogo [2], Gustavo Paneiro [2], Pedro Amaral [3] and António Alves de Campos [2]

[1] Mineral and Energy Resources Engineering Department, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal

[2] CERENA, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal

[3] LAETA, IDMEC, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal

* Correspondence: margarida.figueiredo@tecnico.ulisboa.pt

Abstract

The marble industry relies on proprietary commercial names rather than objective visual categories, creating market inefficiencies for stakeholders who select stones based on appearance. Supervised classification methods perpetuate this problem by replicating inconsistent commercial labels instead of discovering intrinsic visual structure. We propose an unsupervised pipeline combining a two-stage training strategy, pure self-supervised pretraining followed by cluster-aware fine-tuning of a DINO Vision Transformer, with UMAP dimensionality reduction and Ward's agglomerative hierarchical clustering. Systematic ablation studies on 1,540 marble images spanning 10 commercial varieties validate each design choice: cluster-aware training at k=10 yields superior embeddings over the self-supervised baseline (Silhouette Score 0.778 vs. 0.761; Davies–Bouldin Index 0.293 vs. 0.364), UMAP compression to five dimensions resolves high-dimensional noise pathologies, and Ward's linkage produces the most compact partitions. The resulting taxonomy reveals three phenomena invisible to commercial classification: cross-category merging of visually indistinguishable stones carrying different market names, intra-category splitting of heterogeneous sub-populations within single varieties, and coherent grouping where commercial and visual boundaries coincide. We further demonstrate that standard extrinsic metrics are misaligned with unsupervised taxonomy objectives when reference labels encode the inconsistencies the method aims to resolve. This work provides a validated methodology for data-driven visual classification in the natural stone industry and a transferable template for domains with unreliable labelling conventions.

Keywords: self-supervised learning; vision transformer; DINO; deep clustering; hierarchical clustering; marble classification; unsupervised visual taxonomy

1. Introduction

The natural stone industry is characterized by a persistent reliance on proprietary commercial names rather than systematic visual classification [1,2]. This market-driven naming convention creates a fragmented landscape in which the same visual appearance may carry different commercial designations depending on the supplier, quarry location, or regional tradition, while a single commercial name may encompass significant visual diversity [1,3]. For buyers, designers, and specifiers who select marble primarily based on aesthetic appeal [4], this nomenclature chaos introduces substantial inefficiencies: stones cannot be reliably compared across suppliers, visual alternatives are difficult to identify when a preferred variety becomes unavailable, and the absence of objective grouping criteria obscures the relationship between a stone's appearance and its underlying geological properties [5]. Consequently, the

 doi:10.20944/preprints202603.1344.v1

industry lacks a standardized visual taxonomy that would enable transparent, systematic organization of marble varieties based on their intrinsic appearance.

Various classification approaches have been explored to address this challenge [6], but they remain fundamentally misaligned with the objective of discovering natural visual structure. Early methods focused on quality grading and defect detection [7], employing handcrafted feature extraction techniques such as GLCM [8] and Local Binary Patterns [9] to distinguish among quality categories or to identify surface imperfections. While effective for industrial quality control, these systems are not designed to capture the aesthetic variability that defines commercial marble varieties. More recent approaches have employed supervised deep learning to classify stones into predefined commercial categories [10–12], achieving high accuracies on their respective datasets. However, these supervised methods perpetuate the very problem they aim to solve. By learning to replicate commercial labels as ground truth, they encode the inconsistencies and market-driven conventions of existing nomenclature rather than discovering objective visual groupings.

A supervised approach is fundamentally unsuited to our objective. Commercial labels are assigned by traders and quarry operators based on market conventions, regional traditions, and supplier preferences, not on systematic visual criteria. Two visually identical stones may carry different commercial names from different suppliers, while a single commercial name may encompass stones with markedly different veining patterns, base colors, or textural characteristics. Our goal is precisely to transcend these inconsistent labels and identify objective visual groupings based solely on appearance, organized hierarchically to reflect natural relationships, ranging from broad visual families to fine-grained textural distinctions. This requires an unsupervised methodology capable of learning discriminative visual representations without relying on external annotations.

In this paper, we propose an unsupervised pipeline for hierarchical visual taxonomy of marble varieties that combines cluster-aware self-supervised learning with nonlinear dimensionality reduction and agglomerative hierarchical clustering. The methodology is motivated by three converging developments in computer vision and unsupervised learning: the emergence of Vision Transformers (ViTs) as powerful feature extractors for structured visual patterns, the maturation of self-supervised learning as a paradigm for representation learning from unlabeled data, and the integration of clustering objectives directly into representation learning to produce feature spaces optimized for grouping tasks. By synthesizing these approaches, we demonstrate that meaningful visual taxonomies can be automatically discovered and hierarchically organized without any reliance on commercial labels, providing both a practical tool for the marble industry and a methodological template for similar challenges in other material science domains where existing labels are inconsistent or unreliable.

The pipeline integrates three components: (1) cluster-aware self-supervised training using a hybrid Cluster-Aware Distillation with No Labels (CA-DINO) strategy, (2) non-linear dimensionality reduction via UMAP to preserve the local manifold structure of learned embeddings, and (3) agglomerative hierarchical clustering using Ward's linkage to reveal nested relationships in the visual feature space. Each design choice is motivated by specific challenges of marble variety identification and validated through systematic ablation studies.

The main contributions are threefold:

1. A validated unsupervised pipeline for the visual taxonomy of marble. We propose and validate a complete pipeline for hierarchical grouping of marble varieties based solely on visual features. Unlike supervised approaches that replicate commercial labels, our method discovers objective visual structure, producing taxonomies grouping stones by perceptually salient characteristics such as vein density, base coloration, and textural organization. Comprehensive ablation studies demonstrate the contribution of each component and identify optimal hyperparameter configurations.
2. Evidence that visual similarity transcends commercial categories. Through quantitative metrics and qualitative dendrogram analysis, we demonstrate that the learned visual hierarchy does not replicate commercial classifications. The model correctly groups visually similar stones under different commercial names while separating visually distinct subpopulations within a single

 doi:10.20944/preprints202603.1344.v1

commercial category. This validates the pipeline's ability to discover intrinsic visual structure independent of market-driven naming conventions.

3. Methodological insights for unsupervised visual taxonomy evaluation. We demonstrate that standard clustering validation metrics, particularly extrinsic measures such as Adjusted Rand Index and Normalized Mutual Information, are fundamentally misaligned with the task of discovering new visual structure when the ground-truth labels themselves represent the problem being solved. Through concrete examples, we show that these metrics penalize correct visual groupings diverging from commercial labels, establishing that rigorous qualitative evaluation of hierarchical coherence is essential for this class of problems.

Beyond immediate application in the marble industry, this work provides a methodological foundation for similar challenges in materials science, manufacturing, and other domains in which existing categorical labels are driven by tradition, market convention, or regional variation rather than by systematic feature-based criteria.

The remainder of this paper is organized as follows. Section 2 reviews related work in marble classification, self-supervised learning, deep clustering, and hierarchical visual organization. Section 3 describes the dataset and presents the proposed pipeline, detailing the training procedure, feature extraction, and hierarchical clustering approach. Section 4 presents the results of systematic ablation studies that validate each pipeline component, followed by qualitative and quantitative analyses of the final visual taxonomy. Section 5 discusses broader implications, including the role of evaluation metrics, learned feature priorities, and practical applications. Section 6 concludes with a summary of contributions and directions for future work.

2. Related Work

2.1. Automated Marble Classification

Automated marble classification has been investigated for over three decades, initially focusing on quality control through defect detection and slab grading. The earliest methods employed classical computer vision techniques: RGB color-space analysis distinguished between broad categories, while morphological operations enabled surface-defect detection [7]. Subsequent work applied Gabor filters and wavelets for multi-scale textural feature extraction [13]. Among handcrafted descriptors, Gray-Level Co-occurrence Matrices (GLCM) and Local Binary Patterns (LBP) proved particularly effective, with LBPs demonstrating continued utility in modern hybrid systems [8,9].

The inherent subjectivity and time-intensive nature of manual marble classification, combined with proprietary data policies in industry settings, motivated a shift toward unsupervised methods [14]. Extracted features were typically fed to clustering algorithms such as k-means [15,16], followed by classification using Support Vector Machines or neural networks [9,17], often with PCA-based dimensionality reduction [13].

The adoption of Convolutional Neural Networks (CNNs) marked a significant advance in classification accuracy [10–12]. However, CNNs face inherent limitations in identifying marble varieties. Their localized receptive fields prioritize local features that may not capture the global continuity essential for distinguishing visually similar varieties. Differentiation among marble types often relies on subtle relationships among veining density, color gradients, and textural organization, features that require both fine-grained local analysis and an understanding of broader spatial context. More fundamentally, supervised methods share a conceptual flaw: if models are trained on commercial labels as ground truth, they will learn to replicate existing market-driven nomenclature rather than to discover intrinsic visual structure.

Recent work has demonstrated that Self-Supervised Learning (SSL) overcomes this limitation in industrial and geological domains. Brondolo and Beaussant [18] applied a second version of the Distillation with No Labels (DINO) methodology to CT-scanned rock-sample analysis, achieving state-of-the-art performance without annotated data. Scabini et al. [19] benchmarked Vision Transformers as texture feature extractors on material databases, demonstrating that DINO-pretrained models achieve

 doi:10.20944/preprints202603.1344.v1

superior material-classification performance. Zhu et al. [20] applied DINOv2 to concrete crack detection, outperforming supervised models where precise annotation is impractical. These successes establish SSL as a viable paradigm for domains where labels are scarce, inconsistent, or unreliable.

2.2. Self-Supervised Learning and Vision Transformers

In domains where large-scale expert annotation is infeasible or where labels themselves are inconsistent, Self-Supervised Learning (SSL) has emerged as a powerful alternative paradigm. SSL methods learn rich visual representations directly from unlabeled data by solving pretext tasks that reveal inherent structure [21], circumventing the reliance on human-provided labels that constrain supervised approaches [22]. The field has evolved from clustering-based methods such as DeepCluster [23], through contrastive frameworks including SimCLR [24] and Momentum Contrast (MoCo) [25], to distillation-based approaches that avoid explicit negative pairs.

Among these methods, DINO [26] has demonstrated particularly strong performance for fine-grained visual tasks. DINO employs a student-teacher distillation framework in which a student Vision Transformer is trained to match the output distribution of a momentum-updated teacher network across multiple augmented views of the same image. Unlike contrastive methods requiring large batches of negative examples, DINO's knowledge distillation produces semantically meaningful features with strong localization properties, as evidenced by the emergence of object-centric attention maps without a supervised signal. This capability is especially relevant for marble classification, where models must simultaneously capture fine local details like vein direction, density, branching patterns, and global textural organization, including color gradients, spatial continuity, and structural layout. DINO's architectural foundation, the Vision Transformer [27], models global relationships among image patches via self-attention mechanisms [28], making it well-suited to capturing the structural continuity characteristic of marble patterns.

The robustness of DINO-learned representations to variations in scale, rotation, and lighting further enhances applicability to industrial image datasets, which often exhibit inconsistent acquisition conditions, scanner artifacts, and variable illumination. These properties position DINO as an ideal foundation for learning visual features without relying on inconsistent commercial labels.

2.3. Deep Clustering

While SSL provides a mechanism for learning discriminative features without labels, our objective requires organizing them into coherent, interpretable clusters. Traditional clustering algorithms perform effectively in low-dimensional spaces but degrade significantly when applied to high-dimensional data, where the curse of dimensionality renders standard distance metrics uninformative [29]. This necessitates learning a compressed representation before clustering [30].

However, a sequential pipeline in which feature learning and clustering are performed independently suffers from a fundamental representational disconnect [31]. The objective guiding feature learning, maximizing agreement between augmented views in SSL, is divorced from the downstream goal of forming well-separated clusters. This misalignment often produces feature spaces that are suboptimal for clustering, as learned representations prioritize view invariance over cluster separability.

Deep Clustering (DC) addresses this limitation by jointly optimizing feature representations and cluster assignments [31,32]. By integrating a clustering-specific loss directly into deep network training, DC creates a feedback loop: evolving cluster assignments refine the feature space, and improved features yield more accurate clusters. Deep Embedded Clustering (DEC) employs Kullback-Leibler divergence to iteratively sharpen assignments, with gradients backpropagated to produce inherently structured features [30]. Empirical evidence demonstrates this integrated approach substantially outperforms two-stage pipelines [30,31].

Contemporary approaches leverage SSL as the backbone for feature learning. Caron et al. [33] integrate online clustering directly into SSL via learnable prototypes and optimal-transport assignment between augmented views, providing the foundation for cluster-aware DINO variants. Contrastive Clustering performs instance-level and cluster-level contrastive learning simultaneously,

 doi:10.20944/preprints202603.1344.v1

enforcing both feature discrimination and cluster structure [34]. Prototypical Contrastive Learning alternates between discovering cluster prototypes via k-means and optimizing representations to align with assigned prototypes [35]. These methods demonstrate that combining SSL's representational power with explicit clustering objectives produces feature spaces optimized for discovering visual categories without supervision, motivating our use of cluster-aware DINO [36].

2.4. Hierarchical Clustering of Visual Features

Beyond producing discriminative features and encouraging cluster structure, our objective requires organizing the discovered visual groups into an interpretable hierarchy that reveals nested relationships, ranging from broad families to fine-grained distinctions. While flat clustering methods partition data into disjoint categories, hierarchical clustering produces dendrograms that reflect multi-level organization, which is essential for a visual taxonomy that serves diverse stakeholders with varying granularity requirements.

Recent work establishes that agglomerative hierarchical clustering applied to deep visual embeddings produces semantically meaningful taxonomies. Naumov et al. [37] demonstrated that agglomerative methods applied to ResNet features on ImageNet datasets (up to 4.5 million images) yield dendrograms with coherent semantic structure. Yang et al. [38] established that agglomerative clustering and deep feature learning are mutually reinforcing, validating bottom-up hierarchical organization as a natural paradigm for discovering visual categories without supervision. Among linkage criteria, Ward's linkage criterion, with its variance-minimization objective, produces balanced, compact clusters at each hierarchical level, a property particularly desirable for taxonomies intended for practical industrial use, where interpretability and consistency across hierarchical levels are critical [39].

3. Materials and Methods

3.1. Dataset

The dataset used in this study comprises 1,540 digital images of marble slabs at 512×512 pixels, distributed across 10 commercial varieties: Estremoz Creme (65 images), Dante Grey (69), Branco Carrara (13), Arabescato Brown (87), Calacata (28), Bardiglio (266), Irish Black (16), Branco Peletigre (185), Exotic Ambar (333), and Ruivina (478). Figure 1 displays a representative sample from each variety.

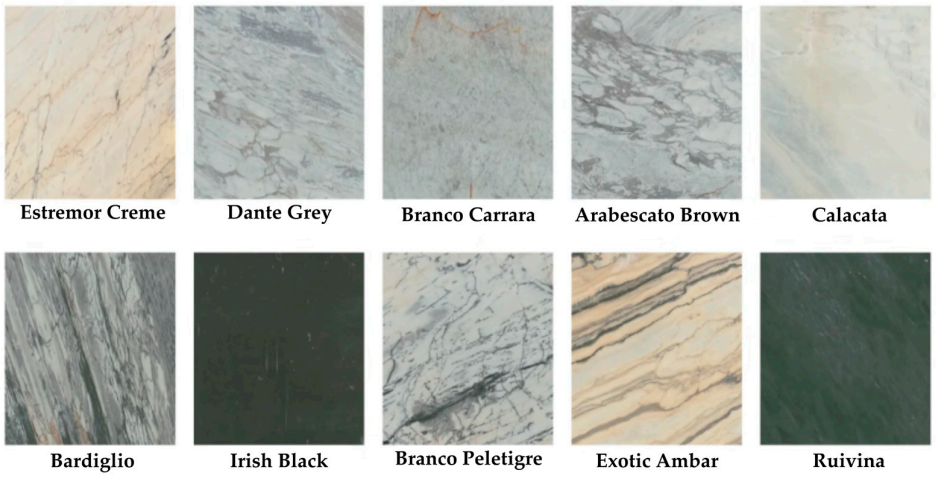

Figure 1. Representative sample image for each of the 10 marble varieties. The total number of images per variety is indicated below each name.

The dataset presents significant real-world challenges, including substantial class imbalance, ranging from 13 images for Branco Carrara to 478 for Ruivina, and minor scanner-induced distortions,

 doi:10.20944/preprints202603.1344.v1

both of which were intentionally retained to assess the pipeline's tolerance to imperfect data. The dataset is particularly well-suited for evaluating an unsupervised visual taxonomy, as it exhibits variability along two key axes. First, it exhibits high intra-class variability, with varieties such as Ruivina, Exotic Ambar, and Bardiglio showing considerable diversity in pattern and color within the same commercial category (Figure 2). Second, it exhibits low inter-class variability, containing visually similar varieties from different commercial designations, such as Branco Peletigre and Dante Grey.

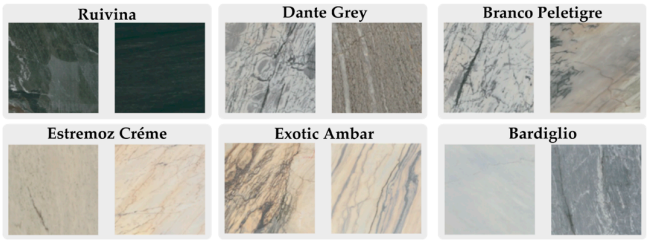

Figure 2. Examples of intra-class visual variability are present in the dataset. Images within the same commercial variety exhibit substantial differences in veining density, base color, and textural organization.

Importantly, the commercial variety labels are used exclusively for post-hoc evaluation and are never provided to the model during training. These labels serve as a reference for computing extrinsic validation metrics and for contextualizing the learned visual groupings; they do not constitute the learning target. To ensure evaluation was performed on original data, no offline preprocessing or data augmentation was applied to the source images prior to the training pipeline.

3.2. Proposed Pipeline

We propose a three-component pipeline for unsupervised hierarchical visual taxonomy of marble varieties: (1) a hybrid self-supervised training strategy using CA-DINO [36], which produces a discriminative, clustering-optimized feature space from unlabeled images; (2) non-linear dimensionality reduction using UMAP, which preserves the local manifold structure of the learned embeddings; and (3) agglomerative hierarchical clustering using Ward's linkage, which organizes the reduced embeddings into a coherent, interpretable visual hierarchy. The design of each component is motivated by the specific challenges of this domain and validated through systematic ablation studies presented in the results section. Figure 3 provides an overview of the proposed pipeline and the corresponding validation studies.

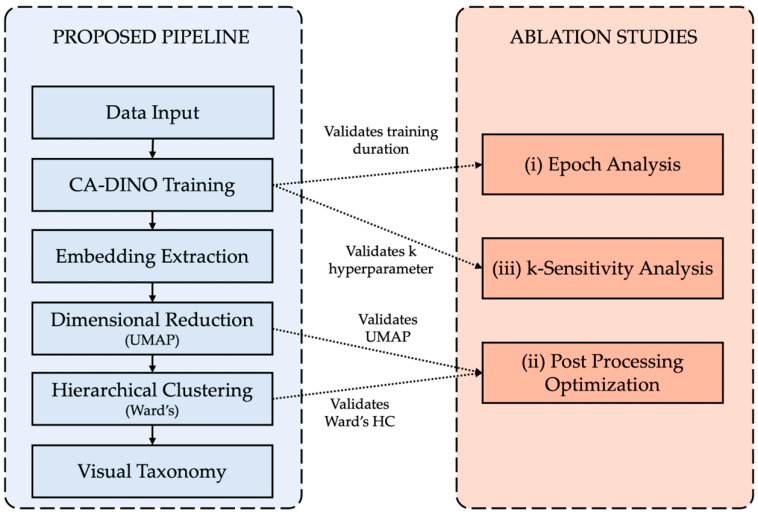

Figure 3. Overview of the proposed pipeline (left) and the ablation studies used to validate each design choice (right).

 doi:10.20944/preprints202603.1344.v1

3.3. Model Architecture and Training

3.3.1. Architecture

The model is built upon a Vision Transformer (ViT) backbone, specifically the Small variant with a patch size of 8×8 pixels (ViT-S/8). We use a patch-size-8 visual transformer model trained using the DINO method [26], pre-trained on ImageNet, followed by a 3-layer Multilayer Perceptron (MLP) projection head with Gaussian Error Linear Unit (GELU) activations and a final weight-normalized linear layer that yields 2048-dimensional output embeddings. The backbone's weights were not frozen and were fine-tuned on the marble dataset throughout training.

3.3.2. Two-Stage Training Strategy

Training proceeds in two sequential stages designed to provide a stable representational foundation before the clustering objective is introduced. The transition point at epoch 90 follows the training schedule adopted in the original CA-DINO framework [36], which demonstrated that dedicating approximately 45% of total training epochs to pure self-supervised pretraining provides a sufficiently stable representational foundation for the subsequent cluster-aware phase. The convergence of the DINO loss to a low-variance regime by epoch 90 (Section 4.1) confirms that this schedule is appropriate for the present dataset. All training was performed using the AdamW optimizer [40] with a base learning rate of 5×10^{-4} scaled linearly by batch size following the formula $lr = lr_{base} \times \frac{2 \times batch\_size}{256}$, yielding an effective learning rate of approximately 7.8×10^{-6}, with weight decay of 0.04. A Cosine Annealing scheduler [41] was applied over the full 200 epochs, and the teacher network weights were updated via Exponential Moving Average with momentum following a cosine schedule from 0.996 to 1.0.

Stage 1 — Pure DINO (epochs 1–90). The model is pre-trained using the standard DINO objective [26]. A multi-crop augmentation strategy is used: the teacher receives two global views (80–100% scale), while the student receives eight views: the same two global views plus six additional local crops (20–80% scale), all resized to 224×224 and subject to random horizontal flips (50% probability) and rotations (up to 10°). Following the original DINO formulation [26], the loss is computed over all cross-view pairs in which the teacher and student process different views; same-view pairs are excluded. The teacher receives $N_t = 2$ global views, and the student receives $N_s = 8$ views (2 global + 6 local crops), yielding a total of $(N_t \times N_s) - N_t = 14$ unique cross-view pairs per image. The Stage 1 loss is thus the cross-entropy between the student and teacher output distributions over these cross-view pairs:

$$L_{DINO} = -\frac{1}{N_t}\sum_{i=1}^{N_t}\ \frac{1}{N_s}\sum_{j=1}^{N_s}\ softmax\left(\frac{t_i - c}{\tau_t}\right) \cdot log\ -softmax\left(\frac{s_j}{\tau_s}\right) \quad (1)$$

where t_i and s_j are the teacher and student outputs over N_t and N_s views, with student temperature $\tau_s = 0.1$ and teacher temperature $\tau_t = 0.04$. Training collapse is prevented via a global center c updated by EMA with momentum $m_c = 0.9$:

$$c_{new} = c_{old} \times m_c + \bar{x}_{batch} \times (1 - m_c) \quad (2)$$

Stage 2 — CA-DINO (epochs 91–200). At each epoch, student embeddings are clustered via k-means to generate pseudo-labels. The CA-DINO phase requires a target cluster count k as a hyperparameter; to assess the pipeline's sensitivity to this choice, models were trained with $k \in \{5,8,10,12,15\}$, spanning a range from under- to over-segmentation of the visual space. A Dynamic Loss Gate (DLG) classifies samples as reliable or unreliable by fitting a two-component Gaussian Mixture Model to the per-sample loss distribution. The gate threshold is set to the minimum of the two fitted component means, placing the boundary between the low-loss (reliable) and high-loss (unreliable) sample populations. For reliable samples, a standard cross-entropy loss is computed against the pseudo-labels:

 doi:10.20944/preprints202603.1344.v1

$$L_{ce} = \frac{1}{|\, I_{reliable} \,|} \sum_{i \in I_{reliable}} CrossEntropy(logits_i, p_i) \quad (3)$$

For unreliable samples whose prediction confidence exceeds the gate condition, with $\tau_2 = 0.5$:

$$\left(L_{ce,i} > gate_{threshold}\right) \wedge \left(max_{\,c}(softmax(logits_i)_c) > \tau_2\right) \quad (4)$$

a sharpened target distribution is generated using sharpening temperature $\epsilon_c = 0.1$:

$$q_i = softmax\left(\frac{logits_i}{\epsilon_c}\right) \quad (5)$$

The Label Correction loss is the Kullback-Leibler divergence between this sharpened distribution and the prediction from a new augmented view $p_{aug,i}$:

$$L_{LC} = \frac{1}{|\, I_{unreliable} \,|} \sum_{i \in I_{unreliable}} D_{KL}\left(q_i^{(detached)} \parallel p_{aug,i}\right) \quad (6)$$

$$p_{aug,i} = softmax(logits_{aug,i}) \quad (7)$$

The total loss for Stage 2 combines all three components:

$$L_{total} = L_{DINO} + L_{ce} + L_{LC} \quad (8)$$

3.4. Feature Extraction and Post-Processing Pipeline

3.4.1. Embedding Extraction

After training, the model is set to evaluation mode, disabling stochastic elements such as dropout. Images are passed through the ViT backbone and MLP projection head, yielding a 2048-dimensional embedding vector from the MLP projection head (DINOHead). These high-dimensional embeddings form the input to the subsequent dimensionality reduction and clustering stages.

3.4.2. Dimensionality Reduction

The 2048-dimensional embeddings are susceptible to the curse of dimensionality, which degrades the reliability of distance metrics used in clustering. We apply UMAP, a non-linear technique that preserves local neighborhood structure in the data manifold, to project the embeddings into a lower-dimensional space. UMAP was selected because the visual differences between marble varieties are encoded non-linearly by the Vision Transformer; a linear projection could discard the manifold structure critical for distinguishing fine-grained textural patterns. To identify the optimal configuration, an initial grid search was conducted over the hyperparameter space shown in Table 1. The best-performing configurations were identified using internal clustering metrics (Silhouette Score and Davies-Bouldin Index) evaluated after Ward hierarchical clustering at k=10. Based on this preliminary screening, three representative target dimensionalities were selected for detailed ablation analysis with the remaining hyperparameters fixed at n_neighbors=5, min_dist=0.0, and Euclidean distance, the combination that consistently yielded the highest internal scores across dimensionalities. The full grid search results informed the parameter selection but are not reported individually, as the ablation in Section 4.3 focuses on the effect of target dimensionality, which emerged as the dominant factor.

 doi:10.20944/preprints202603.1344.v1

Table 1. UMAP hyperparameter search space. All combinations were evaluated against qualitative and quantitative clustering criteria.

Parameter	Range
Output Dimensions	[5, 15, 75]
Number of Neighbors	[5, 10, 15, 30, 50]
Minimum Distance	[0.0, 0.1, 0.25, 0.5]
Distance Metric	Euclidean

3.4.3. Hierarchical Clustering

Agglomerative hierarchical clustering is applied to the UMAP-reduced embeddings, producing a dendrogram that reveals nested visual relationships from broad families to fine-grained distinctions [39]. Prior to clustering, the UMAP-reduced embeddings are L2-normalized to ensure that distance computations are scale-invariant. Ward's linkage is applied, as its variance-minimization objective produces balanced, compact clusters at each hierarchical level, a desirable property for an industrial visual taxonomy where interpretability and consistency across hierarchical levels are critical [39]. This choice is empirically validated against three alternative linkage criteria in Section 4.4. Table 2 summarizes the linkage criteria evaluated. All linkage methods were evaluated using Euclidean distance, as required by the Ward variance-minimization criterion.

Table 2. Linkage criteria evaluated for agglomerative hierarchical clustering.

Linkage Criteria	Description
Ward	Minimizes the increase in total within-cluster variance at each merge
Complete	Distance between the two furthest points across clusters
Average	Mean pairwise distance between all points across clusters
Single	Distance between the two closest points across clusters

3.5. Evaluation Framework

3.5.1. Qualitative Evaluation

Qualitative analysis of the generated dendrograms serves as the primary evaluation tool in this study. Two criteria guide the assessment. First, hierarchical coherence: an optimal dendrogram exhibits a general-to-specific organization, where the highest-level splits separate data into broad, visually distinct families and subsequent splits capture progressively finer distinctions. Second, structural integrity: the dendrogram must avoid common artifacts such as chaining, a staircase-like pattern where points merge sequentially rather than in balanced groups, and early singleton formation, where outliers detach prematurely from the main hierarchy. Figure 4 provides a conceptual illustration of an ideal dendrogram structure.

 doi:10.20944/preprints202603.1344.v1

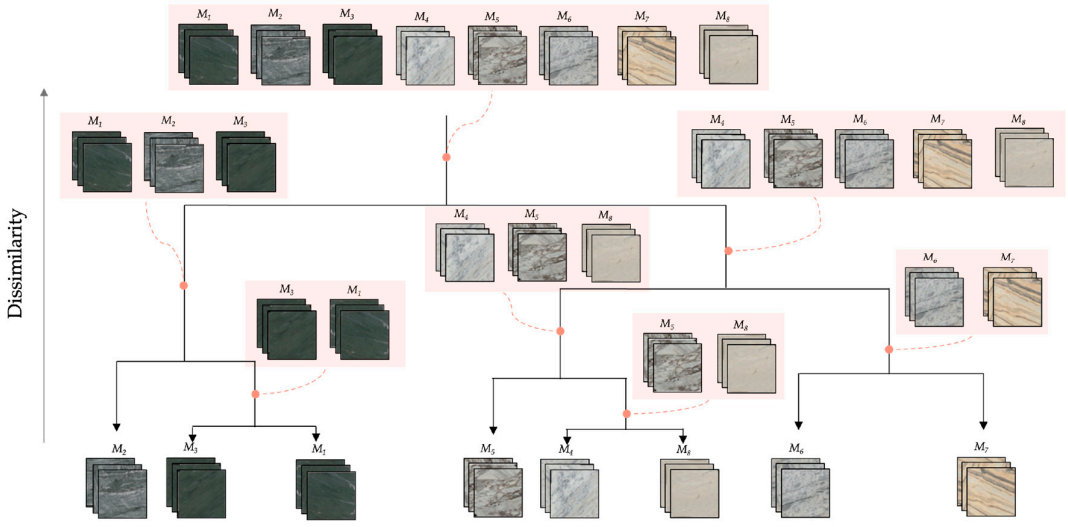

Figure 4. Conceptual illustration of an ideal dendrogram structure, showing clear hierarchical separation and the corresponding general-to-specific visual organization of clusters.

This visual analysis is indispensable because it can identify well-formed clusters containing visually similar marble images from different commercial varieties, a correct outcome for an appearance-based taxonomy that extrinsic metrics would incorrectly penalize.

3.5.2. Quantitative Evaluation

The clustering results are assessed using a suite of standard validation indices organized into three categories. Table 3 summarizes each metric's interpretation, range, and optimal direction.

Table 3. Summary of clustering validation indices used in this study. Internal metrics assess cluster quality from the data structure alone; external metrics compare cluster assignments against commercial variety labels (used for evaluation only, never for training); the hierarchical metric assesses dendrogram fidelity. Standard optimal directions are indicated, with the caveat that lower external metric scores may reflect desirable behaviour in this study (†).

Metric	Abbreviation	Category	Range	Optimal†	What it measures
Silhouette Score	SS	Internal	[−1, +1]	↑ Higher	Mean ratio of intra-cluster cohesion to inter-cluster separation for each sample
Davies-Bouldin Index	DB	Internal	[0, +∞)	↓ Lower	Average similarity between each cluster and its most similar neighbor; penalizes loose, overlapping clusters
Calinski-Harabasz Index	CH	Internal	[0, +∞)	↑ Higher	Ratio of between-cluster dispersion to within-cluster dispersion; favors compact, well-separated clusters
Adjusted Rand Index	ARI	External	[−1, +1]	↑ Higher†	Chance-corrected agreement between predicted cluster assignments and commercial variety labels
Normalized Mutual Information	NMI	External	[0, +1]	↑ Higher†	Mutual dependence between predicted cluster assignments and commercial variety labels, normalized by entropy
V-measure	V	External	[0, +1]	↑ Higher†	Harmonic mean of homogeneity (each cluster contains only one label) and completeness (all samples of a label are in one cluster)

 doi:10.20944/preprints202603.1344.v1

Cophenetic Correlation Coefficient	CCC	Hierarchical	[−1, +1]	↑ Higher	Pearson correlation between pairwise distances in the original feature space and the distances implied by the dendrogram structure

- **Internal measures.** Silhouette Score (SS) [42], Davies-Bouldin Index (DB) [43], and Calinski-Harabasz Index (CH) [44], assess cluster compactness and separation based solely on the data's inherent structure, without reference to external labels.
- **External measures.** Adjusted Rand Index (ARI) [45], Normalized Mutual Information (NMI) [46], and V-measure [47] compare cluster assignments to the ground-truth commercial labels.
- **Hierarchical measures.** The Cophenetic Correlation Coefficient (CCC) [48] evaluates how faithfully the dendrogram structure represents the original pairwise distances in the feature space.

3.5.3. Metric Interpretation and Evaluation Strategy

Quantitative validation indices serve two complementary roles in this study. Internal measures (SS, DB, CH) provide an objective assessment of cluster compactness and separation in the learned embedding space, independently of any labeling convention. External measures (ARI, NMI, V-measure) quantify the degree of correspondence between the discovered groupings and the existing commercial classification, a relationship that is itself informative: high alignment would suggest the model recovers the commercial taxonomy; low alignment, combined with high internal scores and coherent dendrograms, suggests the model has identified a more perceptually consistent organization than the commercial labels impose.

One important caveat applies to the interpretation of external metrics. Because commercial labels were assigned by market convention rather than systematic visual criteria, penalization for deviating from them is not equivalent to penalization for clustering error. For example, grouping Branco Peletigre and Dante Grey together based on their shared grey base and veining pattern constitutes a visually correct decision that all extrinsic metrics would score as incorrect. External metrics are therefore reported as a descriptive reference to characterize the relationship between the learned visual structure and the existing commercial taxonomy, rather than as a primary measure of pipeline quality. Qualitative dendrogram analysis and internal indices serve as the primary evaluation criteria.

4. Results

This section presents the experimental results of the proposed CA-DINO pipeline for unsupervised visual grouping of marble varieties. Section 4.1 reports the training dynamics of the selected model. Section 4.2 presents the multi-criterion convergence analysis that justifies the checkpoint selection at epoch 200. Sections 4.3 and 4.4 present the ablation studies on dimensionality reduction and linkage method, respectively. Section 4.5 reports the main quantitative clustering results across all k configurations, and Section 4.6 presents the qualitative analysis of cluster structure at the selected k=10.

4.1. Training Dynamics

The training of CA-DINO proceeds in two stages: a pure self-supervised DINO pretraining phase (epochs 1–90) and a cluster-aware fine-tuning phase (epochs 91–200). The evolution of both loss components and the Dynamic Loss Gate threshold across training is shown in Figure 5.

 doi:10.20944/preprints202603.1344.v1

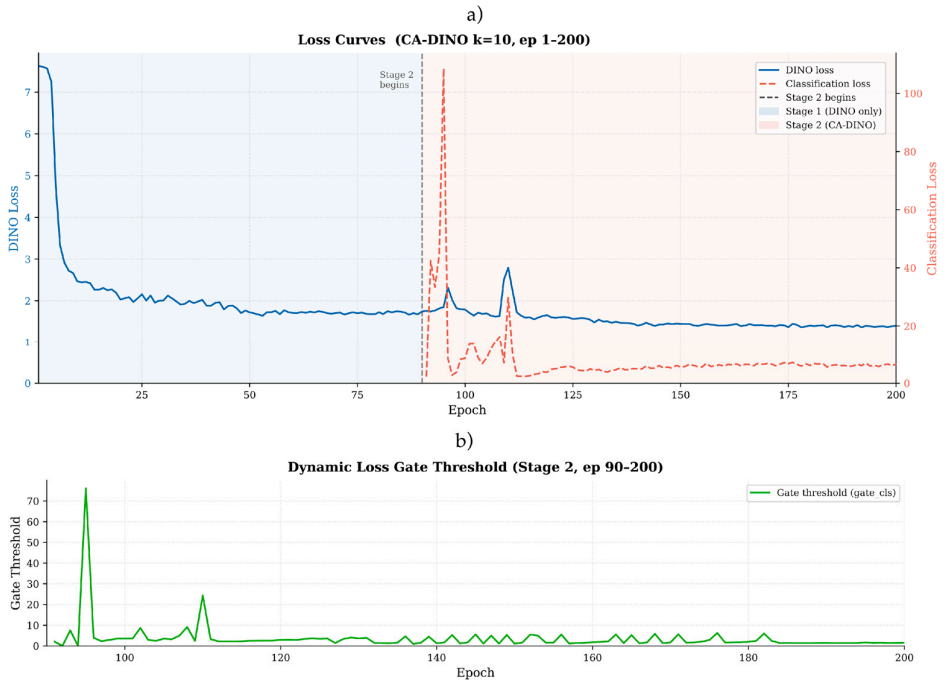

Figure 5. Training dynamics over 200 epochs. a) DINO self-supervised loss (solid line, left axis) and classification loss (dashed line, right axis) across both training stages; the vertical dashed line at epoch 90 marks the transition into Stage 2, and shaded regions highlight each stage. b) Evolution of the pseudo-label acceptance threshold (the minimum confidence required to include a sample in the classification loss) throughout Stage 2.

During Stage 1, the DINO loss decreased from an initial value of 7.623 at epoch 1 to 1.743 at epoch 90, representing a total reduction of 77.1%. The descent was non-monotonic, exhibiting 38 upward perturbations consistent with the stochastic gradient dynamics of self-supervised ViT training with multi-crop augmentation, but converged to a stable regime by the end of Stage 1 with a windowed mean of 1.744 ± 0.055 at epoch 90. At the onset of Stage 2 (epoch 91), the classification loss activated as expected, exhibiting high initial variance, peaking at 108.524 at epoch 95, followed by two secondary spikes at epochs 96 (2.318) and 109–110 (2.794). These early perturbations are consistent with the known instability of pseudo-label assignment during the first cluster centroid updates, when the GMM-based Dynamic Loss Gate has not yet converged. Importantly, the DINO loss recovered below its Stage 1 minimum by epoch 192 (1.365), confirming that the self-supervised objective continued to improve throughout Stage 2 despite the added cluster-alignment pressure. The classification loss stabilised to a low-variance regime by epoch 150, with windowed means of 5.771 ± 0.398 at epoch 150 and 6.222 ± 0.290 at epoch 200, indicating that pseudo-label quality and cluster assignments had converged well before the end of training.

The Dynamic Loss Gate threshold (gate_cls) reached a maximum of 76.154 in the early epochs of Stage 2 before decaying to a final value of 1.560 by epoch 200. This trajectory confirms that the proportion of samples classified as unreliable pseudo-labels decreased substantially over the course of Stage 2, as the cluster centroids stabilised and the classifier became progressively more confident in its assignments.

4.2. Epoch Convergence Analysis

The selection of epoch 200 as the evaluation checkpoint rests on three independent and complementary convergence criteria, grounded in the analysis of self-attention maps across checkpoints at epochs 90, 100, 200, 300, and 400, conducted on the CA-DINO k=10 training run extended to 400 epochs. Aggregated SSIM and cosine similarity (CS) between consecutive epoch-pair attention maps were computed over the full test set of 49 images. The convergence curves are shown

 doi:10.20944/preprints202603.1344.v1

in Figure 6. Figure 7 shows the evolution of self-attention maps across all five checkpoints for three purposefully selected test images, each illustrating a distinct convergence behaviour: one image whose attended regions stabilise rapidly after epoch 100 (SSIM = 0.899 at the 100→200 transition), one that undergoes progressive sharpening in late training (SSIM dropping to 0.603 at 300→400), and one exhibiting an atypical pattern of early convergence followed by a transient disruption at the 200→300 transition before restabilising.

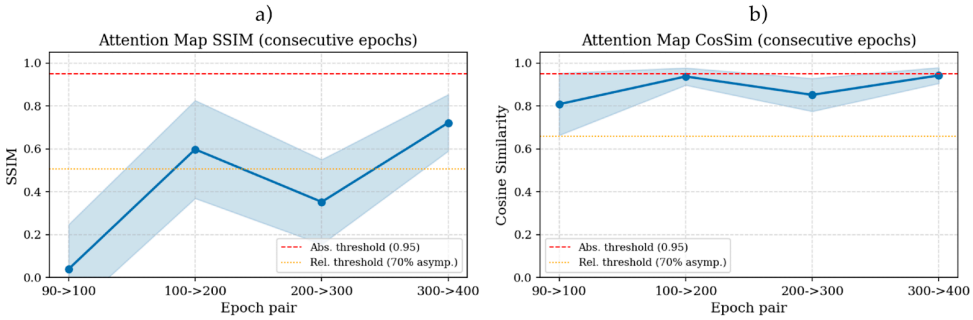

Figure 6. Convergence of attention map representations across training epochs, measured as the mean Structural Similarity Index (SSIM; panel a) and cosine similarity (panel b) between attention maps of consecutive epoch pairs, averaged over 49 test images. Shaded bands indicate one standard deviation and dashed lines mark the convergence reference thresholds.

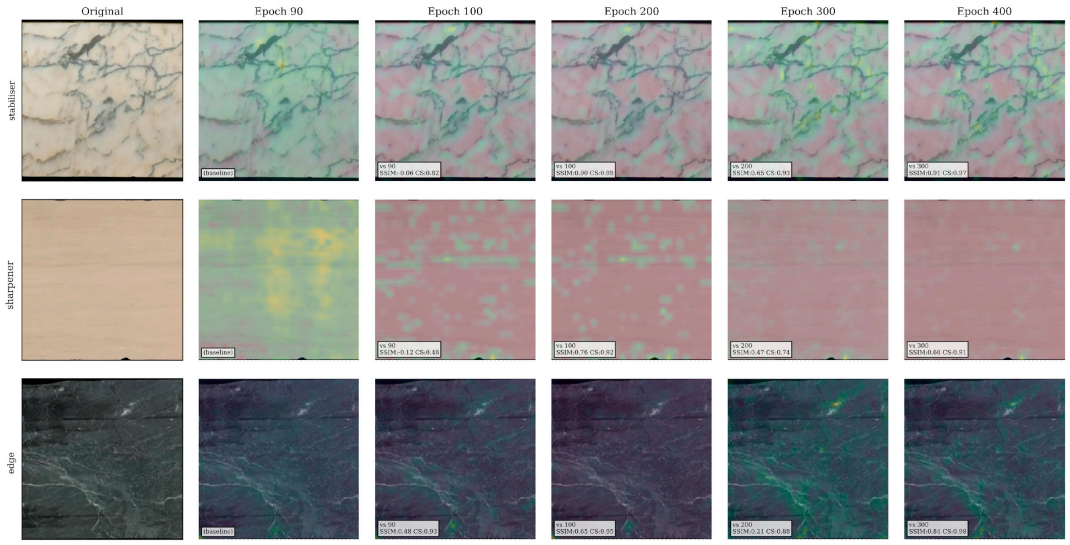

Figure 7. Self-attention map evolution for three representative test images illustrating contrasting convergence behaviours. Each row shows the original image followed by attention maps at epochs 90, 100, 200, 300, and 400, with per-cell SSIM and cosine similarity reported relative to the previous checkpoint. The top row (stabiliser) converges rapidly after epoch 100 (SSIM = 0.899 at 100→200). The middle row (sharpener) continues to evolve during late training, with SSIM dropping to 0.603 at 300→400, indicating progressive overspecialisation toward localised surface features. The bottom row (edge case) shows a transient disruption at 200→300 (SSIM = 0.211) before restabilising.

Primary criterion — directional convergence of embeddings. Since downstream clustering operates via cosine distance in the embedding space, cosine similarity between consecutive checkpoint attention maps is the most operationally relevant convergence signal. At the 90→100 transition, CS was 0.8079 ± 0.1449, representing 85.8% of the asymptotic value observed at epoch 400 (CS=0.9420), confirming that the early Stage 2 period still involved substantial directional

 doi:10.20944/preprints202603.1344.v1

reorganisation of attended features. By the 100→200 transition, CS reached 0.9370 ± 0.0401, corresponding to 99.5% of the asymptotic value, effectively indicating that the embedding directions governing cluster assignments, UMAP projections, and centroid distance calculations had fully stabilised by epoch 200. The remaining 0.5% gain to the final state represents a marginal increment that provides no meaningful improvement to downstream clustering while introducing increasing overfitting risk.

Secondary criterion — non-monotonic SSIM trajectory as a spatial stability indicator. SSIM between consecutive attention maps was 0.0389 ± 0.2081 at the 90→100 transition, reflecting rapid and highly variable spatial reorganisation of attended regions in the early Stage 2 period, with per-image values ranging from −0.36 to +0.82. By the 100→200 transition, SSIM rose substantially to 0.5973 ± 0.2283, indicating that the spatial structure of attention maps had broadly stabilised by epoch 200. Beyond this point, SSIM exhibited a characteristic non-monotonic pattern: it declined to 0.3522 ± 0.1974 at the 200→300 transition before recovering to 0.7214 ± 0.1318 at 300→400. This decline and subsequent recovery is inconsistent with continued representational improvement and instead reflects a late-stage spatial sharpening process driven by EMA teacher drift, whereby the model progressively concentrates attention onto increasingly localised subregions of already-attended areas. Critically, this sharpening is accompanied by a risk of spatial overspecialisation: at 200→300, the rate of SSIM change turned negative (−0.002451 per epoch), and the subsequent recovery at 300→400 reached only 82.8% of the asymptotic SSIM value at epoch 200 normalised against the terminal state, confirming that late-stage training does not improve the broad spatial representational structure established by epoch 200.

Contextual criterion — SSL generalisation on constrained datasets. Consistent with established behaviour in self-supervised ViT training [26], the optimal downstream evaluation checkpoint is not necessarily the terminal one. In the DINO framework, k-NN accuracy on downstream tasks does not increase monotonically with the number of epochs, and intermediate checkpoints have been shown to generalise better when the downstream dataset is small or domain-shifted relative to the pretraining data. In the present setting, a constrained natural stone dataset of 94 test samples, continuing beyond epoch 200, risks the model's self-attention heads progressively encoding training-set-specific spatial patterns that reduce embedding transferability to the full test population, as evidenced by the attention map grid, which shows cases of late-epoch fixation on surface cracks and localised edge artifacts in feature-poor samples. Based on these three criteria jointly, epoch 200 was selected as the evaluation checkpoint, providing effective asymptotic embedding directions, a confirmed spatial stability plateau, and conservative generalisation capacity appropriate for the dataset scale.

4.3. Dimensionality Reduction Ablation

To evaluate the effect of dimensionality reduction on downstream clustering quality, four preprocessing configurations were compared: the raw 2048-dimensional CA-DINO embeddings, and UMAP projections to 5, 15, and 75 dimensions (n_neighbors=5, min_dist=0.0, random_state=42), all applied to the test-set embeddings at epoch 200 (n=94), followed by L2 normalisation and Ward hierarchical clustering at k=10. Quantitative results are reported in Table 4. Raw embeddings (2048D) achieve the highest label-alignment scores (ARI = 0.593, NMI = 0.730) and near-perfect cophenetic correlation (CCC = 0.999), while UMAP-reduced configurations consistently yield better intra-cluster compactness, with Silhouette Scores above 0.786 and Davies-Bouldin indices below 0.255. Among the reduced configurations, UMAP 75D offers the best compromise between compactness and label alignment (ARI = 0.433, NMI = 0.674). The 2D UMAP projection of cluster assignments across all configurations is shown in Figure 8, where colour encodes cluster identity and marker shape encodes commercial variety.

 doi:10.20944/preprints202603.1344.v1

Table 4. Dimensionality Reduction Ablation: Clustering metrics at k=10 (Ward linkage, L2 normalisation, epoch 200, n=94). Best value per metric in bold.

Configuration	SS ↑	DB ↓	CH ↑	ARI ↑	NMI ↑	V-measure ↑	CCC ↑
Raw (2048D)	0.489	0.768	**39887.7**	**0.593**	**0.73**	**0.73**	**0.999**
UMAP 5D	0.786	0.254	6478.2	0.423	0.672	0.672	0.939
UMAP 15D	**0.803**	**0.226**	4808.8	0.397	0.656	0.656	0.974
UMAP 75D	0.797	0.255	8594.5	0.433	0.674	0.674	0.932

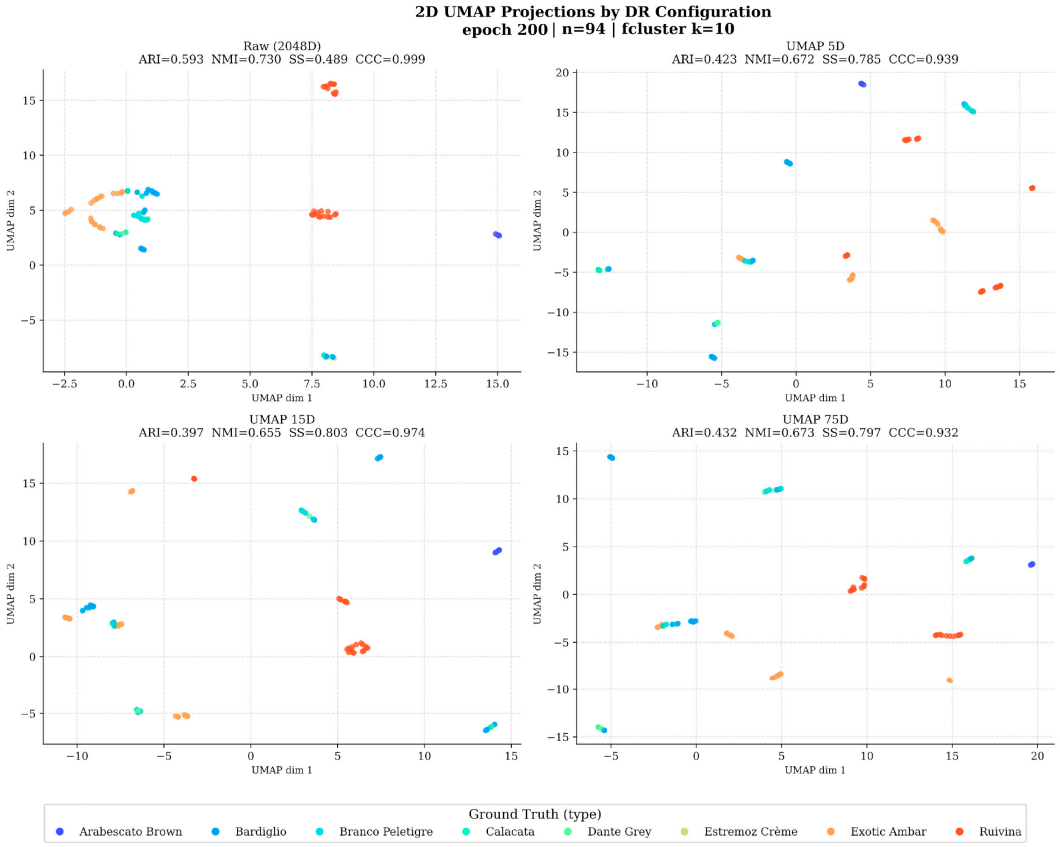

Figure 8. UMAP 2D projections of embedding spaces for all four dimensionality reduction configurations (Raw, UMAP 5D, 15D, and 75D), with cluster assignments indicated by colour and commercial variety by marker shape. Alignment between colour and shape reflects agreement between learned clusters and ground-truth variety labels.

The raw 2048D embeddings achieved the highest label-alignment scores (ARI=0.593, NMI=0.730) and the highest Cophenetic Correlation Coefficient (CCC=0.999), yet simultaneously exhibited the worst cluster geometry of all four configurations, with a Silhouette score of only 0.489 and a Davies-Bouldin index of 0.768. This apparent contradiction is a direct consequence of the extreme feature-to-sample ratio of the raw space: with n=94 samples and d=2048 dimensions, the dataset operates in a regime where approximately 22 features are available per sample, well within the range where the curse of dimensionality produces inflated inter-cluster distances from noise-dominated variance rather than from genuine cluster compactness. In this regime, Ward linkage can spuriously recover label-correlated structure because the high-dimensional noise amplifies the separation between groups sharing common low-frequency visual patterns, making ARI and NMI unreliable selection criteria. The Calinski-Harabasz score of 39,888 for raw embeddings, four to eight times larger than any UMAP variant, further reflects this artefact, as CH grows with ambient dimensionality

 doi:10.20944/preprints202603.1344.v1

independently of actual cluster quality. The near-perfect CCC of 0.999 is similarly diagnostic of trivially separable geometry in high-dimensional space rather than a meaningful property of the cluster hierarchy. Raw 2048D embeddings were therefore excluded from consideration as the operating configuration despite their label-alignment advantage.

UMAP compression consistently resolved this pathology: all three reduced configurations raised the Silhouette score above 0.786 and reduced the Davies-Bouldin index below 0.255, confirming that geometrically compact and well-separated clusters emerge only after dimensionality reduction collapses the noise-dominated axes of the raw embedding space. Among the three UMAP variants, the choice of target dimensionality involved a clearly identifiable trade-off between geometric quality and label-discriminative structure. UMAP 15D achieved the best Silhouette (0.803) and Davies-Bouldin (0.226), but at the cost of the lowest ARI (0.397) and NMI (0.656) of all UMAP configurations, indicating that compression to 15 dimensions discards a portion of the label-discriminative variance present in the original embedding space, a symptom of over-compression relative to the intrinsic dimensionality of the eight-class marble representation. UMAP 75D retained marginally higher label alignment (ARI=0.433, NMI=0.674) but produced the lowest CCC of all UMAP variants (0.932), reflecting reduced hierarchical distance preservation and a less stable dendrogram structure for downstream clustering. UMAP 5D achieved the best balance across the three relevant criteria: the highest ARI and NMI among all UMAP variants after excluding the raw embedding artefact, strong geometric quality (SS=0.786, DB=0.254), and the best dendrogram stability among all configurations except the inflated raw baseline (CCC=0.939). The 0.010-point ARI advantage of UMAP 75D over UMAP 5D does not constitute a meaningful margin given the sample size of n=94, and is offset by the inferior dendrogram stability of UMAP 75D. UMAP 5D was therefore selected as the operating dimensionality reduction configuration for this pipeline, representing the Pareto-optimal point among valid configurations with respect to the joint criteria of label alignment, cluster geometry, dendrogram stability, and computational efficiency in downstream hierarchical clustering.

4.4. Linkage Method Ablation

To identify the optimal agglomerative linkage criterion, Ward, Complete, Average, and Single linkage were compared on the UMAP 5D L2-normalised embeddings at epoch 200, evaluating both internal clustering quality and alignment with ground truth labels at k=10 under controlled and equal preprocessing conditions. Dendrogram structures for all four methods are shown in Figure 9. Metric profiles across k=2–20 are shown in Figure 10. Quantitative results are reported in Table 5.

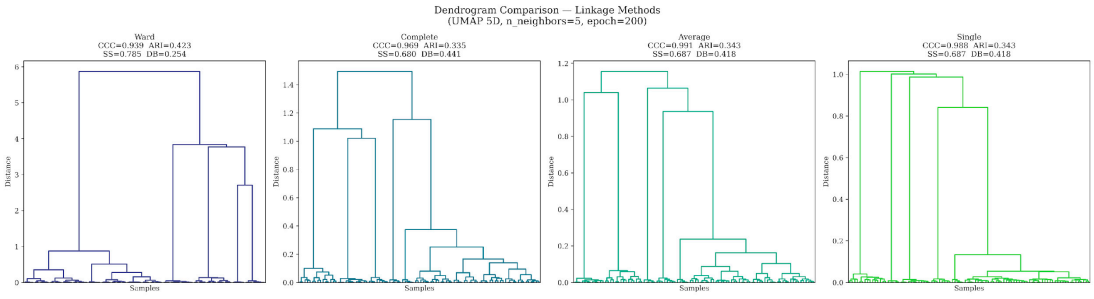

Figure 9. Dendrogram structures for Ward, Complete, Average, and Single linkage applied to UMAP 5D L2-normalised embeddings at epoch 200. Each panel is annotated with cophenetic correlation coefficient (CCC), adjusted Rand index (ARI), Silhouette score (SS), and Davies-Bouldin index (DB) at k=10.

 doi:10.20944/preprints202603.1344.v1

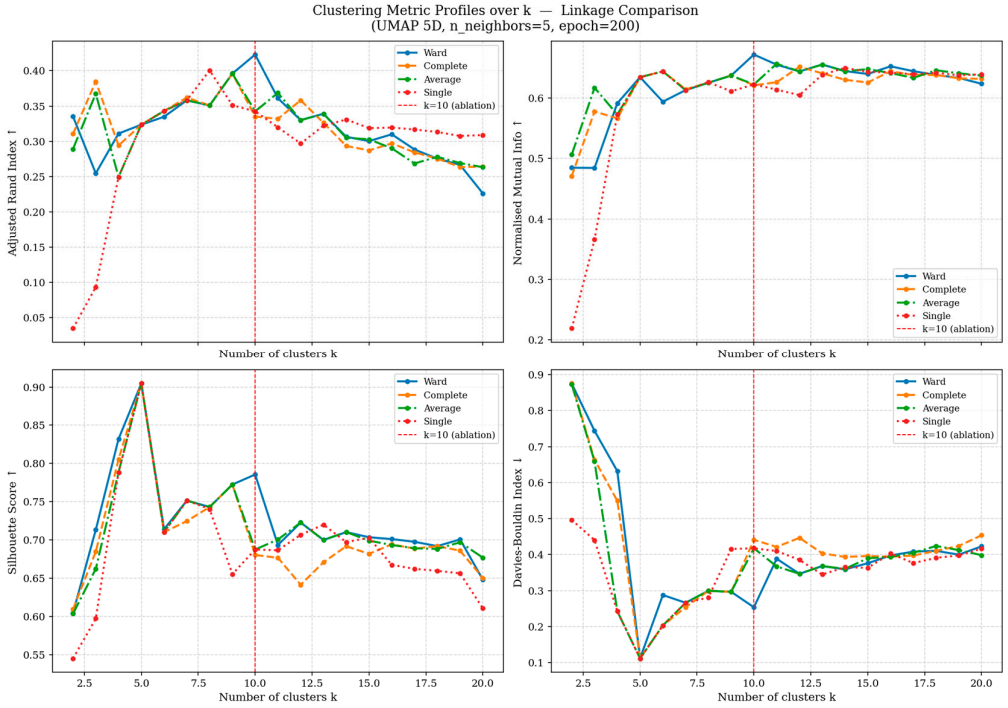

Figure 10. Clustering metric profiles across k=2–20 for all four linkage methods, showing ARI, NMI, Silhouette score, and Davies-Bouldin index. The vertical dashed line marks k=10. ARI peaks at k=8–10 across methods, consistent with the eight ground-truth commercial varieties, while Silhouette scores peak at k=2–3 and decline monotonically.

Table 5. Linkage criterion ablation: clustering metrics at k=10 (UMAP 5D, L2 normalisation, epoch 200, n=94). Best value per metric in bold. ↑ higher is better; ↓ lower is better.

Linkage	SS ↑	DB ↓	CH ↑	ARI ↑	NMI ↑	V-measure ↑	CCC ↑
Ward	**0.786**	**0.254**	**6478.200**	**0.423**	**0.672**	**0.672**	0.939
Complete	0.680	0.441	6149.300	0.335	0.621	0.621	0.969
Average	0.687	0.418	6074.600	0.343	0.623	0.623	**0.991**
Single	0.687	0.418	6074.600	0.343	0.623	0.623	0.988

Ward linkage achieved the best performance across all metrics, with the exception of CCC, attaining the highest Silhouette score (SS=0.7855), the lowest Davies-Bouldin index (DB=0.2536), the highest Calinski-Harabasz score (CH=6478.2), the highest ARI (0.4229), and the highest NMI and V-measure (both 0.6719). This result confirms that Ward's variance-minimisation objective is the most appropriate merging criterion for this embedding space, yielding clusters that are both geometrically compact and well-aligned with the ground-truth commercial labels. The margin over the next-best method is substantial: Ward's ARI exceeds Complete linkage by 0.088 and Average linkage by 0.080, indicating a clear rather than marginal advantage.

Average and Single linkage produced near-identical results across all reported metrics at k=10 (ARI=0.3425 vs. 0.3425, NMI=0.6225 vs. 0.6225, SS=0.6874 vs. 0.6874, DB=0.4176 vs. 0.4176), with CCC values of 0.9911 and 0.9879, respectively. Inspection of the cluster assignments confirmed that the two methods yield highly similar but not strictly identical partitions; the metric equivalence at this decimal precision reflects the fact that the few reassigned samples lie near cluster boundaries and contribute negligibly to aggregate scores. This convergence is consistent with the geometric properties of the L2-normalised UMAP 5D space, in which the compact, well-separated cluster structure reduces the practical difference between minimum-distance and average-distance merging

 doi:10.20944/preprints202603.1344.v1

strategies at the k=10 cut level. Complete linkage achieved intermediate internal quality (SS=0.6803, DB=0.4409) and the second-highest external metrics (ARI=0.3350, NMI=0.6211), with the highest CCC among the non-Ward methods (0.9692). Single linkage did not exhibit the chaining pathology commonly observed in continuous embedding spaces at this scale, likely because the L2-normalised UMAP 5D space is sufficiently compact to prevent the formation of elongated chain-like structures. The metric profiles across k=2–20 show that all four methods reach their Silhouette peak at k=2–3, then decline monotonically, whereas ARI peaks at k=8–10, consistent with the underlying ground-truth structure of 8 commercial varieties. Based on this analysis, Ward linkage was confirmed as the pipeline default and applied to all configurations reported in Section 4.5.

4.5. Clustering Performance Across k Values

Table 6 reports the full set of clustering metrics for the Pure DINO baseline and all five CA-DINO configurations across $k \in \{5, 8, 10, 12, 15\}$, evaluated on the held-out test set of 94 samples. All CA-DINO results are reported from the training run that produced the most coherent training dynamics across all k values.

Table 6. Clustering Metrics: Pure DINO baseline vs. CA-DINO across k values (epoch 200, UMAP 5D, Ward linkage). Best CA-DINO value per metric in bold. The arrow indicates the direction of improvement.

Model	k	SS ↑	DB ↓	CH ↑	ARI ↑	NMI ↑	V-measure ↑	CCC ↑
Pure DINO	—	0.761	0.364	1602.7	**0.438**	0.552	0.552	0.88
CA-DINO	5	0.614	0.532	598.4	0.244	0.399	0.399	0.917
CA-DINO	8	0.625	0.495	465.7	0.181	0.366	0.366	0.809
CA-DINO	10	**0.778**	**0.293**	5957.4	0.305	**0.58**	**0.58**	0.903
CA-DINO	12	0.656	0.461	1320	0.134	0.341	0.341	0.831
CA-DINO	15	0.684	0.381	**10467.4**	0.114	0.402	0.402	**0.958**

CA-DINO at k=10 achieved the strongest overall performance across all evaluated configurations. Notably, it is the only CA-DINO variant to surpass the Pure DINO baseline on multiple metrics simultaneously: Silhouette score (0.778 vs. 0.761), Davies-Bouldin index (0.293 vs. 0.364), NMI (0.580 vs. 0.552), and V-measure (0.580 vs. 0.552). This demonstrates that when the number of training clusters is correctly matched to the intrinsic visual structure of the dataset, the cluster-aware objective produces a representation space that is simultaneously more geometrically coherent and better aligned with expert-defined labels than the unconstrained self-supervised baseline. The Pure DINO baseline retained the highest ARI (0.438 vs. 0.305 for CA-DINO k=10), reflecting that without cluster supervision, the model distributes representational capacity uniformly across the embedding space and produces a partition at k=8 that closely mirrors the ground truth commercial label count. CA-DINO at k=10 accepts a reduction in ARI in exchange for a finer-grained partition that captures intra-variety visual subpopulations, a partition structurally distinct from, and complementary to, the expert labelling scheme, as analysed in Section 4.6

The performance of CA-DINO configurations at k≠10 was consistently inferior across both internal and external metrics. k=8 and k=12 exhibited the weakest geometric cohesion (DB=0.495 and DB=0.461 respectively), while k=5 and k=15 showed intermediate internal quality but substantially lower external validity (ARI of 0.244 and 0.114). The anomalously high Calinski-Harabasz value at k=15 (10467.4) is an artefact of the CH index's known sensitivity to the ratio of between-cluster to within-cluster variance at high k in compact, well-separated spaces, and should not be interpreted as evidence of superior clustering quality at that configuration [49].

 doi:10.20944/preprints202603.1344.v1

4.6. Emergent Visual Grouping Structure

The qualitative analysis examines how the CA-DINO pipeline organises the 94 test samples across the visual similarity space learned during training. The Ward dendrogram over the UMAP 5D embeddings encodes a continuous multi-scale structure that can be interrogated at any level of granularity, and this section traces that structure across six levels of resolution, from the root binary split down to the ten-group reference partition, identifying the qualitative patterns that emerge at each scale. Three recurring phenomena are documented: cross-category merging of commercially distinct varieties that share visual properties, intra-category splitting of commercially unified varieties that harbour visual sub-populations, and coherent pure family formation where commercial and visual boundaries genuinely coincide. The full level-by-level decomposition is presented in Figure 12, the cluster composition heatmap at the k=10 reference cut is shown in Figure 11, and the UMAP 2D projection with cluster assignments overlaid on commercial variety markers is shown in Figure 13.

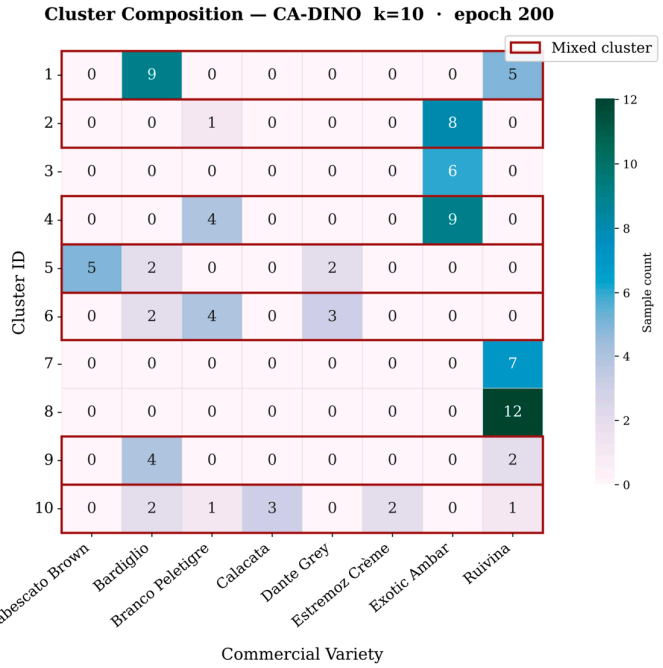

Figure 11. Cluster composition heatmap at the k=10 reference cut (Ward linkage, UMAP 5D, L2 normalisation, epoch 200). Rows correspond to cluster IDs; columns correspond to commercial varieties; cell values report sample counts. Red borders indicate mixed clusters containing samples from more than one commercial variety.

 doi:10.20944/preprints202603.1344.v1

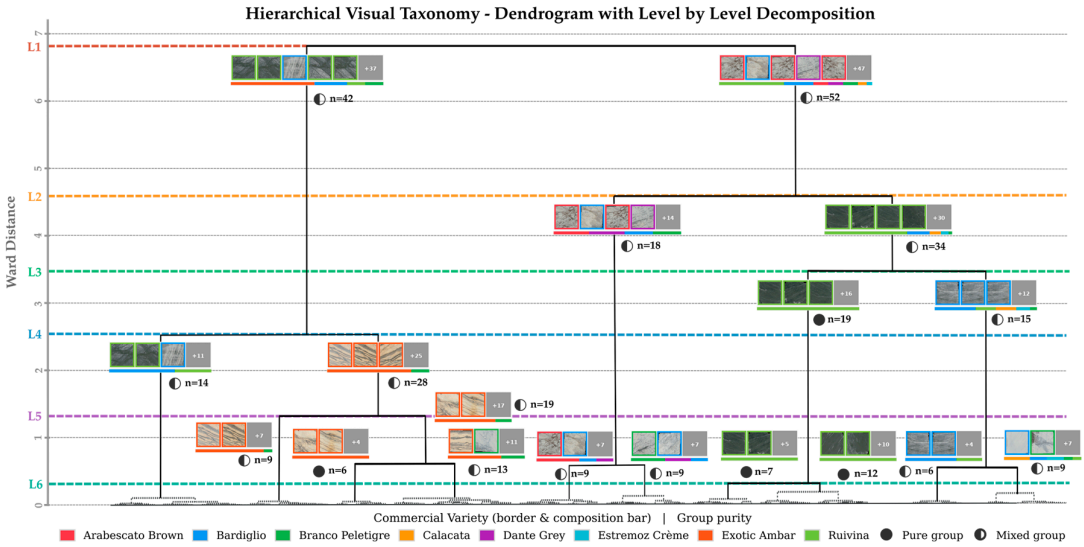

Figure 12. Level-by-level hierarchical decomposition of the CA-DINO k=10 embedding space (epoch 200, n=94). Left: full Ward dendrogram with six coloured dashed cut lines marking Levels 1–6 at decreasing cut heights. Right: image panel strips at each level shown in dendrogram leaf order; group width is proportional to member count; solid-bordered panels indicate pure groups (single commercial variety); dashed-bordered panels indicate mixed groups; the coloured bar beneath each thumbnail strip shows the proportional variety composition of that group.

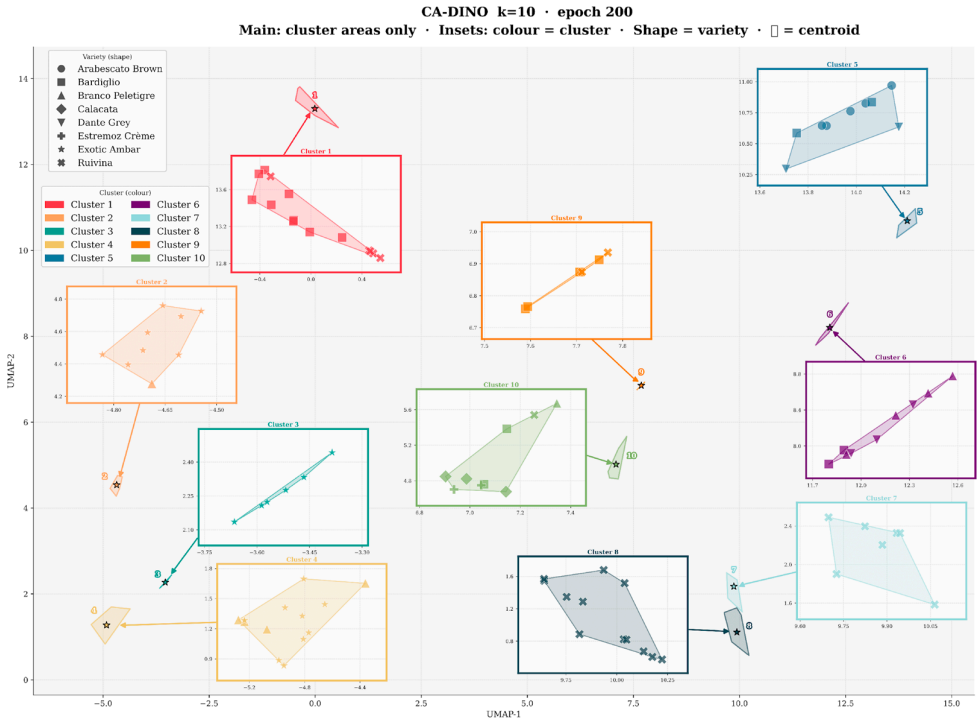

Figure 13. UMAP 2D scatter plot of CA-DINO test-set embeddings (epoch 200). Point colour encodes cluster ID at the k=10 Ward cut; point shape encodes commercial variety. Markers where shape and colour do not align within a visually coherent region indicate cross-category groupings — specimens from distinct commercial varieties placed in the same cluster on the basis of visual similarity.

 doi:10.20944/preprints202603.1344.v1

4.6.1. Level-by-Level Structure

The root binary split at Level 1 (Ward cut height 6.804), visible as the leftmost dashed line in Figure 12, partitions the 94 test samples into branches of 42 and 52 samples. The left branch is dominated by Exotic Ambar (23/42, 55%), with minority contributions from Bardiglio (9), Ruivina (5), and Branco Peletigre (5). The right branch contains the remaining 52 samples, led by Ruivina (22) and Bardiglio (10), along with all samples of Arabescato Brown (5), Dante Grey (5), Branco Peletigre (5), Calacata (3), and Estremoz Creme (2). This primary split broadly separates the warm-toned, high-saturation stones from the cool-toned, grey-veined family, the most salient chromatic and textural distinction in the dataset, confirming that the top-level organisation of the model's embedding space is perceptually grounded.

At Level 2 (cut height 4.619), visible in the Level 2 panel strip of Figure 12, the right branch subdivides to isolate an 18-sample group composed of Arabescato Brown (5), Dante Grey (5), Bardiglio (4), and Branco Peletigre (4), four commercially distinct varieties grouped together on the basis of their shared grey-on-grey veining pattern. This is one of the most commercially significant findings in the hierarchy: the model identifies a visual family that does not correspond to any commercial boundary, revealing a set of stones that are visually indistinguishable at the texture level despite carrying different market designations. At Level 3 (cut height 3.462), the first pure group emerges: 19 Ruivina samples consolidate into a coherent single-variety subtree, as shown in the Level 3 panel strip of Figure 12, establishing that Ruivina's characteristic warm-toned ground and distinctive reddish veining constitute a compact and recoverable visual signature at an intermediate scale of the hierarchy. Level 4 (cut height 2.471) further resolves the Exotic Ambar branch, consolidating 28 samples in which Exotic Ambar represents 82% of membership, with the remaining 18% consisting of Branco Peletigre specimens visually proximate to the amber-toned range. By Level 5 (cut height 1.250), Exotic Ambar begins to internally subdivide, and by Level 6 (cut height 0.295, the k=10 reference cut), the ten-group partition is reached with 3 pure and 7 mixed groups, as detailed in the heatmap of Figure 11.

Variety-level tracking across the six levels of Figure 12 reveals highly differentiated behaviour. Arabescato Brown, Calacata, and Estremoz Creme each appear in a single group throughout all six levels, confirming that these varieties possess compact and distinctive visual signatures consistently recoverable as coherent units at any resolution. Exotic Ambar remains unified from Level 1 through Level 4 before splitting into two sub-groups at Level 5 and three at Level 6, reflecting the structured natural variation in background saturation across this variety's quarry range. Ruivina emerges as a pure 19-sample group at Level 3 but fragments across five groups at Level 6, indicating that the model identified multiple visually distinct sub-populations within this single commercial label. Bardiglio is the most fragmented variety, appearing across five groups at Level 6, consistent with its broad and heterogeneous visual range, also reflected in the scatter of Figure 13, where Bardiglio markers appear across multiple distinct colour regions.

4.6.2. Three Illustrative Cases

The three cases shown in Figure 14 extract specific sub-trees from the full dendrogram to illustrate each of the three phenomena identified above.

 doi:10.20944/preprints202603.1344.v1

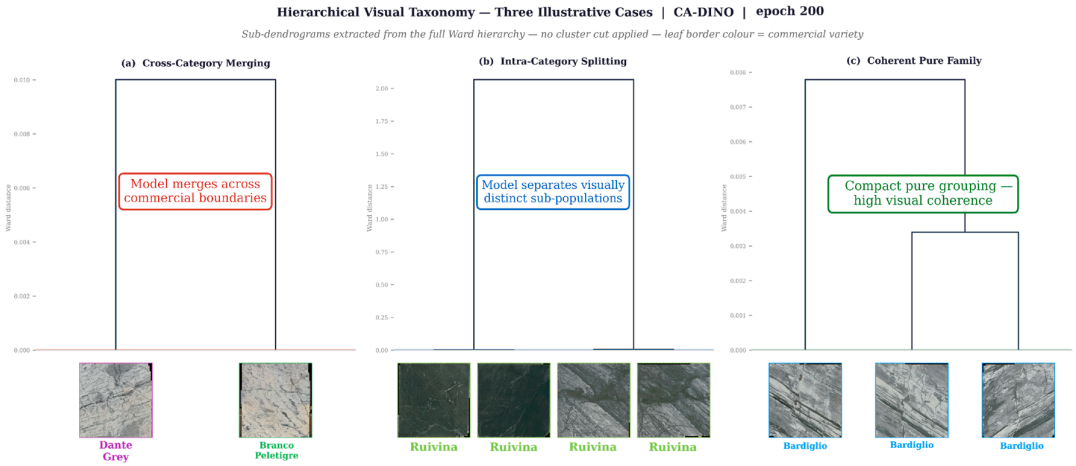

Figure 14. Three illustrative cases extracted from the Ward dendrogram. Panel (a): Cross-category merging — a Branco Peletigre specimen and a Dante Grey specimen joined at Ward distance 0.0100, the smallest inter-variety merge height in the hierarchy, illustrating that the model places commercially distinct stones in immediate proximity when their visual signatures are indistinguishable. Panel (b): Intra-category splitting — two visually distinct Ruivina sub-populations with intra-group distances of 0.0033 and 0.0057, separated by an inter-group distance of 1.4621 (inter-to-intra ratio of 257), demonstrating that a single commercial variety can harbour multiple recoverable visual sub-populations. Panel (c): Coherent pure family — a 3-sample Bardiglio subtree at intra-family distance 0.0057, corresponding to 0.007× the dataset mean pairwise distance, demonstrating that where commercial and visual boundaries align, the model produces tight, stable, and interpretable groupings. Each panel shows the relevant sub-dendrogram above variety-labelled colour-bordered thumbnails in dendrogram leaf order; no cluster cut line is imposed.

Cross-category merging (Panel a of Figure 14). The most precise instance of cross-category merging involves a Branco Peletigre specimen and a Dante Grey specimen joined at Ward distance 0.0100, the smallest inter-variety merge height in the entire hierarchy. These two samples, carrying distinct commercial designations, are the closest pair in the full embedding space. This is not an isolated occurrence: at Level 2 of Figure 12, an 18-sample group consolidates Arabescato Brown, Dante Grey, Bardiglio, and Branco Peletigre (22%, 28%, 22%, 28% respectively), confirming a persistent and robust cross-category affinity among grey-veined stones across multiple levels of the hierarchy. The result is commercially relevant: it identifies pairs and families of stones that are visually indistinguishable at the texture level despite carrying different market prices and origin labels, a discrimination that is structurally invisible to supervised classifiers trained on commercial categories.

Intra-category splitting (Panel b of Figure 14). Within the commercial category of Ruivina, the hierarchy identifies two sub-populations with intra-group distances of 0.0033 and 0.0057, respectively, separated by an inter-group distance of 1.4621, yielding an inter-to-intra ratio of 257. These two sub-populations only rejoin in the dendrogram at Ward height 2.068, well above their individual merge heights, confirming that the separation is genuine and not an artefact of local neighbourhood structure. The finding directly challenges the implicit assumption of commercial taxonomy that each named variety constitutes a visually homogeneous class: within a single commercial label, the model recovers multiple visually distinct sub-populations that would be collapsed and concealed by any label-supervised approach. For marble quarrying and trading applications, this means that intra-variety visual variability is structured and recoverable rather than random noise.

Coherent pure family (Panel c of Figure 14). A 3-sample Bardiglio sub-tree forms a coherent pure family at an intra-family distance of 0.0057, corresponding to only 0.007× the dataset mean pairwise

 doi:10.20944/preprints202603.1344.v1

distance of 0.8677. As visible in Figure 14, the Bardiglio hierarchy contains multiple such pure subtrees at comparable scales, indicating that purity is not isolated to a single instance but is a recurring property of visually coherent sub-populations within this category. Taken together, the three cases of Figure 14 demonstrate that the Ward hierarchy encodes a genuine multi-scale visual taxonomy: it simultaneously reveals the boundaries that commercial classification imposes unnecessarily (Case A), the internal heterogeneity that commercial classification conceals (Case B), and the visual coherence that commercial classification correctly captures (Case C). No single cut of the dendrogram captures all three simultaneously, a property intrinsic to the hierarchical structure, and precisely what a flat k-partition cannot represent.

5. Discussion

The results presented in Section 4 demonstrate that the proposed CA-DINO pipeline discovers a hierarchical visual organization of marble varieties that is both internally coherent and structurally distinct from the commercial taxonomy. This section interprets these findings in the context of existing literature, examines the implications for evaluation methodology, discusses the feature priorities learned by the model, identifies limitations of the current study, and outlines directions for future work.

5.1. Learned Visual Structure versus Commercial Classification

The three phenomena documented in Section 4.6, cross-category merging, intra-category splitting, and coherent pure family formation, collectively demonstrate that the relationship between commercial naming conventions and intrinsic visual structure is neither arbitrary nor fully aligned, but rather partially overlapping. The pipeline correctly recovers commercial boundaries where they correspond to genuine visual discontinuities, as evidenced by the coherent Bardiglio subtrees and the stable Arabescato Brown grouping across all six hierarchical levels. Simultaneously, it identifies commercially invisible relationships, such as the grey-veined family grouping Branco Peletigre, Dante Grey, Arabescato Brown, and Bardiglio at Level 2, and commercially concealed heterogeneity, such as the multiple Ruivina sub-populations separated by an inter-to-intra distance ratio of 257. These findings validate the central premise of this work: that supervised methods trained on commercial labels are structurally incapable of uncovering this partial misalignment, as they would be forced to either collapse the grey-veined family into separate categories or merge the distinct Ruivina sub-populations.

This result extends the observations of Brondolo and Beaussant [18] and Scabini et al. [19], who demonstrated the effectiveness of DINO-based SSL in geological and material science domains, by showing that self-supervised representations are not only competitive with supervised approaches in classification accuracy but are qualitatively superior when the objective is to discover rather than replicate categorical structure.

5.2. The Role and Limitations of Evaluation Metrics

A central methodological insight of this study concerns the fundamental misalignment between standard extrinsic clustering metrics and the objective of unsupervised taxonomy discovery. As demonstrated in Sections 4.3 and 4.5, configurations achieving the highest ARI and NMI scores, such as the raw 2048-dimensional embeddings (ARI=0.593), did so as an artefact of high-dimensional noise inflation rather than genuine clustering quality, as confirmed by their poor internal geometry (SS=0.489, DB=0.768). Conversely, CA-DINO at k=10, the configuration producing the most perceptually coherent hierarchy, achieved a moderate ARI of 0.305, precisely because it correctly grouped visually similar stones from different commercial categories.

This observation has broader implications for unsupervised learning research in domains with unreliable labels. When the ground truth itself encodes the inconsistencies that the method aims to resolve, extrinsic metrics measure conformity to a flawed reference rather than discovery quality. We

 doi:10.20944/preprints202603.1344.v1

therefore advocate for a multi-criterion evaluation strategy combining internal geometric measures with systematic qualitative dendrogram analysis, an approach aligned with the broader recognition that hierarchical structure requires evaluation tools sensitive to multi-scale organization rather than flat partition agreement [37,38].

5.3. Feature Priorities in the Learned Embedding Space

The level-by-level dendrogram analysis reveals a consistent hierarchy of visual feature priorities encoded by the CA-DINO representations. The root split at Level 1 separates warm-toned from cool-toned stones, indicating that base chromaticity constitutes the dominant axis of variation in the embedding space. Subsequent splits progressively resolve finer distinctions: veining density and pattern topology at Levels 2–3, and textural granularity and background saturation at Levels 4–6. This general-to-specific organization mirrors the perceptual strategy reported by domain experts who select marble primarily based on colour family before attending to veining and textural details [4], suggesting that the self-supervised objective, combined with the cluster-aware loss, learns a feature hierarchy that is perceptually grounded without explicit supervision.

The superiority of UMAP over raw embeddings for downstream clustering (Section 4.3) further indicates that the discriminative features learned by the ViT backbone are encoded non-linearly, consistent with previous findings that Vision Transformers capture complex spatial relationships through multi-head self-attention [27,28]. The failure of linear dimensionality reduction to preserve this structure underscores the importance of manifold-aware post-processing for ViT-derived embeddings in fine-grained visual domains.

5.4. Sensitivity to the Cluster Count Hyperparameter

The ablation across k values (Section 4.5) revealed that pipeline performance is sensitive to the alignment between the training cluster count and the dataset's intrinsic visual complexity. CA-DINO at k=10 substantially outperformed all other k configurations across both internal and external metrics, while k=8 (matching the number of test-set commercial varieties) produced inferior geometric cohesion. This counterintuitive result suggests that the optimal training cluster count does not correspond to the number of commercial categories but rather to the number of visually distinguishable sub-populations in the data, a quantity that exceeds the commercial category count due to intra-variety heterogeneity. Future deployments of this pipeline should therefore treat k as a hyperparameter to be tuned against internal clustering quality, rather than setting it to the expected number of output categories.

5.5. Limitations

Several limitations of the current study should be acknowledged. First, the dataset comprises 1,540 training images spanning 10 commercial varieties and 94 test images, a scale that, while sufficient to demonstrate the methodology, does not yet represent the full diversity of commercially available marble. Second, the evaluation relies on a single geological material; generalisability to other natural stones (granite, travertine, slate) or to broader material science domains remains to be empirically validated. Third, the pipeline currently operates on individual slab images and does not model spatial continuity across adjacent slabs from the same block, a property relevant to industrial matching applications. Fourth, while the qualitative dendrogram analysis provides strong evidence of hierarchical coherence, it is inherently subjective; developing quantitative metrics specifically designed for unsupervised taxonomy evaluation remains an open challenge. Additionally, the pipeline was developed using the original DINO architecture [26] as integrated within the CA-DINO framework [36]; evaluating whether more recent self-supervised ViT variants yield further improvements in embedding quality for this domain remains a direction for future investigation. The reported metric comparisons between CA-DINO k=10 and the Pure DINO baseline are based on a single held-out test set of 94 samples without bootstrap confidence intervals; however, the

 doi:10.20944/preprints202603.1344.v1

consistency of the CA-DINO k=10 advantage across multiple independent metrics (SS, DB, NMI, V-measure) and the magnitude of the improvements relative to the inter-configuration variance observed across all five k values (Table 6) provide converging evidence that the observed differences reflect genuine representational gains rather than sampling variability.

6. Conclusions

This work presented and validated an unsupervised pipeline for hierarchical visual taxonomy of marble natural stone, combining cluster-aware self-supervised learning (CA-DINO) with UMAP dimensionality reduction and Ward's agglomerative hierarchical clustering. Through systematic ablation studies on a dataset of 1,540 marble images spanning 10 commercial varieties, we demonstrated that each pipeline component contributes to the quality of the final taxonomy: the cluster-aware training objective produces geometrically superior embeddings compared to pure self-supervised learning (SS=0.778 vs. 0.761; DB=0.293 vs. 0.364), UMAP 5D compression resolves high-dimensional noise pathologies while preserving discriminative structure, and Ward's linkage yields the most compact and well-separated hierarchical partitions.

The resulting visual taxonomy exhibits three properties that distinguish it from commercial classification: it groups visually similar stones across commercial boundaries, it separates visually distinct sub-populations within single commercial categories, and it preserves commercially meaningful groupings where visual and commercial boundaries genuinely coincide. These findings provide empirical evidence that the commercial naming system is partially, but not fully, aligned with intrinsic visual structure, a nuanced characterisation that neither supervised classification nor simple rejection of commercial labels can capture.

From a methodological standpoint, we demonstrated that standard extrinsic clustering metrics (ARI, NMI) are fundamentally misaligned with the objectives of unsupervised taxonomy when the reference labels encode the inconsistencies the method aims to resolve. This finding has implications beyond the marble domain, applying to any unsupervised learning task where existing categorical labels reflect convention rather than systematic criteria.

Future work will pursue three directions: (1) scaling the pipeline to larger and more diverse natural stone datasets, including granite and travertine, to assess cross-material generalisability; (2) integrating slab-level spatial continuity to support industrial block-matching applications; and (3) developing quantitative evaluation metrics specifically tailored to hierarchical taxonomy coherence in the absence of reliable ground truth. The pipeline's architecture is modular and domain-agnostic, providing a transferable template for visual taxonomy problems in materials science, manufacturing, and other domains characterised by unreliable or convention-driven labelling systems

Author Contributions: Conceptualization— M.F., A.A.C.; methodology— M.F., A.A.C.; software— M.F., and C.D.; validation— M.F., A.A.C.; formal analysis— M.F.; investigation— M.F.; resources— G.P. and P.A.; data curation— C.D., M.F.; writing— original draft preparation— M.F.; writing—review and editing— M.F, A.A.C., M.F., C.D., G.P. and P.A.; visualization— M.F., A.A.C.; supervision— A.A.C.; project administration— P.A.; funding acquisition— G.P. and P.A. All authors have read and agreed to the published version of the manuscript.

Funding: The authors acknowledge to the Sustainable Stone by Portugal project, proposal number C644943391 - 00000051 co-financed by the PRR – Recovery and Resilience Plan of the European Union (Next Generation EU). The authors gratefully acknowledge the support of the CERENA through FCT Project UID/04028/2025 (https://doi.org/10.54499/UID/04028/2025).

Informed Consent Statement: All participants provided informed consent prior to evaluation. This study was conducted in accordance with institutional ethical guidelines for non-invasive behavioral research, and formal ethical review was waived as the study involved only expert-quality assessment tasks without the collection of personal or sensitive data.

 doi:10.20944/preprints202603.1344.v1

Data Availability Statement: While the raw industrial scans remain proprietary, the extracted binary masks and trained model weights will be made available upon publication to support reproducibility.

Acknowledgments: The authors acknowledge to the Sustainable Stone by Portugal project, proposal number C644943391 - 00000051 co-financed by the PRR − Recovery and Resilience Plan of the European Union (Next Generation EU). The authors gratefully acknowledge the support of the CERENA through FCT Project UID/04028/2025 (https://doi.org/10.54499/UID/04028/2025).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

CA-DINO	Cluster-Aware Distillation with No Labels
DINO	Distillation with No Labels
ViT	Vision Transformer
SSL	Self-Supervised Learning
DC	Deep Clustering
DEC	Deep Embedded Clustering
UMAP	Uniform Manifold Approximation and Projection
MLP	Multilayer Perceptron
GELU	Gaussian Error Linear Unit
EMA	Exponential Moving Average
DLG	Dynamic Loss Gate
GMM	Gaussian Mixture Model
KL	Kullback–Leibler
PCA	Principal Component Analysis
GLCM	Gray-Level Co-occurrence Matrix
LBP	Local Binary Patterns
CNN	Convolutional Neural Network
SS	Silhouette Score
DB	Davies–Bouldin Index
CH	Calinski–Harabasz Index
ARI	Adjusted Rand Index
NMI	Normalized Mutual Information
CCC	Cophenetic Correlation Coefficient
SSIM	Structural Similarity Index Measure
CS	Cosine Similarity

References

1. Pereira, D.; Marker, B. The Value of Original Natural Stone in the Context of Architectural Heritage. *Geosciences (Switzerland)* **2016**, *6*, doi:10.3390/geosciences6010013.
2. Navarro, R.; Pereira, D.; Gimeno, A.; del Barrio, S. Verde Macael: A Serpentinite Wrongly Referred to as a Marble. *Geosciences (Switzerland)* **2013**, *3*, 102–113, doi:10.3390/geosciences3010102.

 doi:10.20944/preprints202603.1344.v1

3. Muñoz-Cervera, M.C.; Rodriguez-Garcia, M.Á.; Cañaveras, J.C. Aesthetic Quality Properties of Carbonate Breccias Associated with Textural and Compositional Factors: Marrón Emperador Ornamental Stone (Upper Cretaceous, Southeast Spain). *Applied Sciences (Switzerland)* **2022**, *12*, doi:10.3390/app12052566.
4. Strzałkowski, P.; Köken, E.; Sousa, L. Guidelines for Natural Stone Products in Connection with European Standards. *Materials* **2023**, *16*, doi:10.3390/ma16216885.
5. Badouna, I.; Koutsovitis, P.; Karkalis, C.; Laskaridis, K.; Koukouzas, N.; Tyrologou, P.; Patronis, M.; Papatrechas, C.; Petrounias, P. Petrological and Geochemical Properties of Greek Carbonate Stones, Associated with Their Physico-Mechanical and Aesthetic Characteristics. *Minerals* **2020**, *10*, doi:10.3390/min10060507.
6. Alper Selver, M.; Akay, O.; Alim, F.; Bardak, S.; Ölmez, M. An Automated Industrial Conveyor Belt System Using Image Processing and Hierarchical Clustering for Classifying Marble Slabs. *Robot. Comput. Integr. Manuf.* **2011**, *27*, 164–176, doi:10.1016/j.rcim.2010.07.004.
7. Elbehiery, H.; Hefnawy, A.; Elewa, M. Surface Defects Detection for Ceramic Tiles Using Image Processing and Morphological Techniques. *International Journal of Information, Control and Computer Sciences* **2007**, *1*.
8. Haileslassie, F.; Leta, A.; Desalegn, G.; Kalayu, M. Classification of Marble Using Image Processing. *International Journal on Data Science and Technology* **2019**, *5*, 57, doi:10.11648/j.ijdst.20190503.11.
9. Turan, E.; Ucar, F.; Dandil, B. A Novel Marble Recognition System Using Extreme Learning Machine with LBP and Histogram Features. *Concurr. Comput.* **2021**, *33*, doi:10.1002/cpe.6428.
10. Ouzounis, A.G.; Sidiropoulos, G.K.; Papakostas, G.A.; Sarafis, I.T.; Stamkos, A.; Solakis, G. Interpretable Deep Learning for Marble Tiles Sorting. In Proceedings of the Proceedings of the 2nd International Conference on Deep Learning Theory and Applications, DeLTA 2021; SciTePress, 2021; pp. 101–108.
11. CANAYAZ, M.; ULUDAĞ, F. MARBLE CLASSIFICATION USING DEEP NEURAL NETWORKS. *European Journal of Technic* **2020**, 52–63, doi:10.36222/ejt.671527.
12. Ouzounis, A.G.; Taxopoulos, G.; Papakostas, G.A.; Sarafis, I.T.; Stamkos, A.; Solakis, G. Marble Quality Assessment with Deep Learning Regression. In Proceedings of the 5th International Conference on Intelligent Computing in Data Sciences, ICDS 2021; Institute of Electrical and Electronics Engineers Inc., 2021.
13. Selver, M.A.; Akay, O.; Ardali, E.; Yavuz, B.A.; Önal, O.; Özden, G. Cascaded and Hierarchical Neural Networks for Classifying Surface Images of Marble Slabs. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews* **2009**, *39*, 426–439, doi:10.1109/TSMCC.2009.2013816.
14. Selver, M.A.; Akay, O. Evaluating Clustering Methods for Classification of Marble Slabs in an Automated Industrial Marble Inspection System;
15. Al-Zoubi1, H.R.; Al-Khassaweneh, M.A.; Altawil, I.A. *An Image Processing Approach for Marble Classification*; 2015; Vol. 1;.
16. Sipko, E.; Kravchenko, O.; Karapetyan, A.; Plakasova, Zh.; Gladka, M. The System Recognizes Surface Defects Of Marble Slabs Based On Segmentation Methods. *Scientific Journal of Astana IT University* **2020**, doi:10.37943/aitu.2020.1.63643.
17. Sidiropoulos, G.K.; Ouzounis, A.G.; Papakostas, G.A.; Lampoglou, A.; Sarafis, I.T.; Stamkos, A.; Solakis, G. Hand-Crafted and Learned Feature Aggregation for Visual Marble Tiles Screening. *J. Imaging* **2022**, *8*, doi:10.3390/jimaging8070191.
18. Brondolo, F.; Beaussant, S. DINOv2 Rocks Geological Image Analysis: Classification, Segmentation, and Interpretability. **2024**.
19. Scabini, L.; Sacilotti, A.; Zielinski, K.M.; Ribas, L.C.; De Baets, B.; Bruno, O.M. A Comparative Survey of Vision Transformers for Feature Extraction in Texture Analysis. *J. Imaging* **2025**, *11*, doi:10.3390/jimaging11090304.
20. Zhu, T.; Braytee, A.; Thiyagarajan, K.; Zi, X.; Mustapha, S.; Tao, X.; Prasad, M. Autonomous Detection of Concrete Cracks Using Self-Supervised DinoV2. *Machine Intelligence Research* **2026**, *23*, 168–184, doi:10.1007/s11633-025-1553.
21. Gui, J.; Chen, T.; Zhang, J.; Cao, Q.; Sun, Z.; Luo, H.; Tao, D. A Survey on Self-Supervised Learning: Algorithms, Applications, and Future Trends. **2024**.

 doi:10.20944/preprints202603.1344.v1

22. Jaiswal, A.; Babu, A.R.; Zadeh, M.Z.; Banerjee, D.; Makedon, F. A Survey on Contrastive Self-Supervised Learning. *Technologies (Basel).* 2021, *9*.
23. Caron, M.; Bojanowski, P.; Joulin, A.; Douze, M. Deep Clustering for Unsupervised Learning of Visual Features. **2019**.
24. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. **2020**.
25. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. **2020**.
26. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. **2021**.
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR 2021 · The Ninth International Conference on Learning Representations* **2021**.
28. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A Survey on Vision Transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 87–110, doi:10.1109/TPAMI.2022.3152247.
29. Leiber, C.; Miklautz, L.; Plant, C.; Böhm, C. An Introductory Survey to Autoencoder-Based Deep Clustering -- Sandboxes for Combining Clustering with Deep Learning. **2025**.
30. Xie, J.; Girshick, R.; Farhadi, A. Unsupervised Deep Embedding for Clustering Analysis. **2016**.
31. Wang, R.; Li, L.; Wang, P.; Tao, X.; Liu, P. Feature-Aware Unsupervised Learning with Joint Variational Attention and Automatic Clustering. In Proceedings of the Proceedings - International Conference on Pattern Recognition; Institute of Electrical and Electronics Engineers Inc., 2020; pp. 923–930.
32. Zhou, S.; Xu, H.; Zheng, Z.; Chen, J.; li, Z.; Bu, J.; Wu, J.; Wang, X.; Zhu, W.; Ester, M. A Comprehensive Survey on Deep Clustering: Taxonomy, Challenges, and Future Directions. **2022**.
33. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)* **2021**.
34. Li, Y.; Hu, P.; Liu, Z.; Peng, D.; Zhou, J.T.; Peng, X. Contrastive Clustering. *Proceedings of the AAAI Conference on Artificial Intelligence* **2021**, *Vol. 35 No. 10*.
35. Li, J.; Zhou, P.; Xiong, C.; Hoi, S.C.H. Prototypical Contrastive Learning Of Unsupervised Representations. *ICLR 2021 · The Ninth International Conference on Learning Representations*.
36. Han, B.; Chen, Z.; Qian, Y. Self-Supervised Learning with Cluster-Aware-DINO for High-Performance Robust Speaker Verification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *32*.
37. Naumov, S.; Yaroslavtsev, G.; Avdiukhin, D. Objective-Based Hierarchical Clustering of Deep Embedding Vectors. *Proceedings of the AAAI Conference on Artificial Intelligence* **2021**, *Vol. 35 No. 10*.
38. Yang, J.; Parikh, D.; Batra, D. Joint Unsupervised Learning of Deep Representations and Image Clusters. In Proceedings of the Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; IEEE Computer Society, December 9 2016; Vol. 2016-December, pp. 5147–5156.
39. Kaufman, L.; Rousseeuw, P.J. Finding Groups in Data An Introduction to Cluster Analysis;
40. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. **2019**.
41. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. **2017**.
42. Rousseeuw, P.J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis; 1987; Vol. 20;.
43. Davies, D.L.; Bouldin, D.W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1*, 224–227, doi:10.1109/TPAMI.1979.4766909.
44. Caliński, T.; Harabasz, J. A Dendrite Method Foe Cluster Analysis. *Communications in Statistics* **1974**, *3*, 1–27, doi:10.1080/03610927408827101.
45. Hubert, L.; Arabic, P. Comparing Partitions. *J. Classif.* **1985**, *2*, 193–218.
46. Strehl, A.; Ghosh, J. Cluster Ensembles-A Knowledge Reuse Framework for Combining Multiple Partitions; 2002; Vol. 3;.

 doi:10.20944/preprints202603.1344.v1

47. Rosenberg, A.; Hirschberg, J. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure; 2007;
48. Sokal, R.R.; Rohlf, F.J.; James, F.; Lawrence, R. *The Comparison of Dendrograms by Objective Methods*; 1962; Vol. 11;.
49. Milligan, G.W.; Cooper, M.C. An Examination of Procedures for Determining the Number of Clusters in a Data Set; 1985; Vol. 50;.