

Concept Retrieval—What and How?

Ori Nizan
Technion, Israel
snizori@campus.technion.ac.il

Oren Shrout
Technion, Israel
shrout.oren@gmail.com

Ayellet Tal
Technion, Israel
ayellet@ee.technion.ac.il



Figure 1. Given a query image (left) our method retrieves images from the dataset that share key concepts, relying exclusively on the visual characteristics of the input. The extracted concepts can be interpreted as: (1) Astronaut in space with a planet in the background (2) Human with a backpack on an exploration journey (3) Sunset (4) Heroic figure in an otherworldly atmosphere (5) Astronaut cartoon.

Abstract

A concept may reflect either a concrete or abstract idea. Given an input image, this paper seeks to retrieve other images that share its central concepts, capturing aspects of the underlying narrative. This goes beyond conventional retrieval or clustering methods, which emphasize visual or semantic similarity. We formally define the problem, outline key requirements, and introduce appropriate evaluation metrics. We propose a novel approach grounded in two key observations: (1) While each neighbor in the embedding space typically shares at least one concept with the query, not all neighbors necessarily share the same concept with one another. (2) Modeling this neighborhood with a bimodal Gaussian distribution uncovers meaningful structure that facilitates concept identification. Qualitative, quantitative, and human evaluations confirm the effectiveness of our approach. See the package on PyPI: <https://pypi.org/project/coret/>

1. Introduction

A concept is a mental representation that help humans categorize and interpret the world. In this sense, a concept is not merely a collection of visually similar features but rather an abstract, high-level grouping based on shared meaning, function, or context [32]. Concepts can range from concrete (like "Forrest") to abstract (like "happiness") and are fundamental to thinking, reasoning, and learning. In this paper, we introduce the task of retrieving images based on shared concepts. This task can be seen as a generalization of image retrieval, which traditionally focuses on retrieving visually similar images from a dataset [11, 35]. In contrast, our task emphasizes high-level semantics over visual similarity and aims to capture abstract and contextual meaning effectively. For example, given an image of an astronaut exploring space (Fig. 1), concept retrieval may return images of astronauts in various atmospheres and planets (Concept 1), exploratory journey (Concept 2), and so on. In this example, the concept of an 'exploration journey' is better captured

through activity and context rather than mere visual similarity. This capability is especially valuable for AI-driven creative applications. For example, in advertising, brands often seek images that express concepts such as innovation, exploration, or family values, rather than depicting a specific object. In the arts and creative industries, artists and curators frequently retrieve works based on themes or concepts rather than literal objects. In psychology, such retrieval can support studies of visual metaphors and the ways in which people interpret abstract ideas through images. A key consideration is defining the essential requirements of the task. We identify four such requirements: (1) *Relevance*: the retrieved images should reflect a concept present in the input image; (2) *Consistency*: images retrieved for a particular concept should consistently represent that concept; (3) *Inner-concept diversity*: images within a single concept should vary rather than appear nearly identical; (4) *Cross-concept diversity*: images retrieved for one concept should be semantically different from those of other concepts. Once the requirements are defined, the next step is to develop an effective approach to address the problem. Our method is based on the following observations. Each image in the embedding space neighborhood of a query tends to share at least one concept with it. At the same time, not all neighboring images necessarily share the same concept with one another. However, if a sufficient number of neighbors express a particular concept, that concept can be reliably identified as a primary concept of the query. The key challenge is to effectively leverage these relationships within the embedding space to extract meaningful concepts. To isolate a concept embedding, we aim to partition the embeddings in the local neighborhood into two subsets: one containing images that represent the concept and another excluding it. The challenge lies in achieving this without prior knowledge of the concept itself. The key idea is to identify a *surrogate* embedding within the neighborhood, which will allow us to model the similarity distribution of neighboring embeddings as a bimodal Gaussian. In this distribution, one Gaussian corresponds to the concept, while the other represents images that do not share the concept. Since each image contains multiple concepts, once images sharing a particular concept are identified, the dataset embeddings should be updated to reduce that concept’s influence. Following this adjustment, the search for a new concept resumes. As the embeddings change, the input image’s neighborhood also shifts, allowing new concepts to emerge. This iterative process leverages the evolving embedding space to ensure diverse and relevant concepts. But how should the results be evaluated? As this is a new task, dedicated evaluation metrics are needed. We introduce four evaluation metrics, each corresponding to one requirement. It is important to note that these requirements may not always align; for example, consistency and inner-concept di-

versity can sometimes be in conflict. Therefore, measuring each requirement separately is crucial, allowing applications to determine the appropriate balance based on their specific needs. Additionally, a human evaluation methodology is proposed to capture subjective opinions on how well the results adhere to each requirement. This paper makes the following contributions:

1. Defining a new problem that generalizes the image retrieval task, along with establishing its key requirements.
2. Proposing a novel approach to address the task, which is both efficient and scalable. Quantitative, qualitative, and human evaluation results demonstrate the method’s effectiveness across heterogeneous datasets.
3. Introducing new evaluation metrics designed to assess these requirements effectively.

2. Related Work

This paper aims to retrieve images that share the underlying concepts as a given image. The term *concept* has been interpreted in various ways in computer vision, often diverging from its psychological definition, as a mental representation that forms abstract, high-level groupings based on shared meaning, function, or context. Within computer vision, our work can be viewed as a generalization of image retrieval.

Concepts in Computer Vision. The term “concept” has been interpreted in various ways in computer vision. In most cases, it refers to an object or a style and has been applied in tasks such as image generation and editing [15, 16, 21, 22, 25, 26, 40]. In Concept Bottleneck Models (CBMs) [24, 42], object-based concepts have been used to enhance explainability by learning interpretable representations that improve the transparency of decision-making. These approaches can be broadly classified into: (1) Language-guided extraction [1, 34, 44], which leverages textual descriptions to define concepts. (2) Vision-guided extraction [5, 8, 31, 33], which derives concepts directly from images without relying on textual supervision. Other works have explored concepts for explainability. In particular, Ghorbani et al. [17] define concepts as image segments. For example, a “wheel” of a vehicle is considered a distinct concept. In Chattopadhyay et al. [4], concepts are represented as sets of words from a predefined dictionary. Our approach differs from the above by using the term concept in a broader sense, relying solely on images as input, without predefined linguistic guidance or a fixed vocabulary, and by extracting multiple concepts per image.

Content-based image retrieval. This is one of the classical tasks in computer vision. The goal is to identify and return the most relevant images from a database based on specific visual features or content. The classical methods retrieve the most similar images by comparing their feature vectors [14, 23, 28, 37–39]. Recent approaches em-

ploy multimodal retrieval, where text and image features are jointly learned [7, 19, 20] leveraging vision-language models (VLMs) [13, 27, 36]. We refer the readers to recent surveys that highlight the evolution of retrieval methods [3, 6, 9, 12, 43]. Similarly, our objective is to retrieve images that share common traits with a given input image. However, we seek images that share multiple semantically concepts rather than a single visual similar neighborhood. Our concepts are discovered post-hoc from an existing embedding space, without language labels, attributes, or detectors.

3. Method

Given an image, our aim is to retrieve images that share the same concept(s). The proposed method is designed to meet four key requirements:

1. **Relevance:** Each retrieved image should contain a concept that is present in the input image.
2. **Consistency:** Images retrieved for a given concept should accurately represent that concept.
3. **Inner-concept diversity:** The retrieved images for a given concept should exhibit variation.
4. **Cross-concept diversity:** The set of images retrieved for concept i should visually and semantically differ from those retrieved for all previous concepts.

Our approach is built on two key ideas: (1) *Bimodal neighborhood structure*: Within the neighborhood of a given embedding, some images share a specific concept while others do not. This distinction can be modeled as a bimodal Gaussian distribution, though the exact concept remains unknown. (2) *Concept surrogates*: The similarities of certain embeddings to their neighbors clearly exhibit a bimodal Gaussian structure. These embeddings may serve as concept surrogates. By analyzing the neighborhoods of these surrogates (within the query neighborhood), our method isolates the most prominent concepts in the image, consistent with the dataset. The central question, then, is how to analyze this structure. Our method operates in 4 steps: (1) Identify the neighbors of the query embedding. (2) Within this neighborhood, determine the most suitable surrogate neighbor. (3) Extract images that share the concept common to both the surrogate neighbor and the query. (4) Update the dataset and repeat from Step 2 to extract the next concept. We elaborate below.

1. Identify the query neighbors. Images located near each other in the embedding space are likely to share an underlying visual concept. Therefore, given a query image, we begin by identifying its set of neighboring images, ensuring the *relevance* requirement is met. We use cosine similarity:

$$\text{Sim}(\mathbf{e}_i, \mathbf{e}_j) = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\| \|\mathbf{e}_j\|}, \quad (1)$$

where $\mathbf{e}_i$ and $\mathbf{e}_j$ denote the embeddings of two images, I_i

and I_j . To ensure that relevant neighbors are captured, we adopt a relatively large neighborhood size. This allows us to include images that share a common concept but may lie farther apart due to the influence of diverse attributes. In our implementation, proximity is defined using a threshold $T = \sigma$, computed relative to the mean embedding distance μ of the dataset. Thus, the neighborhood of the input embedding e in the embedding space $\mathcal{D}$ is defined as follows:

$$\text{Neighborhood}_T(\mathbf{e}) = \{\mathbf{e}_j \mid \text{Sim}(\mathbf{e}, \mathbf{e}_j) \geq T, \mathbf{e}_j \in \mathcal{D}\}. \quad (2)$$

2. Determine a surrogate neighbor and its concept set.

The goal of this step is to select an effective surrogate image from the neighborhood. This surrogate image helps retrieve a subset of images that share a common concept, thereby ensuring the *consistency* requirement is met. To achieve this, we first compute pairwise similarities between all image embeddings e_i and e_j in the neighborhood. We then construct a similarity-based histogram for each image in the set. Recall our observation that certain embeddings exhibit a clear bimodal Gaussian distribution in their similarity values, making them suitable surrogates s for identifying a concept. Formally, we apply Gaussian Mixture Modeling (GMM) to each histogram:

$$\text{GMM}(\text{Sim}(s, e)) = \pi_1 \mathcal{N}(\mu_1, \sigma_1^2) + \pi_2 \mathcal{N}(\mu_2, \sigma_2^2), \quad (3)$$

where μ_i and σ_i are the means and standard deviations of the two Gaussians and π_i are the mixture coefficients satisfying $\pi_1 + \pi_2 = 1$. If two well-separated Gaussian distributions emerge, this indicates a potential separation between images that contain a certain concept and those that do not. Figure 2 illustrates the Gaussian distribution corresponding to the input image e , presented in Fig. 1, and its surrogate s . Next, to choose the concept from the multiple candidates, we apply the following two criteria to each GMM: (1) The number of samples in each Gaussian must exceed a threshold. This ensures that both sets are sufficiently representative of the data. (2) The input image must belong to the right Gaussian. This guarantees that the identified subset focuses on a concept that is present in the input image. We rank the candidate surrogates to select the image that best separates the two Gaussian distributions, ensuring a clear distinction between the subsets. This step directly supports inner-concept *consistency*, as a greater separation between the Gaussians indicates more distinct distributions, leading to a more consistent concept. To quantify this separation, we define a separation metric score:

$$\text{SepScore}(s) = (\mu_1 - \sigma_1) - (\mu_2 + \sigma_2), \quad (4)$$

where $\mu_1 - \sigma_1$ is the lower bound of the first Gaussian, and $\mu_2 + \sigma_2$ is the upper bound of the second Gaussian. A higher *SepScore* indicates greater separation. The top-scoring image is chosen. Finally, we define the sub-space

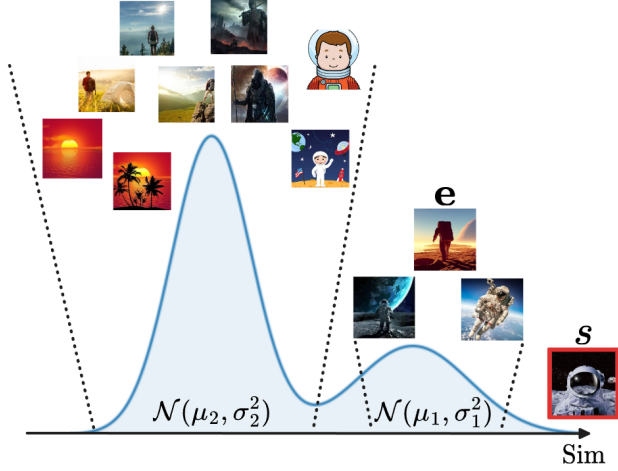


Figure 2. **Similarity score distribution.** This image shows a bi-modal Gaussian of similarity scores between a surrogate s and the input’s neighbors from Fig. 1. The smaller (right) mode defines the ‘concept’ set; the larger (left) defines the ‘non-concept’ set.

associated with the concept. To achieve this, we focus on the Gaussian with the larger mean similarity, as it indicates stronger similarity among the samples and therefore represents the concept. We denote this Gaussian as $\mathcal{N}_s(\mu_s, \sigma_s^2)$. For each embedding $e_j \in \mathcal{N}_s(\mu_s, \sigma_s^2)$ in the neighborhood, we compute its probability of belonging to this Gaussian as follows:

$$Pr(e_j) = \frac{\pi_s \mathcal{N}_s(e_j | \mu_s, \sigma_s^2)}{\pi_s \mathcal{N}_s(e_j | \mu_s, \sigma_s^2) + (1 - \pi_s) \mathcal{N}_{\bar{s}}(e_j | \mu_{\bar{s}}, \sigma_{\bar{s}}^2)}. \quad (5)$$

The concept subspace $\mathcal{C}(e, s, \tau)$, defined for an input embedding e and surrogate s , consists of samples whose membership probabilities exceed the threshold τ , that is, embeddings with higher similarity scores:

$$\mathcal{C}(e, s, \tau) = \{e_i | e_i \in Neighbor_T(e), Prob(e_i) > \tau\}. \quad (6)$$

3. Concept extraction. The objective of this step is to extract images that share a concept emerging from Eq. 6. Although the set $\mathcal{C}(e, s, \tau)$ captures a single general concept, we may wish to focus on specific variations. For example, each column in Fig. 3 illustrates a distinct variation of the concept ‘a dog jumping for a frisbee.’ We propose a three-step process to extract a subset of $\mathcal{C}(e, s, \tau)$: (1) constructing a subspace of the concept, (2) projecting the input onto this subspace, and (3) extracting images corresponding to the concept by identifying the closest embeddings within this subspace. We elaborate on these steps below. To create a concept subspace, we aim to capture the main attributes of the concept. This is achieved by applying *Principal Component Analysis (PCA)* to $\mathcal{C}(e, s, \tau)$, allowing us

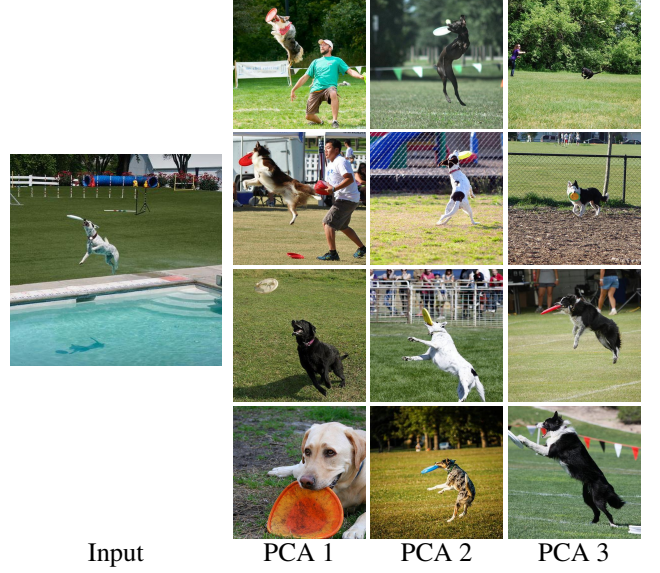


Figure 3. **PCA navigations.** Each column represents a direction within the PCA subspace. They display variations in (1) jump height, (2) breed, and (3) distance from the camera. However, they all share the underlying concept of ‘a dog jumping for a frisbee.’

to extract the dominant components that define the concept. The PCA-concept subspace is defined by the top k principal components, corresponding to the largest singular values, and is spanned by these vectors, represented in the matrix $W_k \in \mathbb{R}^{d \times k}$. Fig. 3 demonstrates that navigating in different directions within the PCA space leads to variations of the concept. PCA 1 varies with the dog’s height, PCA 2 with breed, and PCA 3 with camera distance. Sampling within the PCA subspace yields *inner-concept diversity*. Up to this point, we have constructed a subspace that represents the concept, independent of the details of the input image. To further enhance the relevance of the retrieved images to the input, we project the input embedding $e \in \mathbb{R}^d$ onto the PCA-concept subspace. This isolates the attributes relevant to the concept, as the image may contain multiple overlapping concepts. This is achieved by computing:

$$e_c = e W_k W_k^\top. \quad (7)$$

Recall that the first principal component captures the largest possible variance, while subsequent ones capture progressively less. The key question is how many components to select to capture the concept’s essence while minimizing noise. The ratio between the sum of the variances of the top k PCA components and the total variance quantifies how well the reduced representation preserves the original spread of the data. Specifically, we use a captured variance threshold of $\tau_1 = 25\%$, ensuring that the reduced representation remains meaningful and informative. Thus, we

choose the smallest k that satisfies this criterion:

$$k = \min \left\{ k \mid \sum_{i=1}^k \frac{\sigma_i^2}{\sum_{j=1}^d \sigma_j^2} \geq \tau_1 \right\} \quad (8)$$

where σ_i is the std of the i^{th} principal component. Finally, the concept embedding $\mathbf{e}_c$ is used to retrieve the nearest neighbors. This is accomplished by applying cosine similarity to identify the most similar embeddings.

4. Update the dataset and iterate. After processing the current concept, we shift focus to the next to promote *cross-concept diversity*. The aim is to isolate concept-specific attributes and reduce their influence in later iterations. This is achieved by subtracting the identified concept from a targeted subset of dataset embeddings, ensuring the next selected Gaussian captures a distinct concept. The subtraction is carried out as follows: for each image embedding $\mathbf{e}_i$ in the subset, we begin by computing its projection onto the concept subspace, as defined in Eq. 7. Next, we subtract this projection from the embedding of the corresponding image and update the dataset accordingly:

$$\mathbf{e}_i \leftarrow \mathbf{e}_i - \mathbf{e}_i \mathbf{W}_k \mathbf{W}_k^\top. \quad (9)$$

The subset to which we apply Eq. 9 contains the top dataset embeddings that best match (Eq. 1) the mean embedding of $\mathcal{C}(e, s, \tau)$. We focus on a subset, not the full dataset, since the removed concept may co-occur with others. Our aim is to avoid re-extracting it as a standalone concept, not to eliminate it entirely. For example, the astronaut in Concept 1 of Fig. 1 may also appear in other concepts, like the cartoon in Concept 5. This iterative process ensures that every concept is different from the previous ones, thereby allowing the exploration of multiple, potentially overlapping, concepts.

4. Evaluation metrics

There are no established metrics for our task. Although related, image retrieval focuses on finding visually similar images, so standard metrics like precision and recall are not directly applicable. Instead, evaluation should consider the four requirements in Section 3. We propose a quantitative set of scores that is based on the concept embedding (Section 4.1) Additionally, Section 4.2 also outlines our human study method.

4.1. Quantitative metric

Relevance Score (RS). This metric evaluates a concept's relevance to the input image. A straightforward approach would be to compute the similarity between the input embedding $\mathbf{e} \in \mathbb{R}^d$ and the concept embedding $\mathbf{e}_c \in \mathbb{R}^d$ from Eq. 7, using $\text{Sim}(\mathbf{e}, \mathbf{e}_c)$ (Eq. 1). But this raw similarity is too general, yielding retrieval-like results rather than capturing a specific concept. To address this, we normalize similarity relative to concept distributions across the dataset.

The challenge is defining this normalization. We suggest extracting concepts for all images in the dataset (or approximating this using a random subset). Given this set, the similarity scores between the input image and the concepts form a Gaussian distribution with mean μ and standard deviation σ . The normalized relevance is then defined as:

$$\text{RS}(\mathbf{e}, \mathbf{e}_c) = \Phi \left(\frac{\text{Sim}(\mathbf{e}, \mathbf{e}_c) - \mu}{\sigma} \right), \quad (10)$$

where Φ is the cumulative distribution function of the standard normal distribution. This measures how the similarity score deviates from the overall distribution. The image-level relevance score, $\text{ImRS}(I)$, is defined as the sum of the relevance scores of all extracted concepts for image I :

$$\text{ImRS}(I) = \frac{1}{|\text{Conc}(\mathbf{e})|} \sum_{\mathbf{e}_c \in \text{Conc}(\mathbf{e})} \text{RS}(\mathbf{e}_c). \quad (11)$$

Consistency Score (CS). This metric measures the consistency of images retrieved within a concept. Since the exact decomposition of an image into concepts is unknown, we compute the normalized sum of embeddings. Unrelated concepts, being largely uncorrelated, cancel out when combined. In contrast, a consistent concept shared across images reinforces certain entries, yielding a higher magnitude. Formally, given n retrieved image embeddings $\{\mathbf{e}_j\}_{j=1}^n$ for concept $\mathbf{e}_c$, we define the concept-level consistency score $\text{CS}(\mathbf{e}_c)$ and the image-level score $\text{ImCS}(I)$ as:

$$\begin{aligned} \text{CS}(\mathbf{e}_c) &= \left\| \frac{1}{n} \sum_{j=1}^n \mathbf{e}_j \right\|, \\ \text{ImCS}(I) &= \frac{1}{|\text{Conc}(\mathbf{e})|} \sum_{\mathbf{e}_c \in \text{Conc}(\mathbf{e})} \text{CS}(\mathbf{e}_c). \end{aligned} \quad (12)$$

Inner-Diversity Score (IDS). This metric quantifies variance among retrieved images for a concept, aiming to span a subspace that captures it. As shown in Fig. 3, variations arise by exploring the subspace in different directions. Therefore, we approximate this subspace with Principal Component Analysis (PCA) on the embeddings of retrieved images. Diversity is then measured as the cumulative variance explained by the top K principal components. Formally, given a concept embedding $\mathbf{e}_c$ and its retrieved image embeddings, we fit a PCA and obtain eigenvalues $\sigma_1^2 \geq \dots \geq \sigma_d^2$. With K leading components retained, the concept-level $\text{IDS}(\mathbf{e}_c)$ and image-level $\text{ImIDS}(I)$ inner-diversity scores are defined as:

$$\begin{aligned} \text{IDS}(\mathbf{e}_c) &= \sum_{k=1}^K \sigma_k^2 / \sum_{k=1}^d \sigma_k^2, \\ \text{ImIDS}(I) &= \frac{1}{|\text{Conc}(\mathbf{e})|} \sum_{\mathbf{e}_c \in \text{Conc}(\mathbf{e})} \text{IDS}(\mathbf{e}_c), \end{aligned} \quad (13)$$

where a larger value indicates higher inner diversity. **Cross-Diversity Score (CDS).** This metric captures variance across concepts. To promote diversity, we first quantify their pairwise distinctiveness as:

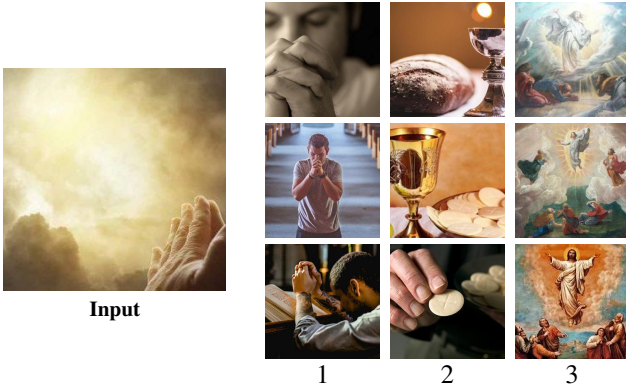
$$\text{CDS}(\mathbf{e}_{c_i}, \mathbf{e}_{c_j}) = \frac{1}{2} (1 - \text{Sim}(\mathbf{e}_{c_i}, \mathbf{e}_{c_j})),$$



(a) The images retrieved by our algorithm can be interpreted as: (1) 'Street art,' (2) 'Woman in white,' and (3) 'Revolution' (communist red star).



(b) The images retrieved by our algorithm can be interpreted as: (1) 'A surfing dog,' (2) 'Surfing boards,' and (3) 'A dog in funny costume.'



(c) The images retrieved by our algorithm can be interpreted as: (1) 'A praying person,' (2) 'Sacramental bread,' and (3) 'Heavenly ascension.'



(d) The images retrieved by our algorithm can be interpreted as: (1) 'Origami,' (2) 'Autumn,' and (3) 'A drawing of autumn leaves.'

Figure 4. **Qualitative results.** Each subfigure shows an input image with three columns, each depicting a distinct extracted concept. All concepts are relevant, consistent, diverse and quite creative. See supplementary material for additional examples.

where higher values indicate greater diversity. We define the concept-level $\text{concCDS}(\mathbf{e}_c)$ as the minimum CDS to other concepts, and the image-level $\text{ImCDS}(I)$ as the normalized sum across concepts:

$$\begin{aligned} \text{concCDS}(\mathbf{e}_c) &= \min_{\mathbf{e}_{c_j} \neq \mathbf{e}_{c_i}} \text{CDS}(\mathbf{e}_{c_i}, \mathbf{e}_{c_j}), \\ \text{ImCDS}(I) &= \frac{1}{|\text{Conc}(\mathbf{e})|} \sum_{\mathbf{e}_c \in \text{Conc}(\mathbf{e})} \text{concCDS}(\mathbf{e}_c). \end{aligned} \quad (14)$$

4.2. Human evaluation methodology

Since visual concepts are inherently abstract, quantitative metrics may not fully capture human intent. We therefore propose a human evaluation to assess the quality of retrieved concepts against the key requirements.

Evaluating consistency. First, participants are asked to evaluate the extent to which a given set of images share a common concept. Next, they are instructed to describe the identified concept in 1–4 words of free text. This setup was designed to test whether the retrieved images convey a

consistent concept on their own. Importantly, at this stage participants do not have access to the input image.

Evaluating relevance. Next, a source image is shown to participants, who must determine whether their identified concept is fully contained, partially contained (or related), or not contained in the source image. The source image may be the original input, a randomly selected image, or one with the same objects arranged differently. The random image is expected to show no relevance, the re-arranged image partial relevance, and the input image to share the concept. Thus, the first two serve as negative controls.

Evaluating inner-concept diversity. Participants are presented with the input image alongside two retrieved sets: one generated by our concept-retrieval algorithm and the other by classical retrieval [13, 36]. They are then asked to compare their diversity by choosing one of the following: (1) Both sets exhibit similar diversity. (2) The first set (concept-retrieval) is more diverse. (3) The second set (classical retrieval) is more diverse.



Figure 5. Domain-specific concepts.

Evaluating cross-concept diversity. To assess concept distinctiveness, participants are shown the input image with three related concept sets. For each set, they first repeat the relevance evaluation and then choose one of the following: (1) Yes, the concepts are different. (2) No, the concepts are not different. (3) Similar to one but different from another.

5. Results

Datasets. To evaluate the proposed method, we used diverse datasets spanning different domains and semantic complexities. *COCO* [29] contains 330,000 images with 80 object categories across various contexts. *LAION-Aesthetics* is a subset of the *LAION-5B* [41] dataset, curated to include images of high aesthetic quality. It comprises millions of images rated based on aesthetic scores, capturing diverse visual styles. *DeepFashion* [30] consists of over 800,000 diverse fashion images, each annotated with 50 clothing categories and 1,000 descriptive attributes.

Qualitative results. Figure 4 shows some examples of our algorithm applied to images from [29, 41]. The input images span diverse contexts, including urban artistic expressions, animals and sports activities, religious scenes, and nature in art. For instance, in Figure 4a, our method extracts three sets of images, whose shared concepts may be interpreted as: (1) ‘street art,’ (2) ‘a woman wearing a long, white dress,’ and (3) ‘revolution (an arm projecting a sense of power, with political/social symbolism such as a red star, flag, flowers, or stone).’ The retrieved images satisfy the four key requirements. They are relevant, containing visual elements from the input image. Each set maintains a consistent theme. Inner-diversity is reflected in variations of pose and scene details. Finally, the three concept sets remain visually and semantically distinct, ensuring cross-diversity. Figure 5 demonstrates a domain-specific result on an input image from *LAION-Aesthetics* [41], using the *Deep Fashion* dataset [30] for concept extraction. Our method efficiently decomposes the input im-

Method	ImRS _μ	ImCS _μ	ImIDS _μ	ImCDS _μ
Ours	0.92	0.87	0.59	0.15
K-Means	0.98	0.88	0.56	0.05
Retrieval	0.99	0.83	0.54	0.01

Table 1. **Quantitative results.** As desired, our results show significantly greater diversity.

age into relevant domain-specific concepts, such as ‘women in denim outerwear,’ ‘striped tops,’ and ‘wide-leg grey bottoms.’ This capability has applications in fashion recommendation systems, virtual try-on solutions, and personalized shopping experiences, enhancing user interaction and retrieval accuracy in fashion-related tasks. Additional general and domain-specific results are in the supplementary material. These examples highlight how our problem and results differ from standard image retrieval. Instead of producing a single ranked list of globally similar images, the input image is implicitly decomposed into more abstract concepts.

Quantitative results. We compare our results to two baselines: (1) the retrieval method of [13, 36], which retrieves 60 images divided into three sets, and (2) *k*-means clustering (*k* = 3) in the embedding space, as a potential alternative for concept extraction. Table 1 shows our quantitative results averaged over all images in the diverse *LAION-Aesthetics* [41] dataset. Our method achieves the desired outcome: significantly higher diversity, both within and across concepts, while maintaining high relevance and consistency. This is because we aim at extracting images that share conceptual similarities, regardless of other visual elements, thereby ensuring diversity. As expected, the retrieval-based approach attains the highest Relevance Score, as it retrieves the most visually similar images, effectively capturing overall content rather than a diverse range of concepts. Meanwhile, *k*-means clustering ranks highest in consistency, as it selects the closest images within a small neighborhood. These results align with the human evaluation described next.

Human evaluation. We conducted a user study, as detailed in Section 4.2, to assess whether human intuition regarding concepts aligns with the concepts identified by our method. A total of 32 participants took part in the evaluation, including 15 females and 17 males. The participants’ ages ranged from 18 to 75 years. In this study, participants were presented with 21 sets of concepts derived from seven different input images from *COCO* [29]. *Evaluating consistency.* Given sets of images extracted by our algorithm, each representing a concept, 95% of participants recognized the images within each set as sharing a common concept. This result highlights the effectiveness of our algorithm in identifying meaningful concepts. *Evaluating relevance.* Now,

when presented with the input image, 79% of the participants agreed that the concept represented by the set is indeed present in the input image (62% fully contained, 17% partially contained). We compared this relevance to two baseline methods: (1) A random input image, where only 33% of participants found the concept relevant. (2) A retrieval algorithm based on object matching, where a subset of objects from the input image was selected, and images containing the same objects were retrieved. In this case, only 41% of participants found the retrieved images relevant. These results confirm that a concept is more than just a collection of objects. *Evaluating inner-concept diversity.* When presented with the input image alongside two retrieved sets—one from our concept-retrieval algorithm and one from classical retrieval—our algorithm outperforms the baseline by 14%, reflecting participants’ relative preference for our approach over retrieval baselines. *Evaluating cross-concept diversity.* When shown the input along with three sets of related concepts, 90% of participants agreed the sets represent different concepts. Among them, 67% stated all three sets were completely different, while 23% found two out of the three were different. These results confirm that our approach extracts relevant, consistent, and internally and externally diverse concepts. They also clearly demonstrate that a concept is more than the sum of its objects.

Ablation: key thresholds. Our algorithm uses several parameters, including the neighborhood size T , the reduced representation threshold τ , and the percentage of embeddings modified between iterations, with detailed tests provided in the supplementary material. Overall, our method is stable across a broad range of hyperparameter values. The chosen defaults provide a good balance between relevance, consistency, and both forms of diversity. For update percentage, as expected larger updates reduce relevance and consistency but increase diversity. Higher τ improves relevance but slightly reduces diversity and consistency. Thus, the parameters were empirically selected: the neighborhood size was set to $T = 0.25\sigma$ (Eq. 2), the reduced representation threshold was set to $\tau_1 = 0.25$ (Eq. 8), and the percentage of embeddings modified between iterations was 10%.

Generalization. The results presented in this paper utilize embeddings from ViT-H-14-DFF [13, 36], due to its strong zero-shot capabilities and the rich semantic structure of its embedding space. Our approach generalizes well across different vision models. In addition to ViT-H-14-DFF, we tested it with weaker embeddings from CLIP ViT-L/14 and DINO. As expected, retrieved concept quality decreases with weaker embeddings, as these models capture less semantic structure. Nevertheless, the method performs well across models, as demonstrated in the supplementary material.

Computational Efficiency: Our method relies on repeated nearest-neighbor searches, Gaussian Mixture Model

(GMM) fitting, and PCA computations. Approximate nearest neighbor search has complexity $O(n)$ [2], where n is the number of dataset images. Approximate PCA has complexity $O(nd)$ [18], where n is a subset of the dataset, d is the embedding dimension, and the number of principal components is small. Approximate GMM [10] has complexity $O(n)$, where n again refers to a subset. The overall complexity is thus bounded by $O(nd)$. The running time is 4.29 seconds per image when extracting three concepts on an AMD EPYC 7763 CPU.

Limitations. Our approach lacks user control over the extracted concepts, which may be particularly important for domain-specific applications. This is an intriguing direction for future research. Furthermore, the method might struggle when the concept is extremely rare in the dataset, as there may not be enough supporting samples to reliably separate it as a distinct mode.

6. Conclusion

This paper introduces the problem of image concept retrieval, which can be seen as a generalization of traditional image retrieval. We define the essential requirements for any method addressing this task and propose a novel approach for concept extraction based on these requirements. Our approach is built on two key observations: (1) the similarity distribution within the input’s neighborhood can be modeled as a bimodal Gaussian, and (2) certain neighbors clearly exhibit this structure. Our approach requires no training, which is crucial given the difficulty of obtaining ground truth. Additionally, we introduce new evaluation metrics tailored to this task. Both qualitative and quantitative results, supported by a human study, validate the effectiveness of our approach. As this is a new problem, several future directions are possible. First, incorporating control over which concepts are extracted. Second, extending our concept retrieval method to other embedding spaces, such as sentences or documents. Since the method relies on neighborhood similarity distributions, it should transfer naturally to such domains. Finally, we expect that dedicated datasets for this task will further support community benchmarking.

Acknowledgments. We gratefully acknowledge the support of the Israel Science Foundation (ISF) 2329/22.

References

- [1] Jacob Andreas, Dan Klein, and Sergey Levine. Learning with latent language. *arXiv preprint arXiv:1711.00482*, 2017. 2
- [2] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975. 8
- [3] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*, 2022. 3
- [4] Aditya Chattopadhyay, Ryan Pilgrim, and Rene Vidal. Information maximization perspective of orthogonal matching pursuit with applications to explainable ai. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [5] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019. 2
- [6] Wei Chen, Yang Liu, Weiping Wang, Erwin Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S Lew. Deep image retrieval: A survey. *arXiv preprint arXiv:2101.11282*, 1(3):6, 2021. 3
- [7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 3
- [8] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020. 2
- [9] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)*, 40(2):1–60, 2008. 3
- [10] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. 8
- [11] T Dharani and I Laurence Aroquiarij. A survey on content based image retrieval. In *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*, pages 485–490. IEEE, 2013. 1
- [12] Shiv Ram Dubey. A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2687–2704, 2021. 3
- [13] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023. 3, 6, 7, 8
- [14] Ruigang Fu, Biao Li, Yinghui Gao, and Ping Wang. Content-based image retrieval based on cnn and svm. In *2016 2nd IEEE international conference on computer and communications (ICCC)*, pages 638–642. IEEE, 2016. 2
- [15] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2
- [16] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023. 2
- [17] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019. 2
- [18] Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011. 8
- [19] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*, 2021. 3
- [20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 3
- [21] Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023. 2
- [22] Yangqing Jia, Joshua T Abbott, Joseph L Austerweil, Tom Griffiths, and Trevor Darrell. Visual concept learning: Combining machine vision and bayesian generalization on concept hierarchies. *Advances in Neural Information Processing Systems*, 26, 2013. 2
- [23] Shraddha S Kashid, Dattatray G Takale, Piyush P Gawali, Gopal B Deshmukh, Parikshit N Mahalle, Bipin Sule, Arati V Deshpande, and Bhausaheb S Salve. Optimizing content-based image retrieval system using convolutional neural network models. In *International Conference on Information and Communication Technology for Intelligent Systems*, pages 13–25. Springer, 2024. 2
- [24] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020. 2
- [25] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 2
- [26] Sharon Lee, Yunzhi Zhang, Shangzhe Wu, and Jiajun Wu. Language-informed visual concept learning. *arXiv preprint arXiv:2312.03587*, 2023. 2
- [27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Pro-*

- ceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. [3](#)
- [28] Yang Li, Shichao Kan, and Zhihai He. Unsupervised deep metric learning with transformed attention consistency and contrastive clustering loss. In *European Conference on Computer Vision*, pages 141–157. Springer, 2020. [2](#)
- [29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. [7](#)
- [30] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. [7](#)
- [31] Riccardo Majellaro, Jonathan Collu, Aske Plaat, and Thomas M Moerland. Explicitly disentangled representations in object-centric learning. *arXiv preprint arXiv:2401.10148*, 2024. [2](#)
- [32] Eric Margolis and Stephen Laurence. The ontology of concepts—abstract objects or mental representations? *Noûs*, 41(4):561–593, 2007. [1](#)
- [33] Meike Nauta, Ron Van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14933–14943, 2021. [2](#)
- [34] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023. [2](#)
- [35] Hamed Qazanfari, Mohammad M AlyanNezhadi, and Zohreh Nozari Khoshdaregi. Advancements in content-based image retrieval: A comprehensive survey of relevance feedback techniques. *arXiv preprint arXiv:2312.10089*, 2023. [1](#)
- [36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. [3](#), [6](#), [7](#), [8](#)
- [37] Sunita Rani, Geeta Kasana, and Shalini Batra. An efficient content based image retrieval framework using separable cnns. *Cluster Computing*, 28(1):56, 2025. [2](#)
- [38] Homayoun Rastegar and Davar Giveki. Designing a new deep convolutional neural network for content-based image retrieval with relevance feedback. *Computers and Electrical Engineering*, 106:108593, 2023.
- [39] Zakhayu Rian, Viny Christanti, and Janson Hendryli. Content-based image retrieval using convolutional neural networks. In *2019 IEEE International Conference on Signals and Systems (ICSigSys)*, pages 1–7. IEEE, 2019. [2](#)
- [40] Mehdi Safaee, Aryan Mikaeili, Or Patashnik, Daniel Cohen-Or, and Ali Mahdavi-Amiri. Clic: Concept learning in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6924–6933, 2024. [2](#)
- [41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. [7](#)
- [42] Chenming Shang, Shiji Zhou, Hengyuan Zhang, Xinzhe Ni, Yujiu Yang, and Yuwang Wang. Incremental residual concept bottleneck models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11030–11040, 2024. [2](#)
- [43] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 157–166, 2014. [3](#)
- [44] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197, 2023. [2](#)