Chapter 2

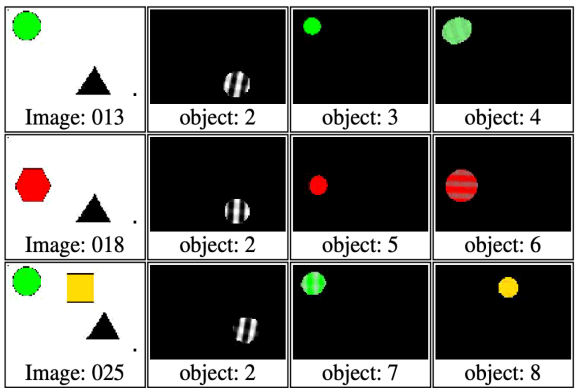
Related Work

The goal of this section is to introduce *image data mining*. We will discuss its definition as seen in our related work, which we will structure around three criteria: **(a)** discovery of visual structure (Sec. 2.1), **(b)** mining of visual structure (Sec. 2.2), and **(c)** human interpretable visual summarization (Sec. 2.3). These three criteria will be discussed further in their respective sections.

Image Data Mining. Already by the end of the previous century, [Omiecinski and Ordonez, 1998] saw Data Mining in Images as the discovery and grouping of repeated visual structure. Their approach, presented in Fig. 2.1 highlights both the motivation for such a task, as the real world is composed of repeatable objects (Fig. 2.1a), but also the significant limitations of Computer Vision techniques available at the time (Fig. 2.1b). However, what this example makes clear is that the purpose of image data mining, is the discovery of visual structure that can be used to summarize a dataset. Something that it doesn't make explicit, yet which is apparent, is that the visual structure that it aims to discover, is **new** in the sense of not being provided to the model in the form of prior knowledge (e.g., through annotation), and **interpretable** in the sense of being at the level of human understanding. In its core data mining answers a recurring need: we have taught models about data, but what did we learn from them?

Properties. In fact, the combination of these properties (novelty and interpretability) is important to make data-mining a distinct task. In the absence of the second property, the first simply describes what most deep, machine learning approaches do, either explicitly or implicitly: construct (hierarchical) detectors of statistically occurring signal patterns in order to minimize risk [Rosenblatt, 1958; Vapnik, 1998]. While

(a) Motivation (Fig. 1.)



(b) Results (Fig. 4.)

Figure 2.1: **Image Mining**. Both images come from [Omiecinski and Ordonez, 1998]. **(a)** Introductory figure of the paper motivating the need for data mining as all aerial images contain distinct objects which need to be discovered and counted. **(b)** Results of their proposed (blob detection) algorithm.

some research has been done, in discovering interpretable structure in the neurons of existing deep models [Olah et al., 2020], in their totality state-of-the-art deep learning models are hardly interpretable[1]. The existence of the first property is also crucial. Data mining should be useful for the production of epistemic surplus: both discovery and summarization when employed by non-expert users should be accurate and reliable enough so that their observations can become part of the human knowledge production loop [Sculley and Pasanek, 2008].

Chapter Organization. In our discussion above, we outlined three important properties of image data mining, which will be explored in this chapter. In Sec. 2.1 we will first focus on the way we can automatically discover visual structure in input images. Then in Sec 2.2 we will focus on mining, i.e., how we can produce new knowledge that is contextually informative. Finally, in Sec. 2.3 we will conclude by discussing different approaches of automatic image summarization that are human interpretable.

[1] In fact early statisticians thought that these two properties could eventually come together via a sleight of hand on interpretability, called sparsity. Even if, e.g., rule-based models were not interpretable, they were at least sparse. "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk", was an informal criticism [Dyson et al., 2004].

2.1 Discovering Visual Structure

This Section will summarize the discovery of visual structure through two main approaches: recognition based approaches and synthesis based approaches. In Sec. 2.1.1 we will first discuss a set of approaches that discover repeated visual structure across images using recognition methods that predict bounding boxes, segmentation masks, and correspondences. In Sec. 2.1.2 we will then discuss a set of approaches which can discover visual structure in scenes by learning to decompose and recompose them. Our work will mainly focus on the second set of approaches as they provide interpretable and compressed summaries of their input scenes.

Background. Discovery methods aim at localizing, statistically informative visual structure. Fundamental theories of gestalt visual perception [Wertheimer, 1938; Blake and Zisserman, 1987] show that humans tend to group consistent and continuous visual information. The most common form of visual structure discussed in the literature is that of objects [Roberts, 1963]. An object is typically a 3D structure, visible inside a 2D scene. However, as with any conceptual categories, one can derive multiple definitions. For example, an object can be defined as something separable from its background, something that can be removed by an action in 3D space, or something that localizes the instance of a semantic category (e.g., a "cat" or the letter "a"). Instead of trying to develop object detection methods from laws of perception [Shi and Malik, 2000], data-driven object detection explored how such intuitive definitions can be instilled into models through object detection datasets. For example COCO [Lin et al., 2014], annotated objects as things, i.e., countable elements, and background as stuff, i.e., uncountable elements [Adelson, 2001]. It included multiple common objects such as vehicles, animals, or furniture, and was purposed for supervised object recognition. On the other end, ClevrTex [Johnson et al., 2017; Karazija et al., 2021] saw objects as composable prototypes in 3D-space, rendered in 2D, creating a competitive synthetic dataset whose purpose was to evaluate unsupervised object recognition, and study its conditions of emergence. However, inside the literature, discovery of visual structure is of course not only tied to objects. For example, another common task is that of detecting semantic regions, as in the ADE dataset [Zhou et al., 2017b], motivated by human visual parsing, which could be a useful acquired skill, e.g., for robotics.

Data Representation. In all these datasets the common way visual structure is annotated for the purposes of evaluation, is through: **(a)** localization metadata, often in

(a) Cosegmentation (b) Mining Approximate Duplicates in Artworks

Figure 2.2: **Recognizing Repetition**. Identifying consistent repeating elements in images is a task fundamental to mining. **(a)** Cosegmentation can extract the common objects across different images (Fig. 5 [Joulin et al., 2010]). **(b)** Discovering rare duplicates across artwork collections by mining outliers with high similarity (Fig. 1 [Shen et al., 2019]).

the form of bounding boxes or segmentation masks; and **(b)** as a label, that identifies whether objects belong to the same class or characterize a different instance. Some datasets including those of panoptic segmentation [Kirillov et al., 2019], often contain additional instance annotation in the case where more than one object of the same class is present in the scene.

2.1.1 Discovery via Recognition

As discovery via recognition, we describe any method that for an input set of images, can produce a localization of one or more objects present in them. To do so, it simply needs to infer annotation metadata (e.g., bounding-boxes or segmentation masks) from pixels. For this reason, these methods are mostly addressed through regression and don't depend on a generative prior as those concerned by our work, which are presented in the next section 2.1.2.

Background. The core methodology lying behind discovery via recognition, is using similarities of pixels or regions of the input images to cluster them, into instances or categories. A fundamental approach was that of applying normalized cuts in graphs made from pixels intensities, inspired by the abovementioned gestalt theories of human visual perception [Shi and Malik, 2000]. Although this method worked on a single image, it made further sense to design approaches that perform segmentation in the context of whole datasets. One way was to convert the images of a dataset into image descriptors and discover frequently matching elements using techniques of text analysis by converting the descriptors into bags of words [Sivic et al., 2005]. Similarly,

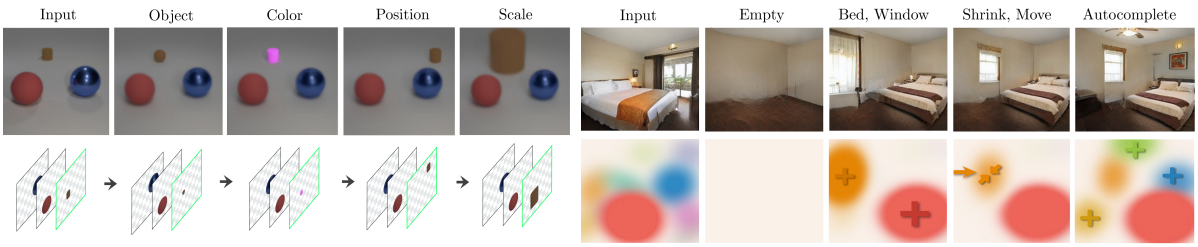
techniques of topic modelling were used to both classify and segment objects from input scenes [Cao and Fei-Fei, 2007]. However, scaling these types of approaches to datasets with multiple and diverse examples remains challenging as the size of the vocabulary can increase significantly, turning model-based approaches that operate directly at the level of images a better research direction.

Matching Pixels. Another set of works approached object discovery through co-segmentation. Given a small set of images that are intentionally selected to have a common element (e.g., a flower in Fig. 2.2a), the task of co-segmentation is to localize the visual support that is common between them [Joulin, 2012]. This task has seen a lot of progress, first approached with discriminative clustering [Joulin et al., 2010], then expanded to multiple classes [Joulin et al., 2012], then into performing object discovery across a dataset [Rubinstein et al., 2013], and finally in discovering multiple foreground objects [Chang and Wang, 2015].

In the Wild. Using feature pyramids, object discovery became more robust and was extended into the wild [Cho et al., 2015]. Later unsupervised discovery and localization was extended to approximate the solution of a combinatorial optimization problem with candidate proposals [Vo et al., 2019, 2020] and deemed to be the state of the art when using self-supervised features for computing proposal similarity and ranking proposals via PageRank [Vo et al., 2021]. One could even perform object localization in large datasets directly by using off-the-shelf transformer features, pretrained through self-supervised learning [Siméoni et al., 2021].

2.1.2 Discovery via Synthesis

As Discovery by Synthesis, we define any method that performs discovery by decomposing visual scenes through a visual synthesis prior, learned via reconstruction. We will organize this section as *explicit* and *implicit*. Explicit approaches directly hard-code priors in their architectures or bottleneck representations (Fig. 2.3a). In implicit approaches this form of decomposition emerges through training generative architectures, that can later be used in order to analyze input scenes (Fig. 2.3b). While the first part of our work, Chap. 3, is based on explicit discovery by synthesis as it uses a sprite-based deformable architecture prior, the second part, Chap. 4, builds on implicit discovery by synthesis as it relies on a conditional diffusion model.



(a) DTI-Sprites (b) BlobGAN

Figure 2.3: **Discovery via Synthesis**. An image can be decomposed into a latent representation that allows for its consistent analysis and high-level editing. **(a)** Explicit, sprite based via DTI-Sprites (Fig. 4 [Monnier et al., 2021]). **(b)** Implicit, latent based via BlobGAN (Fig. 3 [Epstein et al., 2022]).

Background. Discovery via synthesis is based on a historical research approach known as Analysis by Synthesis. Its goal is to analyze input scenes by learning to decompose them into high-level (semantic, templatic etc.) layers with the goal of reconstruction. This approach is often motivated by human understanding [Biederman, 1987] and from its origins it is associated to Bayesian learning [Yuille and Kersten, 2006]. Its fundamental assumption is that of inverse graphics, *i.e.*, that a set of input images can be decomposed through a process of decomposition and reconstruction into a common set of composable units, which [Biederman, 1987] called *geons*. This idea has been highly influential to multiple approaches of computer vision which may not directly associate themselves with Bayesian learning.

Explicit. Standard analysis by synthesis methods learned stochastic grammars of images [Zhu et al., 2007] and a part based decomposition [Zhu et al., 2010]. They went as far as exploiting the property of compositionality to perform clustering [Faktor and Irani, 2012]. Others learned to infer bottleneck assignments of texton tokens in order to perform texture synthesis [Zhu et al., 2009]. From this early set of works, it became evident that abstracting scenes into high-level representations can enable the ability of categorical and spatial reasoning. Following this motivation multiple later approaches revitalized this research area using standard neural methods [Greff et al., 2016, 2017] or more specialized spatial transformers [Jaderberg et al., 2015b; Dalca et al., 2019] and extended this for the discovery of multiple objects in 2D [Greff et al., 2019a; Monnier et al., 2021; Locatello et al., 2020], 3D [Yao et al., 2018; Deprelle et al., 2019; Loiseau et al., 2024; Monnier et al., 2023], and video data [Kosiorek et al., 2018; Wu et al., 2023] where they were even used in order to learn intuitive physics [Chen et al., 2022]. Due to their high-level priors, they can also help to infer accurately occluded objects and

thus have been found beneficial for unsupervised classification [Kosiorek et al., 2019] and robust pose estimation [Angtian et al., 2021].

Implicit. As implicit discovery by synthesis methods we can first think of those that arrive at a disentaglement of the latent space of generative models, e.g., [Higgins et al., 2017; Peebles et al., 2020] which show that parts of the input scene can be transformed by amplifying or suppressing learned directions of the latent space. Interestingly [Epstein et al., 2022] arrived at a spatial decomposition of scenes like bedrooms, by learning to generate these images starting from localized blob-sized latents. More recently, methods based on diffusion models, split the image synthesis onto individual denoising steps that if manipulated independently can allow forms of structured decomposition and manipulation, either with text [Brooks et al., 2023] (see Fig. 2.3b), or other controls such as keypoints or masks [Epstein et al., 2023; Luo et al., 2024]. [Epstein et al., 2024] learns to generate scenes with disentagled decompositions by rendering multiple Nerfs and optimizing their composited appearance through a diffusion model. These methods can also be adapted for discovery, both for various tasks such as object detection [Ma et al., 2023], amodal segmentation of occluded objects via inpainting [Ozguroglu et al., 2024], and even disentanglement of the random and cyclic effects in time-lapse videos [Härkönen et al., 2022].

2.2 Mining Informative Visual Structure

This Section will discuss the literature developed around mining of informative visual structure. In Sec. 2.2.1 we will first discuss a set of algorithms that perform discriminative clustering, a task that is foundational for mining. Then, in Sec. 2.2 we will focus on approaches that mine visual structure by identifying visual elements that are informative of their label inside the context of their input datasets. Finally, in Sec. 2.2.3, we will discuss model-based interpretability approaches which try to interpret important components of the input data by interpreting and attributing network predictions.

2.2.1 Discriminative Clustering

As we will see next in Sec. 2.2, the most common popular formalization image data mining, is that of discriminative mining. It is based on subjecting images or image parts, to a variant of clustering, known as discriminative clustering. In standard clustering, elements are grouped in an unsupervised way through an approximation of

a cluster assignment which minimizes the distance of the cluster centroids to their respective data points [Lloyd, 1982]. Discriminative clustering formalizes clustering via supervised learning, allowing the introduction of constraints and label supervision [Bridle et al., 1991; Bach and Harchaoui, 2007; de la Torre and Kanade, 2009]. This allows to use weak-labels that form clusters which group the most discriminative elements of a class, in respect to all others. Attempts to formalize this in the literature, have been multiple: clustering via mutual information [Bridle et al., 1991], supervised clustering [Bach and Harchaoui, 2007], and discriminative k-means [Ye et al., 2007b], all of which we will discuss in the following paragraphs.

Mutual Information. An early unsuccessful attempt for discriminative clustering used a classifier to predict cluster assignments on pseudo labels [Bridle et al., 1991]. In order to make clusters more discriminative, the paper used a mutual information objective $I(X;Y) = H(X) - H(X \mid Y)$, between images X and their predicted cluster labels Y. Here $H(X)$ stands for entropy, while $H(X \mid Y)$ stands for conditional entropy. Maximizing $I(X;Y)$, corresponds to maximizing $H(X)$ which opts for "fairness" while minimizing $H(X \mid Y)$ opts for "firmness" [Bridle et al., 1991]. Yet, directly optimizing this objective can lead to suboptimal results during optimization, as clarified by later works [Ohl et al., 2022]. What this approach introduces, however, is **(a)** that clustering can be integrated inside a framework of classification, but more importantly **(b)** that the learned clusters should be discriminative for their target class while opting for frequent elements. In fact, in Chap. 4 we show how extracting and then clustering patches that maximize mutual information $I(X;Y)$ using a strong pretrained prior, can lead to strong results in visually summarizing class labels across a variety of datasets.

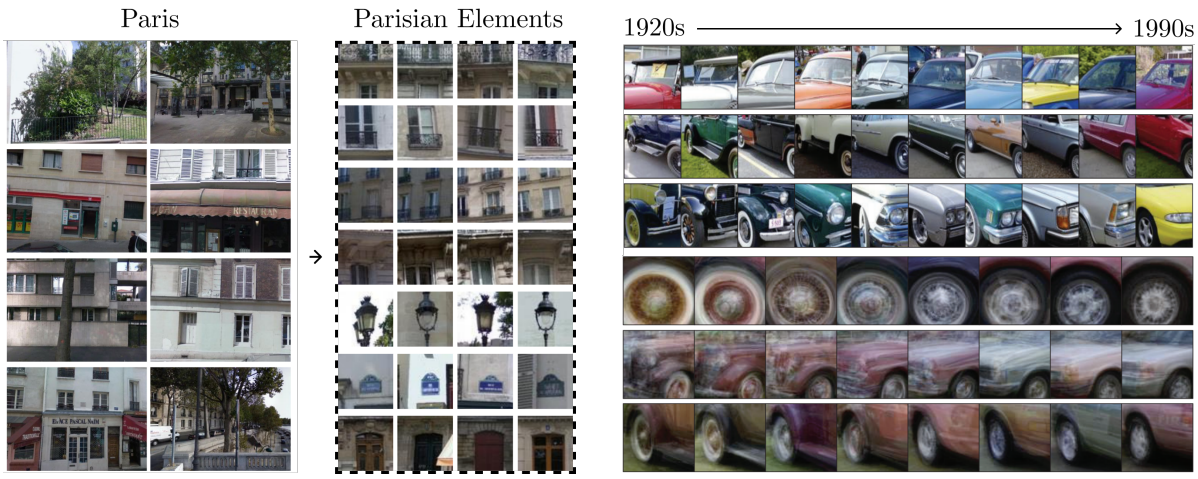
Supervised Clustering. Early successful works formalized clustering as a convex integer programming using labels assigned through SVMs [Xu et al., 2004]. Later, an approach known as DIFFRAC [Bach and Harchaoui, 2007] formalized clustering directly as a supervised problem trained with MSE regression, where the labels are also optimized alongside the projections of data points. Using certain very trivial clustering constraints that would control the size of clusters and would ensure that cluster assignments of points should sum to 1, DIFFRAC becomes much more robust to noise than K-Means [Lloyd, 1982]. Except from cluster constraints, this framework opened the possibility of using weak labels, for example of continuity across video of frames [Bojanowski et al., 2014] or training with any ratio of labeled and unlabeled data [Jones et al., 2022]. As we discussed in Sec. 2.1.1, DIFFRAC was the first basis

cosegmentation [Joulin et al., 2010], but it has also been used for the discovery of "mid-level" patches which improve classification performance [Sun and Ponce, 2013].

Discriminative K-Means. A later set of approaches [Ye et al., 2007a; Ding and Li, 2007; de la Torre and Kanade, 2009; Ye et al., 2007b] used ground truth labels to implement discriminative clustering by combining it with Fisher Linear Discriminant Analysis (LDA) [Mika et al., 1999] transforming it into a framework similar to K-Means [Lloyd, 1982]. In order to discover "mid-level" level visual patches further works realized that one cannot rely on a two stage approach of simply clustering and then performing classification to detect discriminative elements, as "it is infeasible to use a discovery dataset large enough to be representative of the entire visual world" [Singh et al., 2012]. Instead, they relied on negatives to turn the "classification" step into one of "detection", by turning "clusters into detectors" using as a positive dataset of a target class and a negative dataset of random images outside that class. This technique has also been adapted to serve as a proxy for object detection [Bansal et al., 2015]. While the idea of using negative data was already a common technique even for methods as old as face recognition [Viola and Jones, 2001], the idea of using this as a proxy task to perform unsupervised learning, is reminiscent of a later set of approaches known as self-supervised learning [Uelwer et al., 2023]. Yet, a benefit of older methods was that in order to perform detection they remained close to a set of ground truth image patches, allowing for image mining and interpretability.

2.2.2 Image Data Mining

As discussed in Sec. 1.3, Image Data Mining is the main focus of this thesis. Image data mining can involve mining the ("mid-level") vocabulary of an image dataset that best represents existing labels, as well as mining new categories starting from existing ones. While both of our methods mine visual vocabularies, Chap. 3 focuses on how these vocabularies can be used to track minute changes in visual structure that validate and inform existing typologies, while Chap. 4 focuses on the challenge of extracting these vocabularies from versatile in-the-wild datasets. In the first paragraph of this Section we focus on discussing approaches which summarize an input dataset as a visual vocabulary of "mid-level" visual structure, while the second discusses a complementary research task that instead learn to discover new categories by learning to categorize existing ones.



(a) Single Class (b) Across Classes

Figure 2.4: **Mining Patch Summaries**. Images can be summarized into elements discriminative of certain class. **(a)** Mining elements that summarize the label "Paris" (Fig. 6 [Doersch et al., 2012]). **(b)** Mining Elements that are consistent across different decades of car design (Fig. 7,8 [Lee et al., 2013]). Note that averages are used in order to show the consistency of a class.

Mining Visual Vocabularies. As its name may suggest, what mining tries to discover is rare and valuable. Ideally one would have an algorithm that based on a question provided by a user could output the most important elements of the input dataset, by discriminating them against other unimportant elements. When the search space of attributes is fixed and already known, as in the case of Fashion [Chen et al., 2015; Matzen et al., 2017], simply using a classifier should suffice. For example standard techniques such as frequent pattern mining [Han et al., 2000] can be used to mine the most frequent cloth attributes after they have been extracted from an image collection [Chen and Luo, 2017]. However, it becomes apparent that when attributes are not present one has to decide what elements used to discriminate from the input dataset. For example, if one would like to find elements that are related to 'manuscripts' from a dataset containing only 'manuscripts' and 'natural images', the task becomes relatively trivial as even the background of the page could be a high ranking potential candidate. Thus, most mining methods properly structure their datasets to have overlapping attributes, and rely on Discriminative Clustering (Sec. 2.2.1) to find visual structure $x \in \mathcal{X}$ (e.g., patches that come from an image space $\mathcal{X}$) that are both frequent $p(x)$ and discriminative $p(y \mid x)$ of a certain class y. This idea has been highly successful in performing mining in geographical data [Doersch et al., 2012] (see Fig. 2.4a), yearbooks [Ginosar et al., 2017] (see Fig. 2.6a), (New York) fashion [Hidayati et al., 2014], via discriminative k-means [Singh et al., 2012] which we discussed in the previous

Sec. 2.2.1. This technique was also used to track correspondences of discriminative elements across time in cars [Lee et al., 2013] (see Fig. 2.4b) and architecture [Lee et al., 2015a]. Our work directly contributes to this track of research, by relying on strong priors that unlike [Singh et al., 2012] can optimize this procedure by splitting it into a two-step approach. For tracking minute changes of characters, we first learn (cluster) and then compare prototypes in Chap. 3, while for finding typical patches of images, we first mine and then cluster the most discriminative elements in Chap. 4. This pipeline is in fact similar to [Matzen and Snavely, 2015], which uses a foveated filter to discover and then cluster highly discriminative regions of input images. Such two-stage approaches offer a significant improvement to the quadratic complexity of discriminative clustering. Note, that although not explored by our work, hybrid approaches between discrete and continuous mining still exist. For example [Li et al., 2015], showed how binning CNN activations can create a novel semantic vocabulary of attributes onto which frequent pattern mining could be performed.

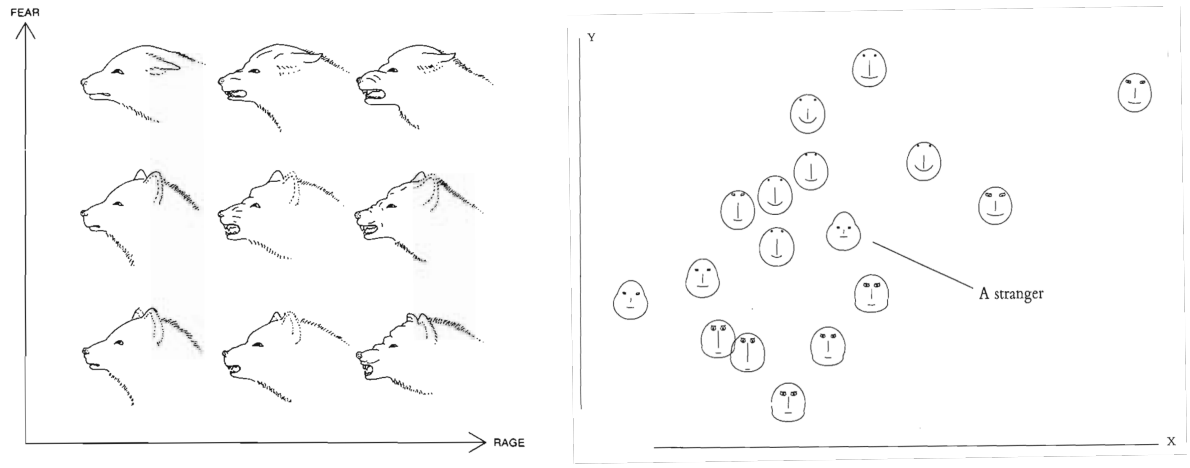
Mining new categories. One task relevant to data mining is that of category discovery. Its objective is, given an existing semi-annotated training set of images to learn a way of representing images at training time such that during test time a model can effectively cluster images into previously unassigned categories [Troisemaine et al., 2023]. Initially two stage approaches, would split data into pseudo-labels [Hsu et al., 2018] or latent space coordinates [Han et al., 2019] during training, and then perform clustering on this intermediary representation. One stage approaches, both learn simultaneously to perform classification on the annotated part of the dataset and on the extracted pseudo-labels of the unsupervised part [Bendale and Boult, 2016; Zhong et al., 2021]. Other works, motivated by the continuous nature of categorical discovery, develop a model that doesn't overfit on known classes, while uses them to learn effective representations [Vaze et al., 2022; Rizve et al., 2022]. Lots of category discovery methods assume mono-categorical image datasets similar to ImageNet [Deng et al., 2009] or iNaturalist [Van Horn et al., 2018], but some have extended novel category discovery of objects in scenes [Zheng et al., 2022; Fomenko et al., 2022; Bharadwaj et al., 2025; Feng et al., 2024]. Note, that in comparison to mining, new categories do not assume an association to a certain parent label and can be clustered independently of any hierarchy. This makes these methods different from our approaches discussed in Chap. 3,4, even if they are still relevant.

2.2.3 Model-centric Interpretations.

A complementary set of approaches that is not explored in our work, tries to mine visual structure by mining emergent neuron activations of neural networks trained on our data. Here, we discuss three approaches to extract model-centric interpretations, starting from early saliency based interpretability, extending them to mechanistic interpretability, and to data attribution.

Background. As interpretable architectures were historically hard to scale a set of approaches in the literature focused in producing interpretations of the outputs of well-performing deep neural networks. One common technique is to produce saliency maps that try to map parts of the input to output predictions for both CNNs [Selvaraju et al., 2017] and transformers [Chefer et al., 2021]. More than visualizing why a network makes a certain prediction in respect to the input image, other approaches realized that certain neurons of the same network were causally responsible for this prediction. For example, visualizing neuron activations related to certain concepts can reveal the most predictive attributes for certain classes, like tires for cars [Olah et al., 2020] or track how their visual representation varies in a multimodal manner [Goh et al., 2021].

Mechanistic Interpretability. Inspired by neuroscience, a parallel set of approaches discovered that object detectors can emerge inside CNNs trained for scene classification [Zhou et al., 2015] and found that certain neurons in GANs were causally related with generating semantic parts of output scenes, like trees [Bau et al., 2019]. This gave rise to the field of *mechanistic* interpretability which aims to "reverse engineering the computational mechanisms and representations learned by neural networks into human-understandable algorithms and concepts to provide a granular, causal understanding" [Bereska and Gavves, 2024]. For example, one can mine common units across models [Dravid et al., 2023] that show some universal properties of discovering visual structure, while others can locate the existence of sparse neurons, like a "Paris" neuron in large VLMs via sparse autoencoders [Lieberum et al., 2024]. More relevant to our work, a sequence of approaches uses sparse autoencoders (SAE) on a CLIP vision encoder [Radford et al., 2021] to discover and annotate emerging mono-semantic neurons [Fry, 2024; Rao et al., 2024; Pach et al., 2025] or decompose polysemantic neurons by clustering the visual circuits that correspond to disjoint classes [Dreyer et al., 2024]. While these approaches tend to be a form of data mining, they have a mixed scope on whether their goal is to analyze the neural network they study or the data that it has been trained on.



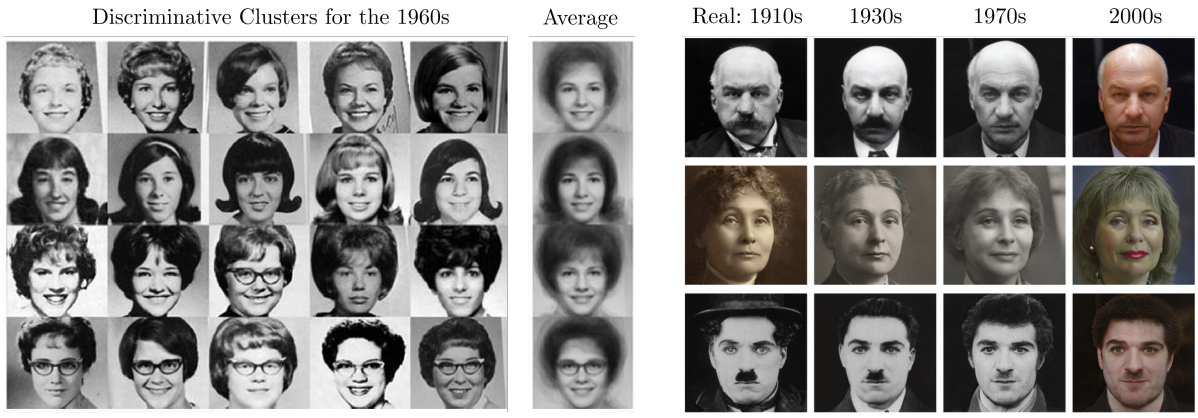
(a) Rational (b) Affective

Figure 2.5: **Human knowledge transmission through visual summarization.** Visual communication of quantitative data [Tufte and Graves-Morris, 1983] is an characteristic way by which humans can be trained to organize information **(a)** or detect abnormalities **(b)**. For example one can really easily understand the effect of emotions of fear and rage in the visual characteristics of an animal as seen on the left [Zeeman, 1976], reminiscent of principal component visualization on the latent space of generative models [Peebles et al., 2020]. At the same time these representations are often a reflection of the underlying human perceptual priorities as is the case with Chernhoff faces [Chernoff, 1973] displayed on the right which are supposed to allow plotting k-dimensional data in a way that abnormalities could be easily be spotted through universal affective cues.

Data Attribution. Trying to pose this question more in regard to a dataset, another set of approaches track tried to measure how predictions are influenced by a set of training data points [Koh and Liang, 2017]. This gave rise to the field of data attribution, for which several methods were proposed that tried to measure the influence of a data point to a model, via personalization methods [Wang et al., 2023], unlearning [Wang et al., 2024], or TRAK-based linearization [Georgiev et al., 2023]. While data-attribution still reveals how training data influences a network to make certain predictions it doesn't focus on discovering new elements in that data.

2.3 Summarizing Informative Visual Structure

As we pointed out in the introduction, in order for knowledge extracted through neural networks to qualify as human knowledge it should function as a form of summarization. While this process is often an integral part of the mining algorithm, for example in the case of discriminative clustering, in this section we will isolate it



(a) Single Label (b) Cross Label

Figure 2.6: **Mining Average Summaries**. One can present mining results with a single image through proper aggregation. (**a**) Discriminative clusters of portraits from the 1960s. Unlike 2.4, here the summaries help reveal cosmetic patterns as is the case with *tailwhip* haircut on the second line (Fig. 9. [Ginosar et al., 2017]). **(**b) Disentangling time from face layout allows a consistent translation of faces, effectively compresses the understanding of change (Fig. 1 [Chen et al., 2023]).

and focus solely on the different ways of visual presentation that have been used for automatic image data summarization. We will first discuss the most relevant form of summarization to our work that compresses visual collections into image summaries; then we will discuss text based summarization, a recent set of approaches that use large multimodal neural networks to provide textual summaries of image collections; finally we will conclude by discussing Exploratory Data Analysis (EDA), that is a more unstructured way of image summarization that allow a user to intuitively navigate large image collections.

Human Knowledge Mining. Visual transmission of knowledge is a core element of human intellectual culture. Academic journal articles, are a characteristic bottleneck through which humans get to visually summarize their findings in a process of producing knowledge. In his book, the "Visual Display of quantitative information" [Tufte and Graves-Morris, 1983], Edward Tufte, investigates multiple innovative ways that humans have devised to transmit visual information, two examples of which we show in Fig. 2.5. What is important to note is that often these visual summaries are even identical to what we consider as knowledge. Although a lot of literature methods that we will describe in this section do not always innovate in the way that Tufte's examples do, they all invent ways to navigate the trade-off between knowledge extraction and visual summarization.

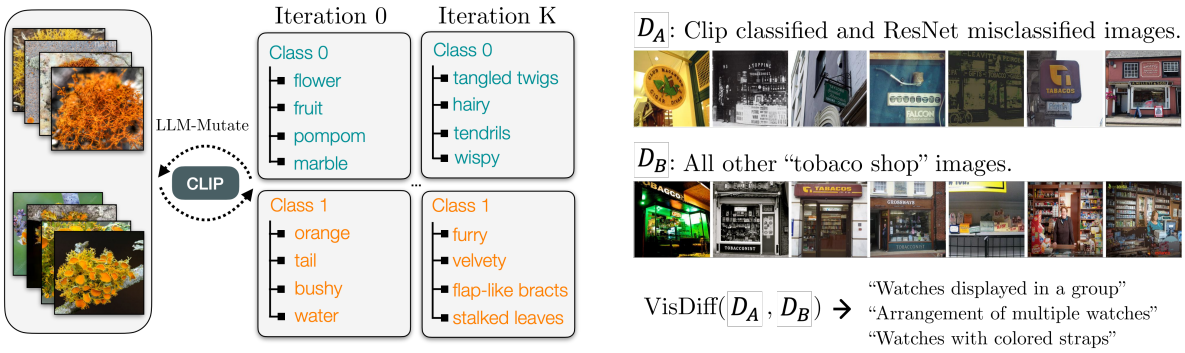
2.3.1 Visual Summaries

The goal of visually summarizing images is to reveal mined concepts through a single visual representation. These images can often serve as a scientific artifact, as is the case with the most important figures of journal publications. When the mined attributes are discrete, as is for example the case with fashion attributes, one can for example compress information into a plot that simply tracks quantitative changes and highlights the most important trends [Chen et al., 2015; Matzen et al., 2017]. However, when there is no known category to describe the mined concept, it often needs to be compressed through a small set of images. There summarization is often performed by visualizing the top patches representative of a certain cluster or via image averages which by computing the mean of a group of images can reveal persistent trends [Doersch et al., 2012; Lee et al., 2013; Ginosar et al., 2017; Matzen et al., 2017] (e.g., see Fig. 2.6a).

On Chap. 3 we show how these average summaries can be learned through a differentiable method, that not only provides results that are human interpretable, but which also enable interpretable quantitative comparison across classes. Complementary, one can understand target images by visualizing attribution of both pre-defined [Kiapour et al., 2014] and discovered attributes [Doersch et al., 2013]. Using recent techniques of generative modeling, one can also translate images across different contexts, for example by translating an input portrait across time [Chen et al., 2023], as seen in Fig. 2.6b. As we will show in Sec. 4.5.1, this can further allow us to discover transformation trends captured inside the latent space of a generative model trained on a geographic image collection. Instead, a similar approach visualizes latent averages to convey the average appearance of an area in a city like Paris or New York [Feng et al., 2025].

2.3.2 Text based summarization

Through the advent of recent large multimodal foundation models [Radford et al., 2021; Achiam et al., 2023], deep-learning models have become increasingly capable of understanding images in relation to text. This has enabled a new trend of mining, where the goal is to generate textual summaries of image collections. This form of summarization can be used to produce categorical descriptions of data that can serve as text classifiers [Chiquier et al., 2024] (see Fig. 2.7a) or to describe differences between datasets [Dunlap et al., 2024] (see Fig. 2.7b), or even summarize the mechanistic functionality of foundation models [Gandelsman et al., 2024; Shaham et al., 2024]. However, as the goal of data mining is often the discovery of new concepts, this



(a) Labeling Images (b) Naming Differences

Figure 2.7: **Minining Text Summaries**. Models that consistently connect images with text, can be used to summarize image collections into text. **(a)** Classifiers can be evolved to classify species of plants through a small set of interpretable text attributes (Fig. 1. [Chiquier et al., 2025]). **(b)** Differences between datasets can be summarized through text by sampling and ranking well correlated statistical proposals (Fig. 13. [Dunlap et al., 2024]).

strand of approaches is often limited by those that are already properly described by language or otherwise the produced textual summaries can end up being vague or ambiguous. Text based summarization is in fact more similar to Aristotle's definition of categories [Studtmann, 2024], while visual summaries are closer to Eleanor Rosch's prototypes [Rosch, 1973] as we discussed in Sec. 1.2.

2.3.3 Exploratory Data Analysis

Image Data Mining can be seen as a special case of Visual Data Mining [Keim et al., 2002; Simoff et al., 2008], which tries to aid a comprehensive understanding of data through an algorithmically produced visual support as the one served by the popular t-SNE visualization [Van der Maaten and Hinton, 2008] of high-dimensional data. They are both special cases of a process called Exploratory Data Analysis (EDA) [Tukey, 1977], where through different statistical processes and forms of visualization a user can explore a data collection to arrive into scientific observations. The main difference of EDA to data mining, is that EDA is never conclusive, yet while being open-ended it compresses information in a way that a user can still arrive to meaningful observations on a target dataset. For example, in digital sociology one can visualize a twitter retweet network and identify the greatest influencers of a network through an intuitive user-interface developed for non-expert users [Pournaki et al., 2020] (see Fig. 2.8a). In images, average explorer [Zhu et al., 2014] (see Fig. 2.8b) provides an interactive interface for image collections using average images computed through

(a) Twitter Explorer (b) Average Explorer

Figure 2.8: **Exploratory Data Analysis**. Often a tool is needed for the structural visual navigation of large collections of data. **(a)** Visual EDA exploring a twitter retweet network with a force atlas layout [Pournaki et al., 2020]. **(b)** Image EDA exploring collections of cats through correspondences and averages [Zhu et al., 2014]. Notice that both methods hint to representations that the user is meant to discover, yet they don't provide a single answer.

correspondences from user inserted keypoints. Similarly, [Rematas et al., 2015] extracts linearly discriminative mid-level visual features and connects them through a browsable structure-specific interface. Note, that visual data mining can often be the process that is performed at each step of EDA, as is what happens on each click of AverageExplorer [Zhu et al., 2014] and in this way the two processes are often intertwined.