Image Mining via Synthesis

Ioannis Siglidis

▶ **To cite this version:**

Ioannis Siglidis. Image Mining via Synthesis. Computer science. École des Ponts ParisTech, 2025. English. ⟨NNT : 2025ENPC0012⟩. ⟨tel-05294120⟩

HAL Id: tel-05294120

https://theses.hal.science/tel-05294120v1

Submitted on 2 Oct 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

ÉCOLE NATIONALE DES PONTS ET CHAUSSÉES

IP PARIS

THÈSE DE DOCTORAT

de l'École nationale des ponts et chaussées

Image Mining via Synthesis

École Doctorale MSTIC, N°532

Signal, Image, Automatique

Thèse préparée au LIGM - IMAGINE, École des Ponts

Thèse soutenue le 16 mai 2025, par **Ioannis Siglidis**

Composition du jury:

Alexei A. Efros Professeur, UC Berkeley	*Président*
Jean Ponce Professeur, École Normale Supérieure	*Rapporteur*
Josef Sivic Chercheur distingué, Czech Technical University	*Rapporteur*
Shiry Ginosar Professeure assistante, Toyota Institute of Technology	*Examinatrice*
Hadar Averbuch-Elor Professeure assistante, Cornell University	*Examinatrice*
Mathieu Aubry Directeur de recherche, École des Ponts	*Directeur de thèse*

One must still have chaos in oneself to give birth to a dancing star.

Abstract

Image archives contain lots of hidden knowledge that researchers in the digital humanities would like to discover at scale. Much of this knowledge is visual, which makes challenging to either describe it using textual descriptions, or to manually ground existing descriptions to visual evidence. Given collections of images with general predefined classes, for example a type of script or the name of a country, the goal of this thesis is to develop machine learning approaches that can mine informative visual structure hiding behind these image collections. In recent years, a great variety of synthesis methods have managed to incorporate weak human labels into unsupervised synthesis methods, capturing a more visually accurate representation of their input images. The key idea of this thesis is to repurpose their abilities of image composition towards reliable human-interpretable visual summarization. Our work focuses on two specific problems of image data mining.

The first problem, is to summarize and help refine existing typologies of handwritten characters. Character morphology has been central to the field of palaeography, where existing typologies are described through textual descriptions. The reliance of the literature on such descriptions doesn't allow a quantitative methodology that could ground qualitative observation and exploration across collections of historical manuscripts. Our first contribution, the "Learnable Typewriter," is a novel approach that learns an explicit decomposition of a manuscript's text lines into small images of characters called sprites. Learning to reconstruct input text lines by composing sprites via differentiable transformations and weak-supervision, our method produces a concrete and interpretable summarization of a character's morphology. This enables palaeographers to perform quantitative comparison that stands very close to the concrete visual evidence of a script, captured through sprites. We validate this approach on various datasets of printed modern and historical fonts and ciphered texts, as well as provide two case studies of palaeographic analysis of medieval manuscripts.

The second problem, is to summarize the visual structure that makes images, typical of their assigned label. Analyzing historical, or cultural image datasets, by counting the presence of predefined attributes often provides very general observations that can't focus on the visual details that are typical of an input label. In our second con-

tribution, "Diffusion Models as Data Mining Tools", we leverage the abstract and scalable compositional synthesis capabilities of diffusion models to mine typical visual vocabularies from labeled datasets. Through a "typicality" score that is based on the influence of a label to the denoising performance of the diffusion model, we rank and then cluster a vocabulary of visual structure that is typical to that label. Unlike prior approaches that discover frequent and discriminative patches through pairwise matching, our diffusion based approach has linear complexity, more robust representations and can scale to versatile datasets of cars, portraits, geographical images, or scenes, that range from thousands to millions of images. As our approach is based on a diffusion model, it also allows us to directly translate images across labels in order to mine elements that are typical across all classes, as well as to interpret the sampling bias of the diffusion model itself.

Keywords: Image mining, Analysis by Synthesis, Digital Humanities, Sprite Methods, Diffusion Models, Deep Learning.

Résumé

Les archives d'images contiennent une grande quantité de connaissances visuelles cachées que les chercheuses en humanités numériques souhaitent découvrir à grande échelle. Une grande partie de ces connaissances est difficile à décrire par des descriptions textuelles, ou à fonder manuellement sur des preuves visuelles. Étant donné des collections d'images accompagnées de catégories générales prédéfinies, telles qu'un type d'écriture ou le nom d'un pays, l'objectif de cette thèse est de développer des approches d'apprentissage automatique permettant d'extraire la structure visuelle informative cachée derrière ces ensembles. Ces dernières années, de nombreuses méthodes de synthèse ont réussi à intégrer des labels humaines faibles dans des approches non supervisées, capturant une représentation plus fidèle et plus précise de leurs images d'entrée. L'idée principale de cette thèse est de réorienter ces capacités de composition d'images vers un résumé visuel interprétable et fiable par l'homme. Notre travail se concentre sur deux problèmes spécifiques liés à l'analyse de données visuelles.

Le premier problème consiste à résumer et à affiner les typologies existantes des caractères manuscrits. La morphologie des caractères est au cœur du champ de la paléographie, où les typologies sont traditionnellement définies à l'aide de descriptions textuelles. Cette dépendance empêche l'adoption de méthodes quantitatives pouvant fonder l'observation qualitative et faciliter l'exploration des manuscrits historiques à l'échelle. Notre première contribution, le "Learnable Typewriter", propose une approche innovante qui apprend une décomposition explicite des lignes de texte d'un manuscrit en petites images appelées sprites. En apprenant à reconstruire les lignes de texte d'entrée par la composition de sprites à l'aide de transformations différentiables sous supervision faible, notre méthode produit un résumé concret et interprétable de la morphologie des caractères. Cela permet aux paléographes de mener des comparaisons quantitatives directement fondées sur les preuves visuelles concrètes capturées par les sprites. Nous validons cette approche sur plusieurs ensembles de données de polices imprimées modernes et historiques, ainsi que sur des textes chiffrés, et nous présentons deux études de cas portant sur des manuscrits médiévaux.

Le deuxième problème porte sur la synthèse de la structure visuelle qui rend certaines images typiques de leur étiquette associée. L'analyse des ensembles d'images historiques ou culturelles à travers le comptage d'attributs prédéfinis fournit souvent des observations générales, sans saisir les détails visuels propres à une étiquette donnée. Dans notre deuxième contribution, intitulée "Diffusion Models as Data Mining Tools", nous exploitons les capacités abstraites et scalables de composition visuelle des modèles de diffusion pour extraire des vocabulaires visuels typiques à partir d'ensembles étiquetés. Grâce à un score de "typicalité" qui mesure l'influence d'une étiquette sur la performance de débruitage du modèle de diffusion, nous classons puis regroupons les éléments visuels les plus représentatifs d'une étiquette donnée. Contrairement aux méthodes antérieures, qui reposent sur des appariements par paires pour découvrir des fragments fréquents et discriminants, notre approche fondée sur la diffusion présente une complexité linéaire, des représentations plus robustes et peut être appliquée à des ensembles d'images très diversifiés — voitures, portraits, images géographiques ou scènes — allant de milliers à plusieurs millions d'images. Étant fondée sur un modèle de diffusion, notre approche permet également de traduire directement une image d'une étiquette à une autre, afin d'identifier les éléments co-typiques à plusieurs classes et d'interpréter le biais d'échantillonnage du modèle lui-même.

Mots clés : Fouille d'Images, Analyse par synthèse, Humanités numériques, Méthodes de Sprite, Modèles de diffusion, Apprentissage profond.

Acknowledgements

> We are all on earth to help others.
> What I can't figure out is what the
> others are here for.
>
> Joke by W. H. Auden.[1]

I'm not sure how I feel about a PhD that lasted three years; I could have done another one! I won't lie, it was painful (and often is) but always for the wrong reasons: alienation, competition, engineers, Paris (this heaven and hell), war; you name it! Yet, I can't but owe my existence to a certain set of people and institutions that helped me reside even temporarily into the imaginary. Don't be fooled: humans are flawed! However, what makes them unique is their mean and their variance [Sapolsky, 2018]. Below, I mine the mean plus the variance.

Administrative. Let's start with the people who devote their time supporting other people bring abstract ideas into papers. Cooks of our *cantine* and employees at our cafeteria, nurturing us with warm meals and (occasionally) smiles. Isabelle and Stephanie, for all their administrative support to my multiple *missions*. My jury members: Jean Ponce, Josef Sivic, Hadar Averbuch-Elor, for inspiring my research, committing their time and providing thoughtful comments to help me reinforce this (vicious) circulation of intelligence. To Karteek Alahari and Marie-Pierre Beal for being in my monitoring committee and spending time making sure that my thesis was on the right track.

Lab. "**Parisians**": Sonat for being caring; Syrine for art-house movies and nerdy papers; Nicolas D. for sharing publications, asking questions, and bringing water; David for being there equally for his colleagues as for his papers; Renauld for always greeting me with a smile; Loic for being egalitarian; Vincent for being exemplary of how to be successful without being toxic, in academia; Antoine for being kind and motivating; Nermin for accepting and for listening; Robin for following your

[1] Recited by Marvin Minsky in "How To Solve All The Problems In The World" at *ideaCity03*.

vocation; Romain for classy soirées; Elliot for (sometimes) being an otter; my office mates Nguyen, Shiyao for baring and eating with me; Tom for being the best travel buddy during our ten-day voyage in the new world. "**BAIRs**": To Yutong for being the spirit of a cat but "scaled the hell up!"; Yossi for sparse disentanglement; Assaf for the idempotency of research; Aleks for his friendship; Grace for her aid in data, code and insights; and to so many more!

World: Friends. To the friends I made along the way. Those that I don't see as often as they deserve: Diego and Alejandro for writers' obsession and witty sarcasm; Betsy for her brain; Emily for her intelligence (and care); Veronica for her worlds; Jenn for semantic radiation; Sonia for AI mysticism; Tyler F. for his moustache occluded smile; Tyler B. for hope in higher institutions; Segolene for sparks and fireworks; Ludwig for constantly turning life into a play; Georgy for god and robots; and so many more, who made my story of immigration feel like an oddyssey. To closer ones: Nikos for immeasurable support and for being human; Armin for nurture and for being the sweetest output of a razor; Marilena for making the unconsious-conscious and stability; Bessy for queer and lots of laughs; Frosso for dignity and lots of laughs; Simon for freedom and lots of laughs; Monse for being a form of Gaia.

Athens: Friends and Mentors. My friends from Athens: Manos for the importance of intellectual loitering and micro-theories; Gregoris for the seriousness of philosophy and being my first advisor; Kosmas for net-utopia irl; Peppa for the art of pulling strings; Buzzy for good bad taste; Alexandra for love; Anastasia for care; Chrysa for honesty in coolness (and /ei/); Diana for being brave; Yorgos for the passion to exist. Elders: Marc for being cool a.f.; Christina for being competent a.f.; my cousin Themis for transforming hardship into wisdom (a smile at a time); my godfather Tasos, for the love of nature (as in science) and for the gift of education; Christos for teaching me grace; Irkos for offering me unbinded books; Ilan for a singularity of inspiration.

PhD Collaborators and Mentors. To Alyosha for so generously gifting me his company, and supporting me as a flâneur in a goal-oriented community; to Benjamin, and the Antıkythera cohort (Michelle, Christina, Connor, Garry, Winnie, Philip, Ivar, Iulia, Cezar, Alasdair, Thomas, and ofc. Nicolai) for giving me an opportunity to be myself and introducing me to so many addictive terms and concepts. My closest collaborators: Matenia, for being the biggest supporter and motivator of my work, an oracular head reconnecting me with the classics; Nicolas G., Julien, and Hyolim for navigating my

storm of ideas and being always kind; Shiry for her commitment in my thesis, making me feel as part of her extended family, and for wanting to build a Japanese garden inside a city; most importantly, to Mathieu without whom this PhD would have neither started nor ended, spending an extensive amount of time, providing practical support in very diverse instances and occasions. Among all, I thank him for teaching me how to be a sculptor.

Most of all, to my mother Eleni, who was always there for me as her highest priority and for teaching me unconditional love, and to my father Panagiotis who always made sure that everything works, as best as possible. To my grandparents and our extended family who nurtured as their pride and legacy: love, generosity and respect. To my godcats: Morisson, Yoko, and Hercules for being /ei/.

Contents

Chapter 1

Introduction

> The great certainty of the natural sciences in comparison with the study of psychology or consciousness comes exactly from the fact that they choose for their object what is strange, while it is almost contradictory and even absurd to try to choose for one's object what is not-strange.
>
> The Gay Science, Friedrich Nietzsche

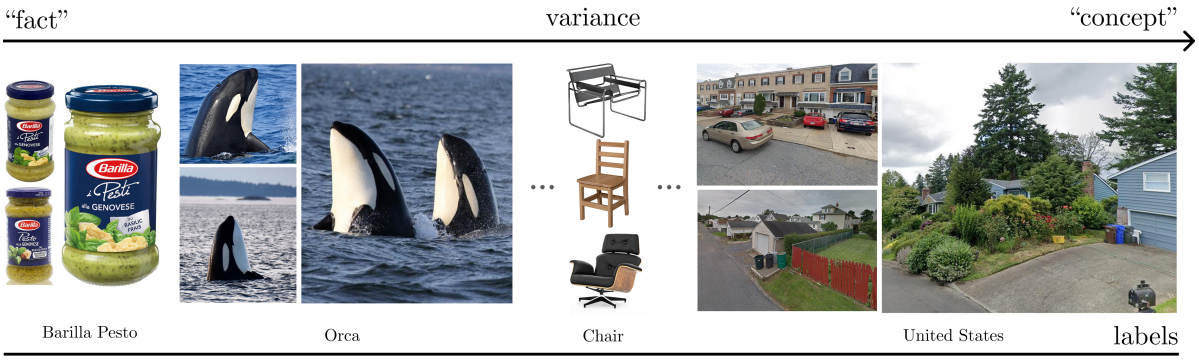
1.1 Philosophical Introduction

The divide between the sciences and the humanities appears foundational in the way institutional knowledge is produced. The one is often asserted as objective or quantifiable while the latter is often cast as subjective or perceptual (phenomenal). Yet for both the goal is to capture what is real, or at least to contour what exists, offering different methodologies with their respective limits. In fact, both can be seen as moments in the production of human knowledge. Human knowledge often starts as a journey from subjective observations, where in an effort of gradually trying to ground them, a form of quantifiable abstraction emerges tasked to provide a guarantee of generality. In the human sciences, one such tradition tries to explain such observables (e.g., social behavior) by trying to abstract their elements into terms of mathematical equations and to draw a parallel between their interactions to those of physical systems [Macy et al., 2024]. A typical example is the German sociologist Nicolas Luhman, who tried to conceptualize social dynamics as dynamics of abstract (mathematical) systems [Luhmann, 2013]. While these approaches have their merit, their migration to such abstractions limits their conclusions from ever bridging back the gap to reality, that these abstractions required in order to function. Their foundational problem is that

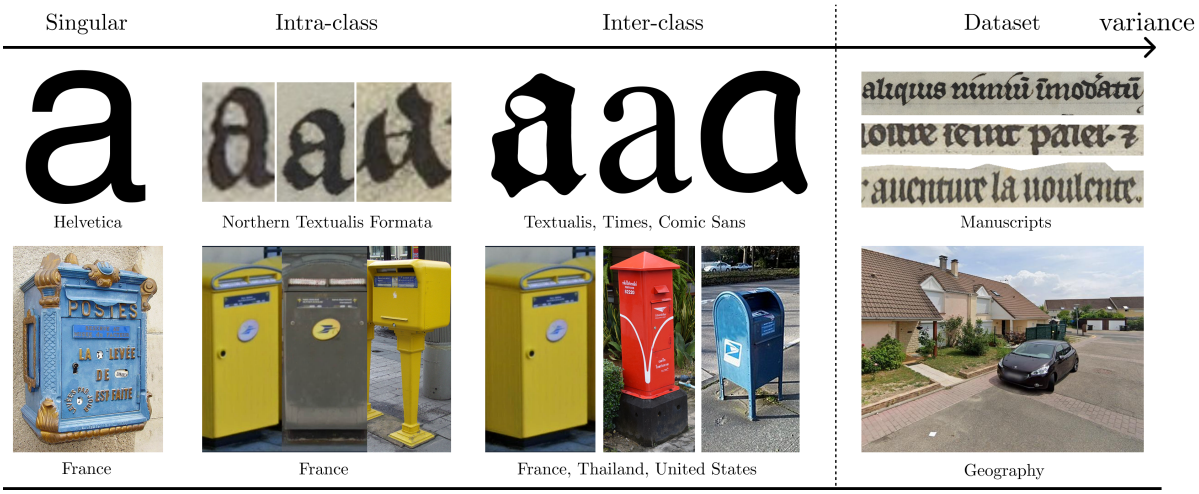
instead of focusing in understanding the *anthropologic complexity* of their input data, i.e., the way their data is both perceived and constructed by humans, their goal is more to capture their *Kolmogorov complexity*, trying to form a compressed representation of them as a physical world that can claim to exist outside humanity. But, to paraphrase Prof. Dawkins, in order to understand that part of the physical reality that is human, models need to be as "drunk" on symbols as they are on "facts" [Dawkins, 2000]. For example, in order for a visual system to understand the "nature" of a church, it would need to incorporate the factual existence of god inside culture irrelevant of whether it is factual in respect to physics. Thus, instead of trying to reduce the humanities into physics, what if we could elevate physics into the humanities? One way would be to create a physical system that has the power of understanding and manipulating two very central modalities of human culture, images and text, just like humans do. Performing such a migration of human perception into machinic perception, can in fact offer a form of *objective subjectivity*. While models are subjective as they depend on the data, architecture and task they were trained on, they remain objective as all their inference is operated via weights and computational processes that can be replicated and analyzed through digital computers. By making technology aware of both the processes of design and emergence that govern human societies, such a technology can allow for the first time in history a proper automation of the production of human knowledge. However, why would someone want to involve digital computation in the production of human knowledge in the first place?

1.2 Motivation

A great part of human knowledge, involves the recurrent synthesis of external information into archives. Yet, as these archives get large, it becomes increasingly hard for humans to manually produce knowledge and ground it at scale. Exploratory Data Analysis [Tukey, 1977], has been a common technique that researchers in the digital humanities have been using for almost 50 years to understand data of a non-human scale. However, their observations could only be as good as their models. This endeavor has been historically hard, as researchers really failed to properly formalize the way humans perceive the world. Yet, after tremendous progress in statistical modeling, machine learning, and optimization, self-supervised, generative, deep learning models are slowly turning the question of aligning a model's perception to human perception into a practical one. Modern approaches are graduating from performing retrieval tasks to meet the challenge of performing reliable mining and summarization of unseen



(a) **Signifier's Paradox.** While a category correctly describes a part of the visual world what it describes may vary significantly.



(b) **Hidden Structure in Visual Variation.** The visual appearance of the signified images may hide distinct structure of visual variation, either behind or across classes.

Figure 1.1: **Motivation.** While labels may be consistent the visual appearance of the signified images, may vary significantly (**top**). The variance of images can be attributed to inter-class variation of a given sub-category or intra-class variation across categories (**bottom**).

information, a fundamental skill required for producing new knowledge. When this involves the high-level abstract representation of text, automating search and compilation of information have been two of the biggest computational breakthroughs of the last 20 years. However, when it comes to low-level forms of information such as images (and sound), this task comes with increased degrees of freedom. For a system to automatically extract knowledge from such data, it is important that it is able to navigate both how humans perceive images in terms of sensory input (e.g., [Wertheimer, 1938; Blake and Zisserman, 1987; Biederman, 1987]), and in terms of labels (e.g., [Barthes, 1990; Rosch and Lloyd, 2024; Barriuso and Torralba, 2012]).

From Fact to Concept. The most established way that humans can abstract, share and communicate the structure of the visual world is to associate it with labels. A label is a signifier, a symbol or a text, that is inferred by instances of the signified, which it is meant to signify. For the process of signification to be useful for a specific task, labels need to effectively abstract or group visual variation. For example, in a text manuscript, the label of the character 'a' should be associated with all the different ways 'a' is inscribed. If this letter comes from a specific typeset, the association is trivial. However, in all other cases, deriving this association necessitates learning to properly organize its variation and discriminate it from other forms. As demonstrated in Fig. 1.1a, this brings us close to a paradox. While labels may correctly abstract the visual world, the visual variation that they capture may diverge significantly. To clarify this further let's pick the example of the *Barilla pesto*. As an industrial object it should carry almost no variation.[1] Of course the illumination, the sensor, the environment, or the background where it is captured may vary, but as an actual, signified object it is almost identical. In other words this *signifier's paradox* is especially relevant even if we had a perfect way of isolating and grouping all the salient parts that associate an image with its label. This creates a spectrum between labels that are more "factual" and labels that are more "conceptual". In the words of the media scholar McKenzie Wark, "a good fact is true for something in particular, while a good concept is slightly true about a lot of things" [Wark, 2023].

Functional Ontology of Categories. This is in fact a very old problem in the philosophy of categorization [Rosch and Lloyd, 2024; Murphy, 2004; Malisiewicz and Efros, 2009]. Arguing against the unchangeable form of Platonic metaphysics [Reeve et al., 2004] and its criticism through the rule-based definitions of categories of Aristotle's [Studtmann, 2024], Wittgenstein famously discussed in his philosophical investigations [Wittgenstein, 2009] that categories such as the word "game", one which he himself would use to describe the way language evolves, group instances of things that may share common properties, for example competitiveness or fun, but which are not necessarily consistent or discriminative across all members of its class.[2] Inspired by Wittgenstein, the cognitive psychologist Eleanor Rosch tried to show that human visual categorization is performed through what she called a prototype theory [Rosch,

[1] In fact this corresponds to a shape of Barilla pesto discontinued in 2022 for reasons of sustainability, which became so recognizable that the company added a warning label so that one could learn its new appearance after the packaging changed.

[2] Don't say: "There must be something common, or they would not be called 'games'"—but look and see whether there is anything common to all. [Wittgenstein, 2009]

1973]. For Rosch a set of empirical data points can all be associated to a centroid that effectively clusters them together, during our perception. Instead of categories being defined by their decision boundary (i.e., "where someone subject draws the line") they maybe associated to a central object that is *typical* of that class to which visual samples are grounded in order to be classified. For a single category these centroids may be multiple in order to properly accommodate all the things a category may represent. In a follow-up work, Medin and Schaffer tried to prove that visual inference is instead performed by associating the input instance to a set of samples stored in memory that are associated to a category [Medin and Schaffer, 1978].

What is an 'a'? While this may seem like an abstract problem it is an integral part of the very nature of categorization, which extends to something as trivial as our example of a signifier: the letter 'a'.[3] In the Middle Ages, professional handwritters, formally called scribes would copy notable texts like the Bible in order to pass knowledge from generation to generation and assert authority [Coulson and Babcock, 2020]. The script they would adopt for writing a text would try to inscribe consistent aesthetic qualities that would reflect its origin and purpose. To date these texts and agree on a common basis of categorization, palaeographers try to visually abstract a certain family of letters into textual descriptions. These Aristotelian definitions of categories, called typologies [Derolez, 2003] suffer from two main problems: **(a)** they are vague in their transmission leaving space for ambiguity and **(b)** as they are defined using text can't grasp the minute but significant variance of the visual data they aim to represent. This way, an alternative approach like the prototypes of Eleanor Rosch [Rosch, 1973] can allow to better organize existing categorization and even locate minute visual structure in its variation.

Classification is (often) not final. The practical motivation of this thesis is that behind concepts like a country's name hides salient visual variation that can be extracted and grouped. For example, behind the name of a country one could discover visual structure that can be grouped as a set of architectural (windows, roofs), infrastructural (road marks, electricity poles), or regulatory (license plates) visual elements. In Fig. 1.1b we locate two types of hidden structure of visual variation. The first type, concerns intra-class variation, or the minute differences between elements that are assigned on the same parent class. For example, given a script type of written charac-

[3] Douglas Hoffstader, famously wrote in 1985 that "the center problem of AI is the question: What is the letter 'a'?" [Douglas, 1985].

ters or equivalently a species of birds, one would like to understand whether there is minor variation that lies behind their underlying morphology, be it ascenders or beaks, which could characterize them as subtypes or subspecies. Even if this classification is arbitrary, this analysis can both allows us to establish its validity in light of a certain context [Stutzmann, 2016], or prove that it is arbitrary (i.e., statistically irrelevant) as in the case of the famous "Salamander's tale" [Dawkins, 2005]. The second type of hidden structure concerns inter-class variation, that lies behind the images of a certain domain. For example, given images that are conceptually identified by their geographical location, there may be elements that well characterize them, e.g., post-boxes [geohints, 2023]. While in the first case each label can be better structured in regard to participating into one of further categories, in this case each label can be decomposed and grouped into a ("mid-level") visual vocabulary [Singh et al., 2012]. Choosing which type of hidden structure of visual variation is more relevant to pursue, is sensitive to the input domain. For example, finding subfamilies of post-boxes for a specific country may be ill-posed due to lack or data or too trivial. Inversely, performing discovery of all the elements that make a manuscript typical to its family may be straight-forward as we already know that characters are expected to be the "visual vocabulary" of a manuscript. In both cases the potential of discovering further visual structure starting from predefined classification, is the fundamental motivation behind *image data mining*.

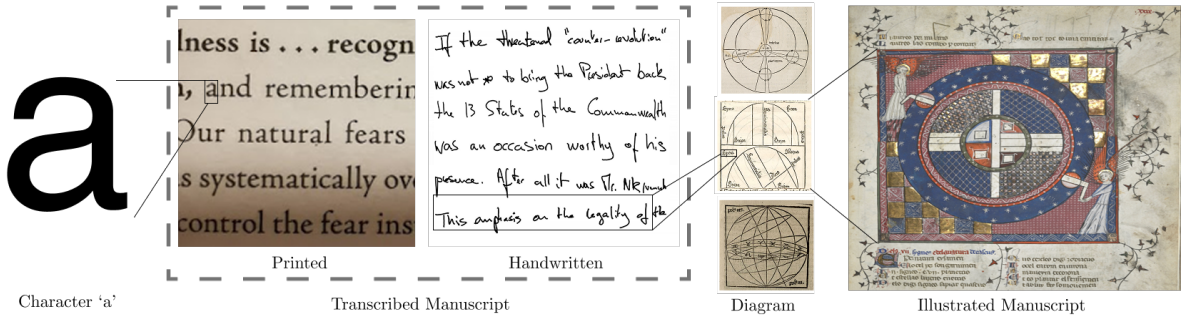
1.3 Goal

The goal of this thesis is to perform Image Data Mining (see Sec. 2). It aims to discover informative visual structure that lies behind the variation of established human categories and present it in a human interpretable way. Visual structure is defined as groups of individual elements (image patches, segments, etc.) that can be clustered together. Humans both perceive the visual world as a combination of such structure and in turn construct it as such, because of how they learn to perceive it. A written manuscript is composed of printed characters that readers learn to recognize (Fig. 1.2a), and a street is composed of road tracks that drivers learn to identify. As the datasets studied in this thesis, concern relevant domains (Fig. 1.2b), our goal is to build on systems that are able to identify this type of structure, as are for example discriminative classifiers, or segmentation networks. However, even if a system has the capacity to detect written characters it may not provide an adequate representation to differentiate and discover new visual structure.

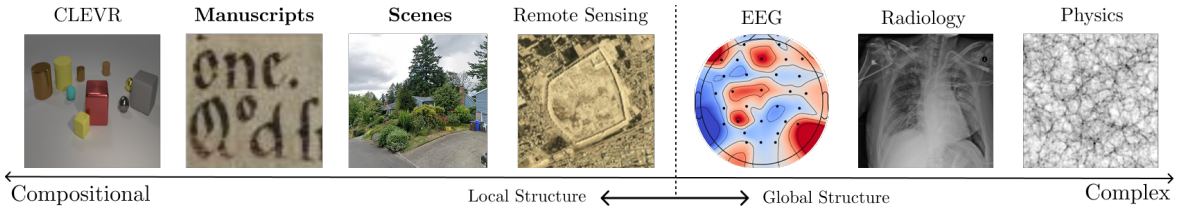
The importance of this problem becomes very apparent in geographical data. Consider for a moment the *"all text in nyc"* artwork that lifts all text appearing in the Google street view scenes of New York in a searchable interface [Zhao, 2024]. Having this annotation one could detect the statistics of certain phrases or find strange words that appear inside the city's billboards[4]. However, unlike text where these compositional attributes are often predefined and few, properly naming elements of an input scene is often ambiguous [Blake and Zisserman, 1987]. Similarly, with searching for text, one could try instead to search for predefined attributes that a system can learn to identify using supervised learning, by relying on standard semantic-segmentation methods [He et al., 2017] or using more open-ended text-based approaches, using multimodal models such as CLIP [Radford et al., 2021]. For example, one could search for images that contain a "car", "traffic lights", or "sky rise buildings that repeat on the side of a large boulevard". Yet even if we had the resources to list and count occurrences across all this combinatory of attributes, they may still be too generic to capture the minute important details under analysis. In other words even if our data are compositional, we may not already know what they are composed of. For example, in the case of UK windows, most buildings in the world have windows, yet the ones found in the UK have a very typical and consistent appearance. While one may describe them as "UK, white, UPVC windows", simply using a description isn't informative either: if a certain type window is only associated with a certain place in the world, then the discovery has already taken place. "As soon as something is named it is recoverable" [Baudrillard, 1999].

Focus of this thesis. More Concretely, this thesis aims at designing methods, that can mine visual structure in two practical cases of visual variation (Fig. 1.3): **(a)** transcribed documents, in order to analyze minute variation across characters (Chap. 3), and **(b)** labeled large-scale image datasets, in order to extract typical visual structure that summarizes a certain label, be it location, time, or type of scene (Chap. 4). We first focus on the micro-variation of written characters (Fig. 1.3a). For this domain, we aim to extract an aggregate representation of the font (or script) of a given set of documents. Later, by spatially aligning these representations for each character across different collections of fonts our goal is to be able to perform visual comparison that is interpretable while being quantifiable. This can provide us with clues about the evolution of the written character morphology that is hard to summarize otherwise. Our second goal, is to discover macro-variation that lies behind labeled datasets

[4] As the author's first name: https://www.alltext.nyc/search?q=Yannis appearing almost 100 times!



(a) **Example of Compositionality.** A letter can be part of a line of a written manuscript, which can be contained in an illustration, that can itself be part of an illustrated manuscript.



(b) **Compositional vs. Complex datasets.** Compositional datasets, as the ones studied on this thesis (**bold**) can be reduced to local structure which is informative for their global analysis. However, this is not the case for complex datasets, where analysis needs to be aggregated into a more global structure for it to be informative.

Figure 1.2: **Nature of the Studied Datasets.** This thesis focuses on compositional datasets, where local visual structure can be extracted and is informative of their parent class.

(Fig. 1.3b). For example, starting from images that come from a specific country, one would like to discover consistent visual structure that is typical of that location. These can be for example road-tracks, utility poles, post-boxes, windows and more. Such an approach allows human users to arrive at a high level understanding of how a certain label is represented inside a visual collection, potentially providing them with visual cues that can help them understand this collection on a higher level.

1.4 Challenges

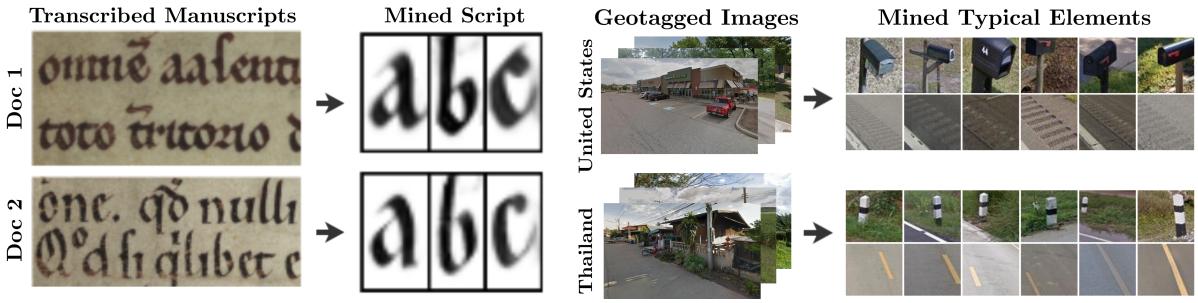
Image Data Mining, is challenging as it requires discovering and summarizing visual structure that is not already predefined by the input labels. Both because it is meant to be applied on datasets that concern the digital humanities, and because it is meant to be used by researchers to aid their analysis, the way we design these processes needs to be informed by the way humans already perceive and navigate visual data,

both in terms of labels and in terms of visual perception. This entails three different challenges.

Reconciling Supervised and Unsupervised Discovery. In order to identify visual structure that is common, one needs to also decide how to group it. In fact these two procedures, counting and grouping, are inseparable. This raises a problem of discovery. A rich unsupervised object discovery literature [Villa-Vásquez and Pedersoli, 2024], tried to show how useful or intuitive visual structure can emerge using certain architectures and training pipelines. However, the default use case of Image Data Mining, isn't simply the unsupervised discovery of visual structure as is for example the case with object discovery, but the identification of informative elements that best correspond to the labels of a weakly-labelled dataset [Singh et al., 2012]. While there exists a rich literature of categorical supervision, it mainly focuses on reliably recognizing predefined human categories [Wang et al., 2022]. This raises the challenge of reconciling both supervised and unsupervised approaches to discover non-annotated elements in our input dataset, while bringing this discovery closer to human perception.

Lack of Ground Truth A more foundational challenge, is that unlike object discovery, where the aim is to discover commonly occurring visual structure, in image data mining the visual structure that is expected to be discovered is in the form of clusters of *frequent outliers*, that are neither pre-assumed nor trivial. This locates Image Data Mining, in a regime of lack of ground truth that relies on subjective evaluation. It turns it into a task for which it is hard to measure progress in high granularity, and where its evaluation benefits from interdisciplinary expertise. Not only that but answers to the question of what an informative cluster is for a certain label may be multiple and sensitive to the input dataset. For example, while for purposes of classification two different measures of similarity, e.g., CLIP [Radford et al., 2021] or DINO [Caron et al., 2021] may have a similar performance, their qualitative performance, i.e., their retrieved nearest neighbors for the same dataset, may be noticeably different.

Interpretable and Faithful Summarization. The discovered output visual structures need to be summarized to a human level, in order to be compiled towards a form of conclusive evidence. Such evidence needs to be the most representative for the posed question and context. Not only that, but it also needs to be intimately connected to the way the model actually represents the input data. This connection



(a) **Inter**: *script* of historical Manuscripts. (b) **Intra**: *typical elements* of different countries.

Figure 1.3: **Goal.** Given labeled datasets, our goal is to summarize and compare their most informative elements in a human interpretable way. **(a)** given a collection of annotated manuscripts, our goal is to extract and summarize their script allowing experts to perform comparative quantitative morphological analysis (Chap. 3). **(b)** given a collection of annotated datasets of street-view images, our goal is to extract and summarize their typical visual structure, so that non-expert users can understand their labeling in relation to further visual elements (Chap. 4).

should be clear to the target user otherwise a future analyst that uses this tool can easily manipulate its results and project their biases in order to arrive to their desired conclusions [Sculley and Pasanek, 2008]. Unfortunately, most deep-learning architectures output high-dimensional representations of their input which can then be aggregated in multiple ways and whose features are often polysemantic [Olah et al., 2020]. This raises the challenge of how we can design approaches, which while being able to properly represent the data can provide interpretable summaries faithful to how the model actually perceives them.

1.5 Contributions

The approach of this thesis is to mine informative visual structure, using implicit or explicit, synthesis based compositional methods. These approaches align more with human perception [Biederman, 1987] and their predictions can be directly interpreted. We use two methods of the recent literature that can best capture the properties of the underlying visual structure. Sprite based methods [Smirnov et al., 2021; Monnier et al., 2021], are able to capture minute details of the input images by effective modeling the input data as a combination of a fixed subset of components. Diffusion models, through denoising learn to compose patches of the input scene into a more complex layering that can reproduce more abstract visual structure [Ozguroglu et al., 2024; Kamb and Ganguli, 2024]. Escaping the purism between unsupervised and supervised learning, our work adapts these methods to guide the discovery of new visual structure starting

from existing weak-annotation. For example, in the case of letters we don't want a system to have to both learn from scratch what a "d" is (which can be ambiguous as we discuss in Chapter 3) when our goal is to capture its shape. Instead, we would like to construct a system that learns the shape of "d", in a way that allows us to compare it with other shapes. Similarly, we would like to prime our system with the ability to recognize common visual structure or discriminate countries, so we can later find what visual structure of each country make it typical. Starting from established categories this enables mining visual structure that is not already annotated in the original training set, yet which corresponds to meaningful information.

Analyzing morphological variation of characters. To study images characterized by intra-class variation we use **sprite**-based methods [Monnier et al., 2020; Smirnov et al., 2021], that learn a decomposition of input images through a set of differential transformations of smaller input images, which in computer graphics are called sprites. Given an input text line, a neural network is trained to select and learn their appearance in order to effectively reconstruct all the text lines of an input dataset, on average. What sprites get to capture are interpretable summaries of the appearance of input characters. As they are very close in reconstructing the target images, they can be used to reliably capture the average written morphology of a given type. We show that this method can operate in a variety of printed and handwritten datasets [Siglidis et al., 2024a] and that using the learned sprites produced by this method can enable quantitative morphological palaeographical analysis [Vlachou-Efstathiou et al., 2024].

Mining typical summaries of labeled datasets. When studying inter-class variation of complex data that can hardly be characterized by a repetition of the same input image we rely on another synthesis method that is known as **Diffusion Models** [Ho et al., 2020] which learn to synthesize images by removing Gaussian noise across multiple resolutions. This allows them to learn hierarchical and robust representations of the input scene [Li et al., 2023a]. In fact, paired with text conditioning these models can learn to compose arbitrary combinations of visual concepts prompted through words, into an output image. By using them as a strong prior, we can identify how informative parts of the input image are for a target class, by averaging the difference in denoising performance using the information of a target class. Those that can be denoised significantly better in presence of the target class are clustered using features of the diffusion model and are then ranked to produce the final visual vocabulary. This type of approach allows us to summarize an input label in a way that is both

interpretable and that is directly connected to what the model perceives and groups as the most typical visual structure of a set of input labels [Siglidis et al., 2024b].

1.6 Thesis outline

This thesis consists of five chapters, which are organized as follows:

Chapter 2: Related Work. This chapter discusses related work in the area of Image Data Mining. It explores image data mining around three axes: discovery, mining and summarization. It outlines an evolution of methods and problems in each axis and clarifies how they are related to our work.

Chapter 3: The Learnable Typewriter, A Generative Approach to Text Analysis. Then, this thesis focuses in how to capture the inter-variation of characters from printed or real manuscripts of input text lines using a sprite-based approach. Through an architecture that is explicitly designed for reconstructing text lines using weak supervision it allows to capture the morphology of written characters [Siglidis et al., 2024a] for various datasets of printed font [Vincent, 2007], ciphers [Knight et al., 2011] and historical fonts [Seuret et al., 2023]. The power of this approach, is further demonstrated through an application to palaeographic analysis [Vlachou-Efstathiou et al., 2024] that allows qualitative and quantitative comparison in both rare [Camps et al., 2022] and established typologies [Derolez, 2003].

Chapter 4: Diffusion Models as Data Mining Tools. In the next chapter the focus turns to detecting visual structure of macro-variation across a variety of annotated image datasets of varying sizes. Starting from a pre-trained diffusion model [Rombach et al., 2022], a diffusion-based score is introduced that allows to mine informative elements from the dataset given an input conditioning. This allows to extract visual summaries for the labels of input datasets, scaling across a variety of data, such as historical cars [Lee et al., 2013] (10K), portraits [Chen et al., 2023] (25K), geographical data [Luo et al., 2022] (350K), and places [Zhou et al., 2017a] (1.8M), enabling an interpretable understanding of the meaning of a label assigned inside the context of an input dataset [Siglidis et al., 2024b].

Chapter 5: Conclusion. This thesis concludes by summarizing our work and discussing future directions, open problems and implications.

1.7 Publications

The following three publications are presented in the manuscript [Siglidis et al., 2024a; Vlachou-Efstathiou et al., 2024; Siglidis et al., 2024b]:

- Ioannis Siglidis, Nicolas Gonthier, Julien Gaubil, Tom Monnier, and Mathieu Aubry [2024]. The Learnable Typewriter: A generative approach to text analysis. ICDAR.
- Malamatenia Vlachou-Efstathiou, Ioannis Siglidis, Dominique Stutzmann, and Mathieu Aubry [2024]. An interpretable deep learning approach for morphological script type analysis. IWCP.
- Ioannis Siglidis, Aleksander Holynski, Alexei A. Efros, Mathieu Aubry, and Shiry Ginosar [2024]. Diffusion models as data mining tools. ECCV.

Our code was open-sourced[5] and presented using specialized webpages[6] which contain additional visualizations and results. Our work on The Learnable Typewriter received **the best paper award** at ICDAR 2024, and our work on Diffusion Models as Data Mining Tools aside from being published as a Poster on ECCV, was invited for a spotlight talk at the *Workshop for Visual Concepts*[7].

Not presented in this thesis. During my PhD, I was a joint first author in the following publication [Astruc et al., 2024], which started as a group hackathon idea from Nicolas Dufour and was performed under the supervision of Loic Landrieu:

- *Guillame Astruc, *Nicolas Dufour, *Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, Lintao Xu, Hongyu Zhou, and Loic Landrieu. [2024]. Openstreetview-5m: The many roads to global visual geolocation. CVPR.

The goal of this work was to release the largest open-source dataset for global scale visual geolocation, that had significantly more well-defined label-image associations, and a more balanced coverage than existing open datasets, in order to facilitate turning geolocation into a standard benchmark for evaluating image models as opposed to well established datasets like ImageNet [Deng et al., 2009].

[5] https://github.com/ysig
[6] https://ysig.github.io/phd/#papers
[7] https://sites.google.com/stanford.edu/visual-concepts-workshops/eccv24

Chapter 2

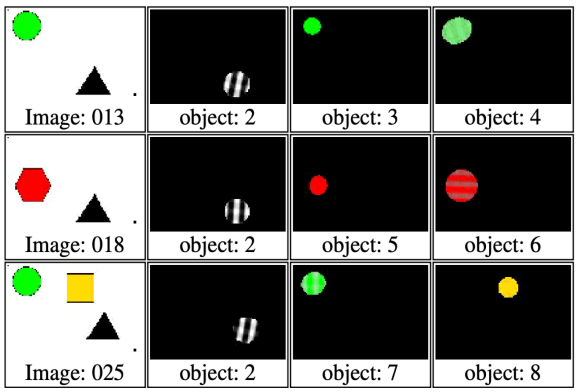
Related Work

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,

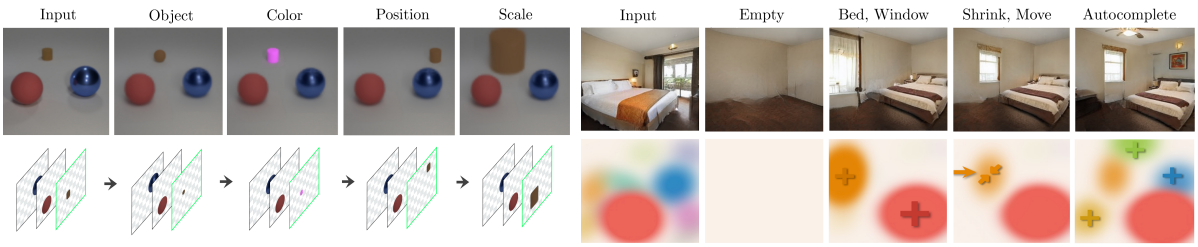
techniques of topic modelling were used to both classify and segment objects from input scenes [Cao and Fei-Fei, 2007]. However, scaling these types of approaches to datasets with multiple and diverse examples remains challenging as the size of the vocabulary can increase significantly, turning model-based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

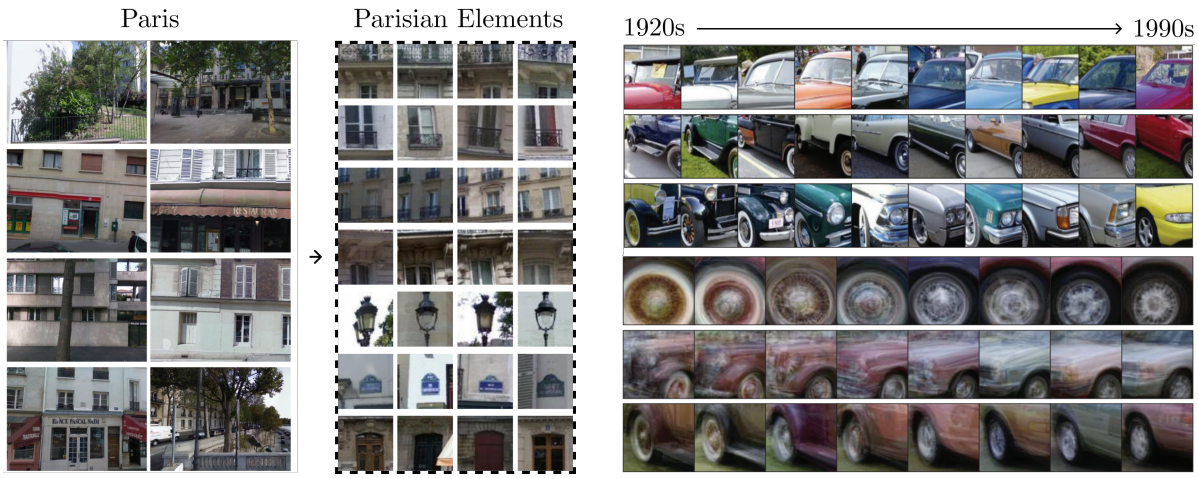
Supervised Clustering. Early successful works formalized clustering as a convex integer programming using labels assigned through SVMs [Xu et al., 2004]. Later, an approach known as DIFFRAC [Bach and Harchaoui, 2007] formalized clustering directly as a supervised problem trained with MSE regression, 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.

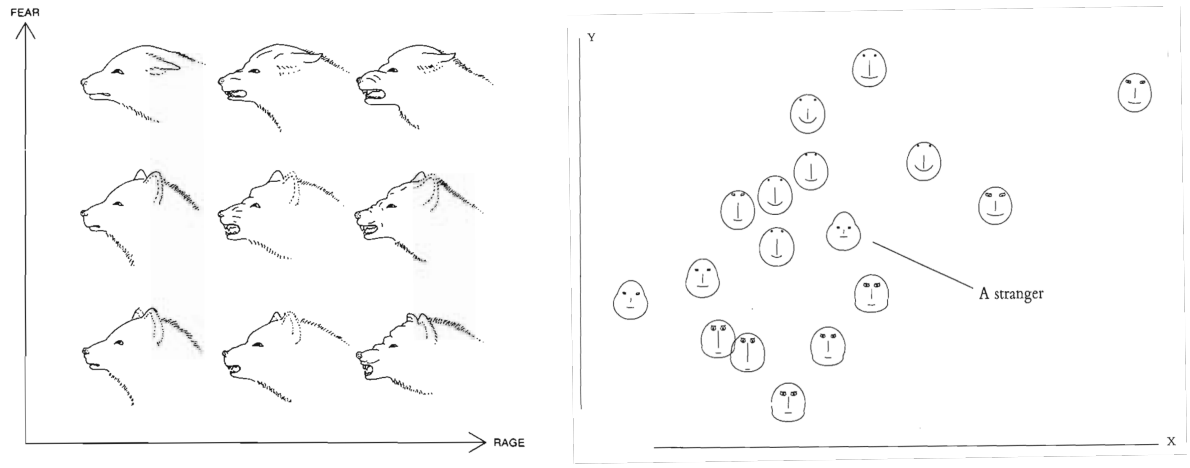
Mining new categories. One task relevant to data mining is that of category discovery. Its objective is, given an existing semi-annotated training set of images to learn a way of representing images at training time such that during test time a model can effectively cluster images into previously unassigned 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.



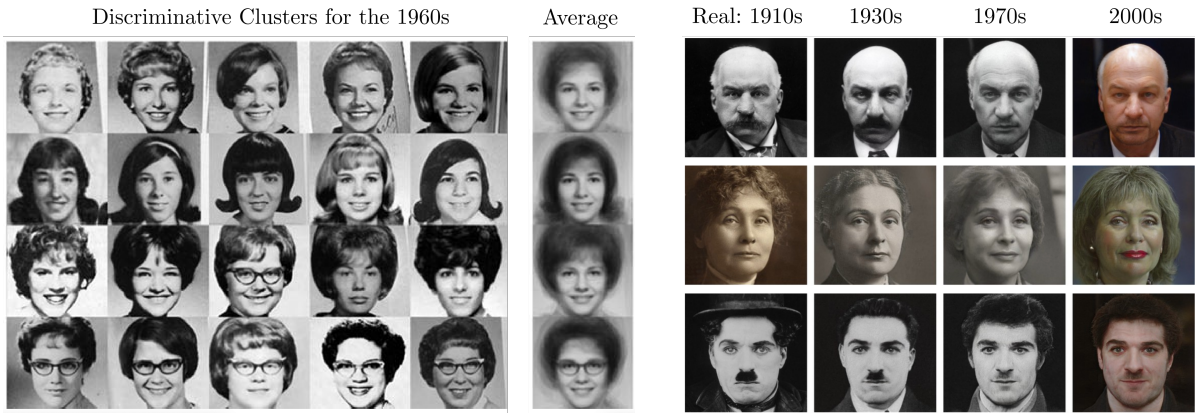
(a) Rational (b) Affective

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.

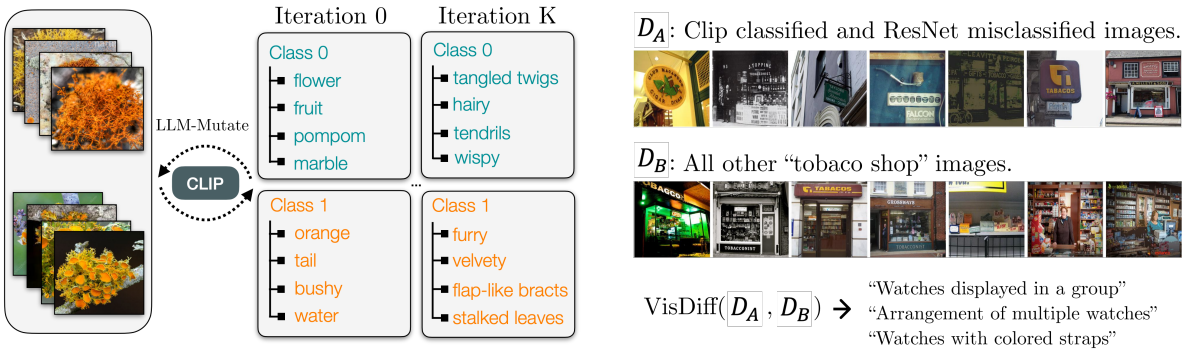
2.3.1 Visual Summaries

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

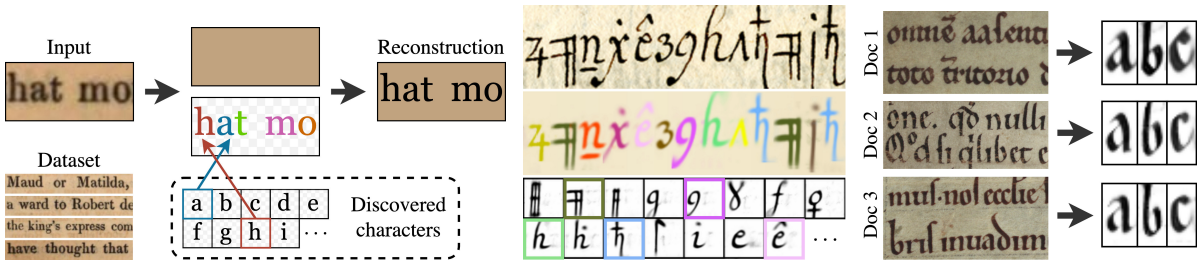
(a) Twitter Explorer (b) Average Explorer

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.

Chapter 3

The Learnable Typewriter: A Generative Approach to Text Analysis



(a) The Learnable Typewriter idea (b) Cipher analysis (c) Paleography

Figure 3.1: **The Learnable Typewriter. (a)** Given a text line dataset, we learn to reconstruct images by discovering the underlying characters. This generative approach can be used both **(b)** to analyze complex ciphers [Knight et al., 2011] and **(c)** as an automatic tool for the study of handwriting in historical documents [Camps et al., 2022].

3.1 Introduction

A popular approach to document analysis in the 1990s was to learn document-specific character prototypes, which enabled Optical Character Recognition (OCR) [Kopec and Lomelin, 1996, 1997; Xu and Nagy, 1999; Baird, 1999] but also other applications, such as font classification [Hochberg et al., 1997] or document image compression and rendering [Nolan and Filippini, 2010]. This idea culminated in 2013, with the Ocular system [Berg-Kirkpatrick et al., 2013] which proposed a generative model for printed text lines inspired by the printing process, and held the promise of achieving a complete explanation of their appearance. These document-specific generative

approaches were however overshadowed by discriminative approaches, whose sole purpose is to perform predictions, and which lead to higher performance at the cost of interpretability, e.g., [Graves and Schmidhuber, 2008; Li et al., 2023b]. In fact, [Simard et al., 2002] showed that using transformations of input exemplars while performing digit classification can allow learning with much fewer samples. Yet, acquiring such exemplars remains costly and ambiguous for more complicated tasks. Inspired by this, in this chapter, we explore how modern deep approaches enable revisiting and extending analysis by synthesis approaches for text line analysis. In particular, we demonstrate an approach that can deal with challenging examples of handwritten documents, enabling a quantitative perspective in analyzing the character morphology in historical handwriting as it is studied in the discipline of palaeography.

While discriminative approaches have been largely dominant in the current deep learning-based computer vision, a recent set of works revisited generative approaches for unsupervised multi-object object segmentation [Burgess et al., 2019; Emami et al., 2021; Greff et al., 2017, 2019b; Yang et al., 2020; Crawford and Pineau, 2019; Deng et al., 2020; Eslami et al., 2016; Jiang and Ahn, 2020; Smirnov et al., 2021; Monnier et al., 2021]. Most of them provide results on synthetic data or simple real images [Monnier et al., 2021], and sometimes demonstrate qualitative results on simple printed text images [Smirnov et al., 2021; Reddy et al., 2022]. Surprisingly, images of handwritten characters, which were notoriously used in the development of convolutional neural networks [LeCun et al., 1989, 1998] and generative adversarial networks [Goodfellow et al., 2014] were largely overlooked by these approaches.

We build on two recent sprite-based unsupervised image decomposition approaches [Smirnov et al., 2021; Monnier et al., 2021] that provide an interpretable decomposition of images into a vocabulary of small images, called sprites. These methods are trained to jointly optimize both the sprites and the neural networks that predict their position and color. Intuitively, we would like to adapt these methods so that from text lines that are extracted from any given document, they could learn sprites that correspond to each character. By adapting MarioNette [Smirnov et al., 2021] to perform text line analysis, we provide quantitative evaluation on real data and an analysis of the limitations of state-of-the-art approaches for unsupervised multi-object segmentation. We argue that text-line recognition should be used as a benchmark for this task in future work.

Because unsupervised performances are not completely satisfactory, we combine this approach with a weak supervision from line-level transcriptions. Transcriptions are widely available and easy to produce with dedicated software, e.g., [Kahle et al., 2017; Kiessling et al., 2019], and we show that this dramatically improves results, while

preserving their interpretability. Through this thesis we motivate the idea that similar weak (i.e., image-level) annotations should also be considered for future problems of image decomposition.

Contributions. To summarize, we present:

- a deep generative approach to text line analysis, inspired by deep unsupervised multi-object segmentation methods, adapted to work in both an unsupervised and a weakly supervised setting,
- a demonstration of the potential of our approach in challenging applications, particularly in ciphered documents and in palaeographic analysis,
- experiments on four different types of datasets: the printed volume of Google1000 [Vincent, 2007; Gupta et al., 2018], the historical fonts of MFGR [Seuret et al., 2023], the Copiale cipher [Baró et al., 2019; Knight et al., 2011], and two sets of historical handwritten manuscripts between the 12th-15th century [Camps et al., 2022] and Tab. F.1.

Our implementation can be found at github.com/ysig/learnable-typewriter.

3.2 Related Work

Text Analysis. Image Text Recognition, including Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR), is a classic pattern recognition problem and one of the earliest successful applications of deep learning [LeCun et al., 1989, 1998]. The mainstream approaches for text line recognition rely on discriminative supervised learning. Typically, a Convolutional Neural Network (CNN) encoder will map the input image to a sequence of features, and a decoder will associate them to the ground truth, e.g., through a recurrent architecture trained with a Connectionist Temporal Classification (CTC) loss [Graves et al., 2006; Graves and Schmidhuber, 2008; Puigcerver, 2017; Bluche and Messina, 2017; de Sousa Neto et al., 2020], or a transformer trained with cross entropy [Kang et al., 2022; Li et al., 2023b].

More related to our work, ScrabbleGAN [Fogel et al., 2020] proposed a generative adversarial approach for semi-supervised text recognition, but their method is neither able to reconstruct an input text line nor to decompose it into individual characters. Also related are approaches that use already annotated sprites (referred to as exemplars or supports) to perform OCR/HTR in common fonts [Zhang et al., 2020], ciphers [Souibgui et al., 2020], and Cuneiform [Mikulinsky et al., 2025], by matching them to text lines. Recent unsupervised approaches, either cluster input

images embedded in a feature space [Baró et al., 2019] or rely on an existing text corpus of the recognized language [Gupta et al., 2018].

Closest to our work are classical prototype-based methods [Kopec and Lomelin, 1996, 1997; Xu and Nagy, 1999; Baird, 1999] and in particular the Ocular system [Berg-Kirkpatrick et al., 2013] which follows a generative probabilistic approach to jointly model text and character fonts in binarized documents and is optimized through Expectation Maximization (EM). Unlike us, it also relies on a pre-trained n-gram language model, originally from the English language and later extended to multiple languages [Garrette et al., 2015]. Other approaches rely on language models to identify characters [Kopec et al., 2001; Berg-Kirkpatrick et al., 2013; Gupta et al., 2018]. However, language models do not exist for unknown ciphers or historical manuscripts. Instead, we propose to disambiguate sprites by relying on line-level transcriptions.

Most related to our application in palaeography, [Goyal et al., 2020] proposes a probabilistic model for printed font analysis, [Srivatsan et al., 2021a] for linear scribal hands of linear-b. However, both works rely on single isolated and binarized characters as input, whereas the goal of our approach is to be directly applicable to colored text lines. Closer to us [Aiolli et al., 1999; Ciula, 2005] uses the tangent distance of [Simard et al., 2002] to learn prototypes from individual characters and cluster them into dendrograms of classes. Unlike us, this approach operates on cropped and segmented black and white characters.

Unsupervised multi-object segmentation. Unsupervised multi-object segmentation refers to a family of approaches that decompose and segment scenes into multiple objects in an unsupervised manner [Karazija et al., 2021]. Some techniques perform decomposition by computing pixel-level segmentation masks over the whole input image [Burgess et al., 2019; Emami et al., 2021; Greff et al., 2017, 2019b; Yang et al., 2020], while others focus on smaller regions of the input and learn to compose objects in an iterative fashion, mostly relying on a recurrent architecture [Crawford and Pineau, 2019; Deng et al., 2020; Eslami et al., 2016; Jiang and Ahn, 2020]. While all of these techniques can isolate objects from their backgrounds by producing segmentation masks, our goal is to summarize their recurring visual appearance.

We thus build on techniques that explicitly model the objects located inside the input image, by associating them to a set of image prototypes referred to as sprites through differentiable transformations [Monnier et al., 2021; Smirnov et al., 2021]. Sprites are color images with an additional transparency channel (RGBA), associated to networks that predict their spatial transformation [Jaderberg et al., 2015b] in order

to compose them onto a target canvas. DTI-Sprites [Monnier et al., 2021] provides good reconstruction fidelity, but can only predict a small amount of sprites for a collection of fixed-size images, and fails to scale when the number of objects within each image increases to those of real documents. At the same time, MarioNette [Smirnov et al., 2021] while being efficient, suffers from a high reconstruction error and fuzzy sprites that suboptimally reconstruct a toy text dataset.

3.3 The Learnable Typewriter

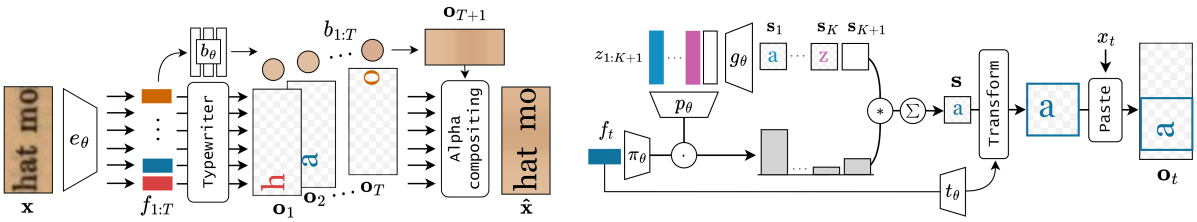
Given a collection of text lines that have a consistent font or script, our goal is to learn a representation of the average shape of all the characters it contains and a deep network that predicts how to transform them in order to reconstruct any input text line. Since complete supervision (i.e., the position and shape of every character found in a document) for would be extremely costly to obtain for our purposes, we propose to proceed in an analysis-by-synthesis fashion by building on sprite-based unsupervised image decomposition approaches [Smirnov et al., 2021; Monnier et al., 2021] which jointly learn a set of character images - called *sprites* - and a network that transforms and positions them on a canvas in order to reconstruct input lines. Due to the intrinsic ambiguity of decomposing a set of characters into sprites, we introduce a complementary weak-supervision from line-level transcriptions.

In this section, we first present an overview of our image model and approach (Sec. 3.3.1). Then, we describe the deep architecture we use (Sec. 3.3.2). Finally, we discuss our loss and training procedure (Sec. 3.3.3).

Notations. We write $a_{1:n}$ the sequence $\{a_1, \ldots, a_n\}$, and use bold letters $\mathbf{a}$ for images. An RGBA image $\mathbf{a}$ corresponds to an RGB image denoted by $\mathbf{a}^c$, alongside an alpha-transparency channel denoted by $\mathbf{a}^\alpha$. We use θ as a generic notation for network parameters and thus any character indexed by θ, e.g., a_θ, is a network.

3.3.1 Overview and image model

Fig. 3.2a presents an overview of our pipeline. An input image $\mathbf{x}$ of size $H \times W$ is fed to an encoder network e_θ generating a sequence of T features $f_{1:T}$ associated to uniformly-spaced locations $x_{1:T}$ in the image. Each feature f_t is processed independently by our *Typewriter* module (Sec. 3.3.2) which outputs an RGBA image $\mathbf{o}_t$ corresponding to a character. The images $\mathbf{o}_{1:T}$ are then composited with a canvas image we denote



(a) Overview of our full pipeline (b) Details of our Typewriter module

Figure 3.2: **Overview. (a)** An image is encoded into a sequence of features, each decoded by the Typewriter module into image layers. They are then fused by alpha compositing with a predicted low-res background. **(b)** The Typewriter module takes a feature as input, computes sprites and associated probabilities from learned latent codes, and composes them into a composite sprite that is then transformed and positioned onto an image-sized canvas.

$\mathbf{o}_{T+1}$. This canvas image $\mathbf{o}_{T+1}$ is a completely opaque image (zero transparency). Its colors are predicted by a Multi-Layer Perceptron (MLP) b_θ which takes as input the features $f_{1:T}$ and outputs RGB values $b_{1:T}$. All the resulting images $\mathbf{o}_{1:T+1}$ can be seen as ordered image layers and are merged using alpha compositing, as proposed by both [Monnier et al., 2021; Smirnov et al., 2021]. More formally, the reconstructed image $\hat{\mathbf{x}}$ can be written as:

$$\hat{\mathbf{x}} = \sum_{t=1}^{T+1} \left[\prod_{j<t} (1 - \mathbf{o}_j^\alpha) \right] \mathbf{o}_t^\alpha \mathbf{o}_t^c. \tag{3.1}$$

During training, the order of $\mathbf{o}_{1:T}$ in the compositing operation is randomized to reduce overfitting, as advocated by the MarioNette approach [Smirnov et al., 2021]. The full system is differentiable and can be trained end-to-end.

3.3.2 Typewriter Module

We now describe in detail the Typewriter module, which takes as input a feature f from the encoder and its position x, and outputs an image layer $\mathbf{o}$, to be composited. An overview of the module is presented in Fig. 3.2b. On a high level, it is similar to the MarioNette architecture [Smirnov et al., 2021], but handles blanks (i.e., the generation of a completely transparent image) differently and has a more flexible deformation model, similar to the one used in DTI-Sprites [Monnier et al., 2021]. More specifically, the module learns jointly RGBA images called *sprites* corresponding to character images, and networks that use the features f to predict a probability for each sprite and a transformation of the sprite. In the following, we detail how we obtain the

following three elements: the set of K parameterized sprites, the sprites compositing, and the transformation model.

Sprite Parametrization. We model characters as a set of K sprites which are defined using a generator network. More specifically, we learn K latent codes $z_{1:K}$ which are used as an input to a generator network g_θ in order to generate sprites $\mathbf{s}_{1:K} = g_\theta(z_{1:K})$. These sprites are images with a single channel that corresponds to their opacity. Similar to DTI-Sprites [Monnier et al., 2021], we model a variable number of sprites with an empty (i.e., completely transparent) sprite which we write $\mathbf{s}_{K+1}$. Instead of directly learning sprites in the pixel space as in DTI-Sprites [Monnier et al., 2021], we found that using a generator network yields faster and better convergence.

Sprite Probabilities and Compositing. To predict a probability p_k for each sprite $\mathbf{s}_k$, each latent code z_k is associated through a network p_θ to a probability feature $z^p_k = p_\theta(z_k)$ of the same dimension D as the encoder features ($D = 64$ in our experiments). We additionally optimize directly a probability feature z^p_{K+1} which we associate to the empty sprite. Given a feature f predicted by the encoder, we predict the probability p_k of each sprite $\mathbf{s}_k$ by computing the dot product between the probability features $z^p_{1:K+1}$ and a learned projection of the feature $\pi_\theta(f)$, and applying a softmax to the result:

$$p_{1:K+1}(f) = \text{softmax}\left(\lambda z^p_{1:K+1} \cdot \pi_\theta(f)^T\right), \tag{3.2}$$

where $\cdot$ is the dot product applied to each element of the sequence, $\lambda = 1/\sqrt{D}$ is a scalar temperature hyperparameter, and softmax is applied to the resulting vector. We use these probabilities to combine the sprites into the weighted average $\mathbf{s} = \sum_{k=1}^{K} p_k g_\theta(z_k)$. During inference, we simply select the sprite $g_\theta(z_k)$ with the highest probability p_k instead of computing a weighted average. Note that this compositing can be interpreted as attention operation [Vaswani et al., 2017]:

$$\mathbf{s} = \text{attention}(\bar{Q}, \bar{K}, \bar{V}) = \text{softmax}\left(\frac{\bar{Q}\bar{K}^T}{\sqrt{D}}\right)\bar{V}, \tag{3.3}$$

with $\bar{Q} = \pi_\theta(f)$, $\bar{K} = p_\theta(z_{1:K+1})$, $\bar{V} = g_\theta(z_{1:K+1})$, D the dimension of the features, and by convention $g_\theta(z_{K+1})$ is the empty sprite and $p_\theta(z_{K+1}) = z^p_{K+1}$. In fact, we show that directly optimizing $z^p_{1:K}$ instead of learning to predict the probability features $z^p_{1:K}$ from the sprite latent codes $z_{1:K}$, as in MarioNette [Smirnov et al., 2021], yields similar results. Note that we learn a probability code z^p_{K+1} to compute the probability of empty

sprites instead of having a separate mechanism as in MarioNette [Smirnov et al., 2021] because it is critical for our supervised loss (see Sec. 3.3.3).

Positioning and Coloring. The final step of our module is to position the selected sprite in a canvas of size $H \times W$ and to adapt its color. We implement this operation as a sequence of a spatial transformer [Jaderberg et al., 2015a] and a color transformation, similar to DTI-Sprites [Monnier et al., 2021]. More specifically, the feature f is given as input to a network t_θ that predicts three parameters for the color of the sprite and three parameters for isotropic scaling and 2D-translation that are used by a spatial transformer [Jaderberg et al., 2015a] to deform $\mathbf{s}$. Finally, using the location x associated with the feature f, we paste the deformed colored sprite onto a background canvas of size $H \times W$ at position x to obtain a reconstructed RGBA image layer $\mathbf{o}$. Positioning the sprites with respect to the position of the associated local features helps us obtain results co-variant to translations of the text lines and independent of the line size. To produce the background canvas, each of the features $f_{1:T}$ is first passed through a shared MLP b_θ, to produce a vector of T background colors $b_{1:T}$. We then use bilinear interpolation to upscale this vector to the full size of the input image x. Specific details concerning the parametrization of the transformation networks can be found in Appendix A (Sec. B.3).

3.3.3 Losses and training details

Our system is designed in an analysis-by-synthesis spirit and thus relies mainly on a reconstruction loss. This reconstruction loss can be complemented by a loss on the selected sprites in the supervised setting where each text line is paired with a transcription. In the following, we define these losses for a single text line image and its transcription, using the notations of the previous section.

Reconstruction loss. Our core loss is a simple mean square error between the input image $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$ predicted by our system as discussed in Sec. 3.3.1:

$$\mathcal{L}_{\text{rec}}(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{x} - \hat{\mathbf{x}}\|^2. \tag{3.4}$$

In the unsupervised setting, we don't train with any additional regularization.

Weakly Supervised Loss. The intrinsic ambiguity of the sprite decomposition problem may result in sprites that do not correspond to individual characters.

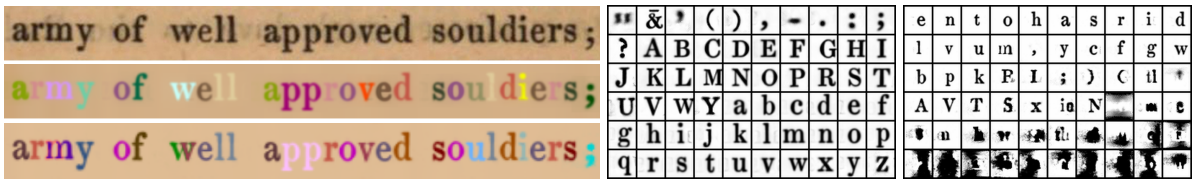
Using line-level annotation is an easy way to resolve this ambiguity. We find that simply adding the classical CTC loss [Graves et al., 2006] computed on the sprite probabilities to our reconstruction loss is enough to learn sprites that exactly correspond to characters. More specifically, we chose the number of sprites as the number of different characters and associate arbitrarily each sprite with a character and the empty sprite with the blank token of the CTC. Then given the one-hot line-level annotation y and $\hat{y} = (p_{1:K+1}(f_1), ..., p_{1:K+1}(f_T))$ the predicted sprite probabilities, we optimize our system's parameters by minimizing:

$$\mathcal{L}_{\text{sup}}(\mathbf{x}, y, \hat{\mathbf{x}}, \hat{y}) = \mathcal{L}_{\text{rec}}(\mathbf{x}, \hat{\mathbf{x}}) + \lambda_{\text{ctc}}\mathcal{L}_{\text{ctc}}(y, \hat{y}) \tag{3.5}$$

where λ_{ctc} is a hyperparameter and $\mathcal{L}_{\text{ctc}}(y, \hat{y})$ is the CTC loss computed between the ground-truth y and the predicted probabilities $\hat{y}$. We use $\lambda_{\text{ctc}} = 0.01$ in all our experiments, increased only for the less challenging Google1000 to $\lambda_{\text{ctc}} = 0.1$.

Implementation and training details. We train on the Google1000 [Vincent, 2007] and Fontenay [Camps et al., 2022] datasets with lines of height $H = 64$ and on the Copiale dataset [Knight et al., 2011] with $H = 96$. The generated sprites $\mathbf{s}_{1:K}$ are of size $H/2 \times H/2$. In the supervised setting, we use as many sprites as there are characters, and in the unsupervised, we set $K = 60$ for Google1000 and $K = 120$ for the Copiale cipher. We train for 100 epochs on Google1000 and for 500 epochs on Copiale with a batch size of 16, and we select the model that performs best on the validation set for evaluation. In the unsupervised setting, we use line crops of width $W = 2H$ and train for 1000 epochs on Google1000 and for 5000 on the Copiale cipher with a batch size of 32 and use the final model. The number of epochs is much higher in the unsupervised case than in the supervised case because the network sees only a small crop of each line at each epoch, but each epoch is much faster to perform. To always avoid learning sprites that reconstruct the background instead of the actual characters, we warm start the training process by only training the background MLP for 3000 gradient steps.

Our encoder network is a ResNet-32-CIFAR10 [He et al., 2016], that is truncated after layer 3 with a Gaussian feature pooling described in Appendix A (Sec. B.2). For our unsupervised experiments, we use as generator g_θ the U-Net architecture of Deformable Sprites [Ye et al., 2022] as it converged quickly, and for our supervised experiments a 2-layer MLP similar to MarioNette [Smirnov et al., 2021] which produces sprites of higher quality. The networks π_θ and p_θ are single linear layers followed by layer-normalization. We use the AdamW [Loshchilov and Hutter, 2019] optimizer with a learning rate of 10^{-4} and apply a weight-decay of 10^{-6} to the encoder parameters.



(a) Semantic Segmentation (↑ sup. ↓ unsup.) (b) Sprites (sup.) (c) Sprites (unsup.)

Figure 3.3: **Results on a printed document (Google1000).** In both the supervised and unsupervised setting our method produces meaningful sprites and accurate reconstructions **(a)**. We show the 60 most used sprites in alphabetic ordering in the supervised setting **(b)** and ordered by frequency in the unsupervised one **(c)**. See text for details and Appendix A for more reconstructions (Sec. A).

3.4 Experiments

In this section, we evaluate the performance of our method on challenging datasets of historical and modern fonts, as well as a handwritten cipher. We first introduce the datasets and metrics used in our evaluation in Sec. 3.4.1. Then we present qualitative results discussing the quality of the learned sprites with and without supervision in Sec. 3.4.2. Finally, in Sec. 3.4.3, to assess the performance of our method and ablate our architectural choices we present quantitative results both for reconstruction and for transcription quality, for both the supervised and the unsupervised setting.

3.4.1 Datasets and metrics

Datasets. We experiment with four datasets with different characteristics: Google1000 [Vincent, 2007], MFGR [Seuret et al., 2023], the Copiale cipher [Knight et al., 2011] and Fontenay manuscripts [Camps et al., 2022]:

• *Google1000.* The Google1000 dataset contains scanned historical printed books, arranged into Volumes [Vincent, 2007]. We use the English Volume 0002 which we process with the preprocessing code of [Gupta et al., 2018], using 317 out of 374 pages and train-val-test set with 5097-567-630 lines respectively. This leads to a total number of 83 distinct annotated characters. Although supervised printed font recognition is largely considered a solved problem, and the annotation for this dataset is actually the result of OCR, this document is still challenging for an analysis-by-synthesis approach, containing artifacts such as ink bleed, age degradation, as well as variance in illumination and geometric deformations due to digitization.

• *MFGR.* The ICDAR-2024 "Multi Font Group Recognition and OCR challenge" dataset [Seuret et al., 2023], contains text lines that were printed with a set of 8 distinct

typefaces: *antiqua, bastarda, fraktur, gotico-antiqua, italic, rotunda, schwabacher, textura*. We focus only on lines that contain fonts from a single group. This dataset is similar to Google1000, but comes with the challenges of older prints, such as non-fully printed letters and allographs. With 12K-45K training lines for each of the 8 different typefaces, it serves as an ideal benchmark to assess the robustness of our model in learning meaningful sprites across a variety of historical prints.

• *Copiale cipher.* The Copiale cipher is an oculist German text dating back to an 18th-century secret society [Knight et al., 2011]. Opposite to Baro et al. [Baró et al., 2019] which uses a binarized version of the dataset, we train our model on the original text-line images, which we segmented using docExtractor [Monnier and Aubry, 2020] and manually assigned to their annotations, respecting the train-val-test split of Baro et al. [Baró et al., 2019] with 711-156-908 lines each. The total number of distinct annotated characters is 112. This dataset is more challenging than printed text because of the handwritten variance of a historical manuscript, and its large number of characters.

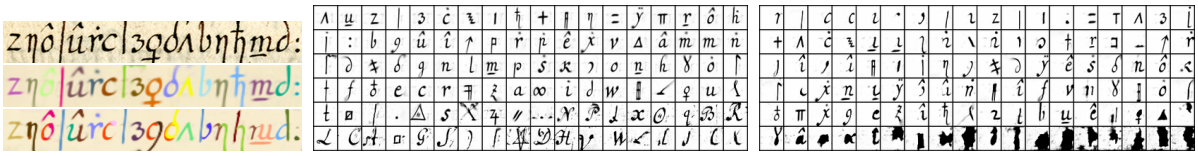
Metrics. Our goal is to capture the shape of all characters and position them precisely on each text line. Such fine-grained annotation is however not available in existing datasets. Instead, to provide a quantitative evaluation of our models, we report mean squared reconstruction error ("Rec." in our tables) and Character Error Rate (CER). CER is the standard metric for Optical Character Recognition (OCR). Given ground-truth and predicted sequences of characters, σ and $\hat{\sigma}$, it is defined as the minimum number of substitutions S, deletions D, and insertions I of characters needed to match the predicted sequence $\hat{\sigma}$ to the ground truth sequence σ, normalized by the size of the ground truth sequence $|\sigma|$:

$$CER(\sigma, \hat{\sigma}) = \frac{S + D + I}{|\sigma|}. \tag{3.6}$$

For simplicity, we ignore spaces. Predictions are obtained by selecting at every position the character associated to the most likely sprite. In the supervised setting, the association between sprites and characters is fixed at the beginning of training. In the unsupervised setting, we associate every sprite to a single character using a simple assignment strategy described in Appendix A (Sec. D).

3.4.2 Qualitative results

Examples of semantic segmentation and sprites in the supervised and unsupervised setting on Google1000 and Copiale are shown in Figs. 3.3,3.4 respectively. In the



(a) ↑ sup. ↓ unsup. (b) Sprites (sup.) (c) Sprites (unsup.)

Figure 3.4: **Results on the Copiale cipher [Knight et al., 2011].** Despite the high number of characters and their variability, our method learns meaningful sprites and performs accurate reconstructions in both settings **(a)**. We show the 108 most used sprites sorted by frequency in the supervised **(b)** and the unsupervised **(c)** settings.

unsupervised setting, several sprites (Figs. 3.3c,3.4c) can be used to reconstruct a single character. For example in Google1000, the "n" and "m" sprites are joined to better reconstruct "m". To account for appearance variation, multiple sprites are learned to reconstruct the most frequent character, e.g., "e" for Google and "c" in Copiale. These effects are even stronger in the handwritten Copiale dataset, where generic sub-character strokes are learned and used to better model characters' variations. In both datasets, some of the least used sprites do not correspond to characters, as they are never selected by the network, and thus are not properly optimized. These behaviors are expected in a completely unsupervised setting, because of the highly unbalanced statistics of the character frequencies and the ambiguity of reconstruction: without additional supervision, there is a clear benefit for the network to model well the variations of common characters, and to approximate or discard rare ones. This is a core limitation of existing unsupervised image decomposition approaches and a motivation for the introduction of our weakly supervised setting. In the (weakly) **supervised** setting, the sprites (Fig. 3.3b, 3.4b) closely correspond to the characters, except for very rare characters like the capital letter 'Z' for Google1000 (as can be seen in Fig. A.1 of Appendix A), while reconstruction is of very high quality and each character is reconstructed with the expected sprite.

Historical Font Reconstruction. In Fig. 3.5 we compare our learned sprites to manually extracted and binarized exemplars, where we observe that the learned sprites are mostly similar to the exemplars. Typefaces are sorted according to the average similarity between all the learned sprites and the manually extracted exemplars (between a-z and A-Z) that is computed using SSIM, as is a standard practice for font comparison [Srivatsan et al., 2021b]. SSIM is the highest for the antiqua font (0.745) and the lowest for the gotico-antiqua font (0.676). This seems correlated to the number of allographs that are present in each typeface. Antiqua is simple and standard, whereas

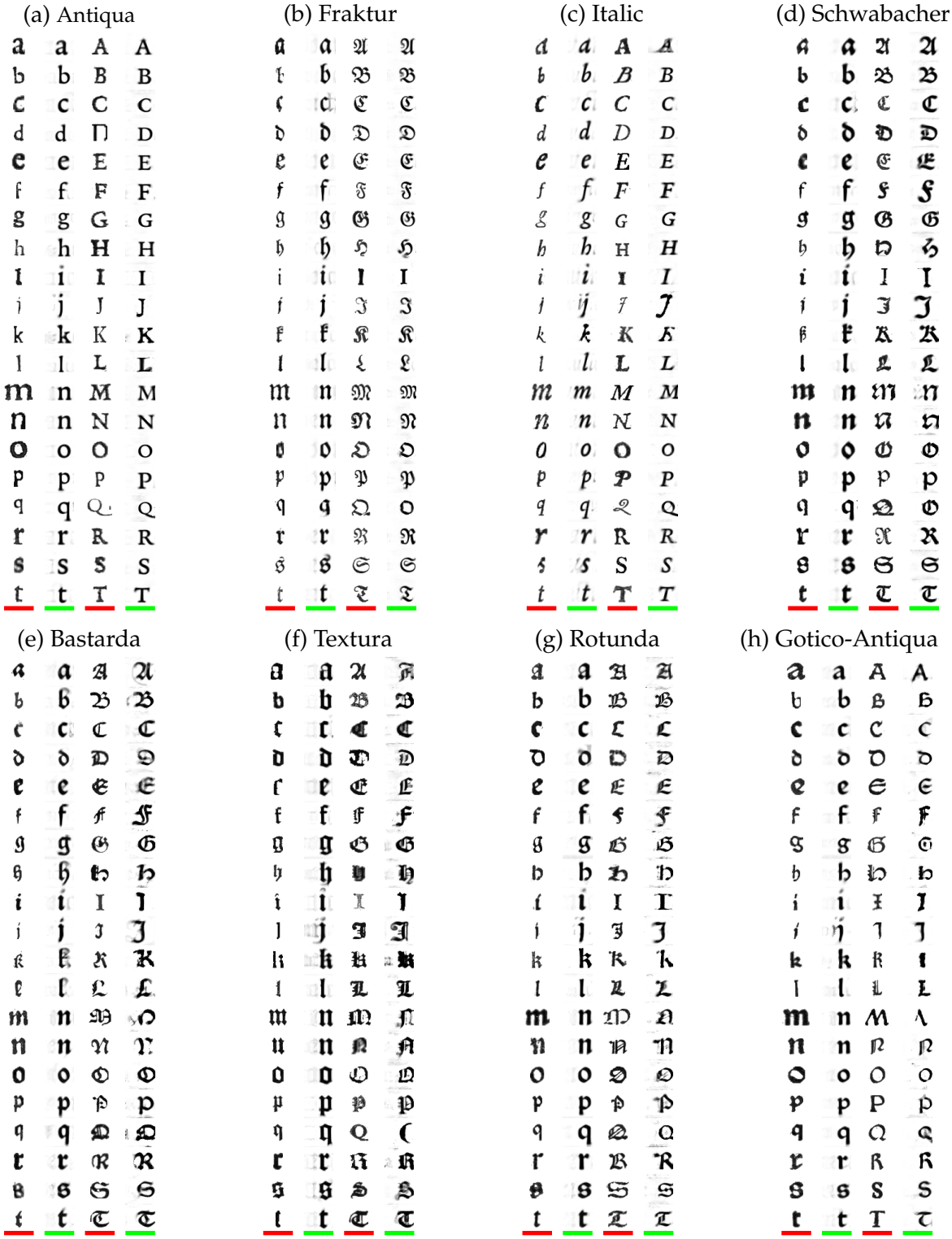


Figure 3.5: **Qualitative Evaluation on MFGR.** We compare a-t, A-T between learned sprites (ours), and manually extracted exemplars. Fonts are sorted by descending SSIM, computed on post-processed sprites (see Sec. C of Appendix A for more details). Note, that although fonts come from a single family, they may present *allographs*. For example, *Italic* contains different variants of "Q", where a random exemplar can significantly differ from the one summarized using our method.

gotico-antiqua is a hybrid between two visually distinct fonts, hence as our model learns a single sprite it fails to summarize them (e.g., see "T", "G", "I", "E", "F"). These results showcase the versatility of our approach, which is crucial for it to be applied for historical analysis.

3.4.3 Quantitative results

Our quantitative results and ablations for Google1000 and Copiale are reported in Tabs. 3.1,3.2 respectively.

For Google1000, the CER in the supervised setting is less than 1%, while it is 7.7% for the unsupervised setting. To provide baselines for these performances, we train and evaluate on our version of the dataset **(a)** ScrabbleGAN [Fogel et al., 2020], a supervised method with a standard recognizer and an additional generator module, **(b)** FontAdaptor [Zhang et al., 2020], a recent 1-shot method that learns to match single character exemplars to text lines, and **c)** an adaptation of the unsupervised DTI-Sprites [Monnier et al., 2020] to text lines which we detail in Appendix A (see Sec. E), where we also show that the vanilla MarioNette [Smirnov et al., 2021] has significantly worse results. Our unsupervised approach performs clearly better than our adaptation of DTI-Sprites and is almost on par with the 1-shot FontAdaptor, while our weakly supervised approach is almost on par with ScrabbleGAN. Our adaptation of DTI-Sprites is better at reconstructing images, but the learned sprites are much less meaningful, as shown by the poor CER performance. Interestingly, reconstruction is much better when using supervision, which hints that a better optimization scheme could improve unsupervised performances. We also evaluate the effect of varying the number of sprites K in the unsupervised setting. For K smaller than the actual number of characters (83), namely $K = 21$ and $K = 41$, we have a significant performance drop of 10% and 26% CER respectively, while increasing the number of characters to 166 and 332 doesn't significantly boost performances.

On the Copiale dataset, we compare our results with HTRbyMatching [Souibgui et al., 2020], a few-shot approach developed specifically for cipher recognition, using the same train/val/test splits. HTRbyMatching was evaluated on a wide range of few-shot scenarios, ranging from a scenario similar to FontAdaptor where a single exemplar is available for every character, to one where 5 exemplars are available for each character together with 5 completely annotated pages. Reported results are only for confident character predictions with different confidence thresholds, but summing the error rate of the predicted symbols and the percentage of non-annotated symbols, one can estimate the CER to vary between 10% and 47% depending on the

Method	Type	Rec. $\times 10^3$	CER
DTI-Sprites [Monnier et al., 2021]	unsup.	2.54	18.4 %
FontAdaptor [Zhang et al., 2020]	1-shot	-	6.7 %
ScrabbleGAN [Fogel et al., 2020]	sup.	-	0.6 %
Learnable Typewriter	sup.	3.5 ± 0.1	$0.85 \pm 0.03\%$
w\o shared z_k	sup.	3.3 ± 0.1	$0.89 \pm 0.06\%$
w\o p_θ	sup.	3.5 ± 0.1	$0.99 \pm 0.05\%$
w\o g_θ	sup.	3.4 ± 0.1	$0.88 \pm 0.04\%$
Learnable Typewriter	unsup.	7.1 ± 0.4	$7.7 \pm 0.6\%$
w\o shared z_k	unsup.	7.4 ± 0.4	$8.0 \pm 0.2\%$
w\o p_θ	unsup.	7.0 ± 0.3	$7.7 \pm 2.0\%$
w\o g_θ	unsup.	10.5 ± 0.7	$27.0 \pm 2.2\%$

Table 3.1: **Quantitative results and ablation on Google1000 [Vincent, 2007].** We report CER and mean squared reconstruction error for all the different approaches. For our method, we report the average of 5 runs and their standard deviation.

scenario[1]. This is consistent with the quantitative results we obtain with our approach, which are much better in the supervised setting (4.2%) and worse in the completely unsupervised one (52.6%). The low performance of the unsupervised approach is consistent with the qualitative results: given that many characters are reconstructed by sub-character sprites, one would have to associate sprite bi-grams to characters in order to obtain good CER performances. Interestingly, the reconstruction error is similar in the supervised and unsupervised setting, hinting that for this specific dataset, optimizing the reconstruction quality might not be enough to obtain relevant decomposition without additional priors. These results enable us to quantify and analyze a limitation of unsupervised image decomposition approaches on a more challenging dataset.

Note that the goal of our approach is not to boost CER performances - which in any case would be futile for Google1000 where the ground truth is already the result of an OCR model - but instead to learn character models and image decomposition. All these comparisons should be thus considered as sanity checks. Yet, it is possible to design post-processing algorithms to improve CER. We tested a simple algorithm where we assign a new sprite to the most frequent bi-grams and tri-grams, which leads to an improved CER for Copiale of 29.9%. However, we find this metric more

[1]Note, that the original paper calls SER our Character Error Rate which excludes spaces, corresponding to Symbol Error Rate, and not Sentence Error Rate.

Method	Type	Rec. $\times 10^2$	CER
HTRbyMatching [Souibgui et al., 2020]	few-shot	-	$10-47\%^*$
Learnable Typewriter	sup.	1.81 ± 0.01	$4.2 \pm 0.3\%$
w\o shared z_k	sup.	1.79 ± 0.01	$4.0 \pm 0.1\%$
w\o p_θ	sup.	1.77 ± 0.02	$4.7 \pm 0.1\%$
w\o g_θ	sup.	1.96 ± 0.07	$4.2 \pm 0.2\%$
Learnable Typewriter	unsup.	1.93 ± 0.02	$52.6 \pm 1.7\%$
w\o shared z_k	unsup.	1.89 ± 0.02	$47.6 \pm 2.8\%$
w\o p_θ	unsup.	1.81 ± 0.06	$51.9 \pm 2.0\%$
w\o g_θ	unsup.	3.99 ± 0.14	$80.6 \pm 0.9\%$

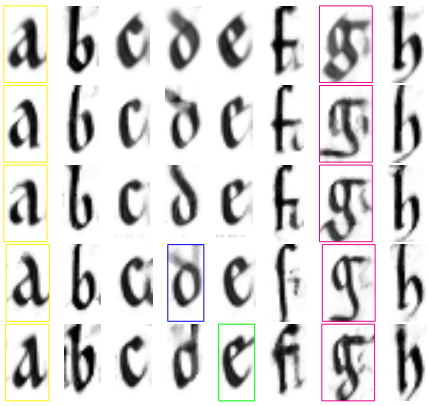
Table 3.2: **Quantitative results on Copiale [Knight et al., 2011].** We report CER and reconstruction error to evaluate both our selected baselines and our method. For our method, we report it across an average over 5 runs alongside its standard deviation. *See text for details.

informative when applied to the raw output of unsupervised image decomposition models.

In particular, we perform on both datasets an ablation of the architecture to better understand which design choices are critical. Interestingly, our results show that both in the supervised and the unsupervised setting, not sharing the latent codes z_k between the generation network and the sprite selection and even completely removing the probability network p_θ has limited influence on the performance clarifying that these design choices of MarioNette [Smirnov et al., 2021] are not of critical importance. Conversely, removing g_θ and directly learning prototypes as network parameters similar to DTI-Sprites [Monnier et al., 2021] has little impact in the supervised case, but leads to a significant drop in performance in the unsupervised one. A more detailed analysis of training curves reveals that in the unsupervised case, training is slower and leads to overfitting. While it might be possible to fix this issue by adapting the learning scheme for the prototypes, this shows that it is easier to learn the prototypes through a generator network than to optimize them directly.

3.5 Application to palaeography

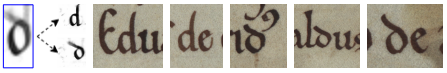
Understanding written character morphology, or *script type* is of central importance to palaeography, which seeks is to employ handwriting as historical evidence. Script types are for a handwritten manuscript what a font is for a printed one, yet as these prototypes don't exist outside the "scribe's mind" [Parkes, 1969; Stokes, 2011] and



(a) Sprites from document finetuning.

(b) Sprites and examples for “a” and “g”.

(c) Sprite and multiple examples for “e’.



(d) Learning dual sprites for “d” allographs.

Figure 3.6: **Sprites learned for similar documents of a Praegothica script.** Left: each line corresponds to a different document. Looking at any column, one can notice the small differences that characterize the handwriting in each document. Right: Colored boxed correspond to cases analyzed in more detail. Sprites summarize the key attributes of a character in each specific document, averaging its variations. Note the complexity of the documents: characters can overlap or be connected ligature, the parchment is often stained, and there are important intra-document character variations. *See text for more details.

are passed-by through a historical tradition, they come with a lot of written variation. Structuring this variation is crucial to adequately navigate such historical evidence [Stutzmann, 2018], because the analysis of script types cannot be reduced to a classification problem [Stutzmann, 2013, 2016; Hassner et al., 2015]. Moreover, such natural language descriptions may often be ambiguous in their process of being communicated across palaeographers and are often arbitrary as a means for comparison [Derolez, 2003]. In other words, as we discussed in Sec. 1.2, they are too Aristotelian [Studtmann, 2024]. Instead of a system of classification, what is more in need is a system for visual grounding [Ciula, 2017]. A simple solution, would be to choose an exemplar from a specific document or to ask a palaeographer to manually draw one that is “typical”. However, such processes are very time-consuming and might reflect their priors or biases, and lack interpretability on how they aggregate the existing variations of written morphology. Instead, we propose to continue previous research approaches such as the *System of Palaeographical Inspection* [Aiolli et al., 1999; Ciula, 2005], which tries to learn a hierarchy of “character prototypes”, similar to the ones of Eleanor Rosch as discussed Sec. 1.2, starting from the segmented characters of a document collection. Instead, in this Section we show how the Learnable Typewriter can extract prototypes directly from input text lines in a way that makes them visually

comparable, enabling a more systematic and interpretable quantitative palaeographic analysis.

Making Prototypes Comparable. Unlike Fig. 3.5 where prototypes are expected to have distinct differences, in palaeography, we expect to compare prototypes with subtle differences. Our goal is to analyze the written morphology across a granularity that can range from an individual document to a complete document collection. In order to properly compare their prototypes, our analysis would benefit from visual alignment. To approach this, our methodology is to first learn a *reference model* using a reference corpus and then finetune the model to reconstruct a subset of that corpus, up to the granularity of a specific document, by finetuning only the parameters that correspond to $g_\theta(z)$. Since the positioning, scaling, and coloring of the prototypes are shared, the prototypes will remain aligned, making them directly comparable. We first demonstrate our analysis approach in a small challenging dataset of Manuscripts from the Fontenay Abbey in Sec. 3.5.1, and then we move to a larger dataset of more established typology of Textualis Formata in Sec. 3.5.2 with the goal of performing a more systematic quantitative analysis.

3.5.1 Fontenay Manuscript

To first demonstrate the potential of our approach for palaeographic analysis, we apply it to a collection of 14 historical charters from the Fontenay abbey [Camps et al., 2022]. It contains digitized charters that originate from the Cistercian abbey of Fontenay in Burgundy (France) [Camps et al., 2022] and were created during the 12th and early 13th centuries. While they were carefully written and preserved, these documents are still very challenging (Fig. 3.6). Although they all use similar scripts from the *Praegothica* type, they also exhibit clear variations. Each of these documents has been digitized and each line has been manually segmented and transcribed. For our experiments, we selected a subset of 14 different documents sharing a similar script which falls into the family of praegothica scripts. These correspond to 163 lines, that use 47 distinct characters. They exhibit degradation, clear intra-document letter shape variations, and letters can overlap or be joined by ligature marks. Moreover, each document represents only a small amount of data, e.g., the ones used in Fig. 3.6 contain between 8 and 25 lines.

Qualitative Results Fig. 3.6a visualizes the sprites obtained for five different documents for the characters “a” to “h” and Fig. 3.6 highlights different aspects of these

results. Fig. 3.6b highlights the fact that the differences in the learned sprites correspond to actual variations in the different documents, whether subtle, such as for the "a" sprite, or clearer, such as for the descending part of the "g" sprite. Fig. 3.6c shows how a sharp sprite can be learned for the character "e", summarizing accurately its shape despite small variations throughout its different occurrences. Finally, Fig. 3.6d shows the case of a document in which two types of "d" co-exist. In this case, the learned sprite, shown on the left, reassembles an average of the two, where both versions of the ascending parts are visible on medium transparency. However, this limitation could be overcome by learning several sprites per character. Using our approach, we can learn two sprites per character, simply by summing their probabilities when optimizing the CTC-loss. We find that when allographs exist, i.e., different appearances of the same letter, these two different sprites do learn recover its two distinct appearances. In our example, we show the two different learned "d" sprites on the right of the original one.

3.5.2 Textualis Formata

Textualis formata is an established medieval gothic script which was continuously used for over three centuries (13th-15th century). To analyze the results of our approach, we adopt the taxonomy formalized by A. Derolez [Derolez, 2003], which provides a framework based on morphological criteria. Derolez makes a distinction between two subtypes of Northern and Southern Textualis (denoted as *NT* and *ST*) following their geographical location. However both types are followed by multiple distinct subtypes, that concern date, geographical origin, or language, which often intersect, questioning its fundamental distinction. To provide an analysis of this script type we compile a dataset of four train subfamilies and three test subfamilies for each class, providing a total of 892 lines. Our dataset selection is detailed in Appendix A (Sec. F.1).[2]

Quantitative Paleography We start by training multiple models as in Sec. 3.5.1 in order to obtain character prototypes at different levels of granularity: **(a)** a script type model for *Textualis*, **(b)** script subtype models for Northern and Southern *Textualis*, and **(c)** document level models for each document in our dataset. To quantitatively compare prototypes between **(b)** and **(c)**, we need to finetune from a common point of reference, for which we set the model **(a)** that was trained on all Textualis. Aligned

[2]Note, that this section is a reduction of [Vlachou-Efstathiou et al., 2024] focusing on its methodology which was my main contribution, as opposed to the palaeographic analysis which was the primary expertise of the first author.

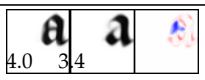
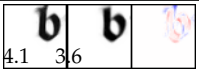
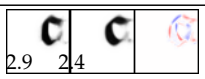
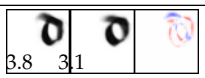
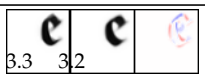
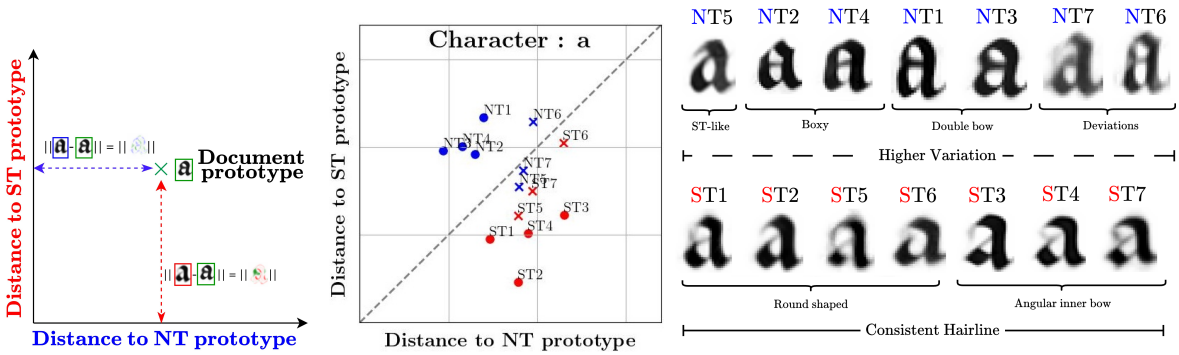
‹Ch.›	Derolez'criteria	NT | ST | diff. σ_{NT} σ_{ST}
‹a›	**NT**: Closed form with variations like "box-‹a›" **ST**: Open form or slightly closed with hairline	
‹b›	**NT**: Sloped or forked ascender tops **ST**: (i) Flat ascender tops, (ii) round lobe	
‹c›	**NT**: Angular or broken lobe curves **ST**: Semi-circular lobe	
‹d›	**NT**: (i) Lengthened and (ii) concave shaft **ST**: (i) Shorter shaft and (ii) almost horizontal, (iii) round bowl	
‹e›	**NT**: (i) Diagonal direction of the hairline and (ii) angular or broken lobe curves **ST**: (i) Horizontal or no hairline, (ii) semicircular lobe form	

Figure 3.7: Derolez' criteria for Northern and Southern *Textualis* and our subtype prototypes.

prototypes can now allow us to subtract them in pixel space and thus quantify their difference in an interpretable way. To optimally and reliably compare prototypes, we develop a post-processing procedure and quality evaluation which we describe in Appendix A (Sec. F.2). To make this difference easier to understand, we use a color map that represents zeros as white, and positive and negative values as two distinct colors, typically red and blue. By revealing pixel-wise differences, this method facilitates an initial qualitative examination of morphological disparities.

Consistency with classical Palaeography. In Fig. 3.7 we systematically check how the extracted observations from prototype comparisons conforms to Derolez's general morphological criteria for both Northern and Southern *Textualis* prototypes, highlighting their variations by visualizing their difference. While only ‹a›-‹e› are shown in the table, more complete results can be found on our original publication [Vlachou-Efstathiou et al., 2024]. In short, we find that Derolez's observations closely align with the variations that our prototypes enable us to visualize, which shows that our method conforms to the classical palaeographic analysis. Additionally, we report the standard deviation across prototypes for σ_{NT} and σ_{ST} for each letter, which were consistently higher for Northern *Textualis*, which is consistent with Derolez's claim that this script subtype generally exhibits higher intra-class variation.

Morphological EDA with Character Graphs. Having prototypes that characterize different groups of instances from certain manuscripts allows us to perform intra-class



(a) Graph interpretation (b) Character graph (c) Sprite Observations

Figure 3.8: **Comparison graph pipeline.** For a given document we can locate how far a manuscript prototype is to the prototypes of two reference families by computing their L2-distance **(a)**. Mapping this way allows us to locate sub-groups of prototypes **(b)** which have clear qualitative interpretation **(c)**. *See text for more details.

analysis. To allow quantitative exploratory data analysis, we introduce a comparison graph, illustrated in Fig. 3.8a. In this graph, each point represents a specific document character prototype, with its coordinates defined as its distance in pixel space to two selected prototypes computed by finetuning on the full manuscript that corresponds to each typology. Blue and red, markers signify Northern, and Southern *Textualis* documents, respectively. Each document's identifier is written near its marker.

Analyzing the Character ‹a›. The letter ‹a› is often considered as a discriminative criterion between script types, so much so that W. Oeser [Oeser, 1971] distinguished seven categories within the Northern *Textualis* script subtype mainly based on allographs of ‹a›. The dispersion of the characters on the graph in Fig. 3.8b provides insight into the variability of ‹a› in this subtype. The group associated to NT1-4 corresponds to the closed "box-a" form in NT2 ( ) and NT4 ( ) and the double-bow variant in NT1 ( ) and NT3 ( ). NT6 presents a more vertically elongated form which stands out ( ). Most striking in our ‹a› character graph is that the prototypes for NT5 ( ) and NT7 ( ) are actually closer to the ST prototypes. This is consistent with the observation that open ‹a› forms are standard for ST. While there are morphological variations across ST documents, with round shapes (ST1 ; ST2 ; ST5 ; ST6), or with more angular inner bows (ST3 ; ST4 ; ST7), the consistent use of an open form, or only closed with a hairline, distinguishes them from the NT subtype, and all ST documents prototypes are consistently closer to the ST prototype. Plotting this graph across documents in Fig. 3.9 allows us to easily identify that **NT5** stands out in the graphs as an outlier document, as seven character prototypes are closer to ST than to

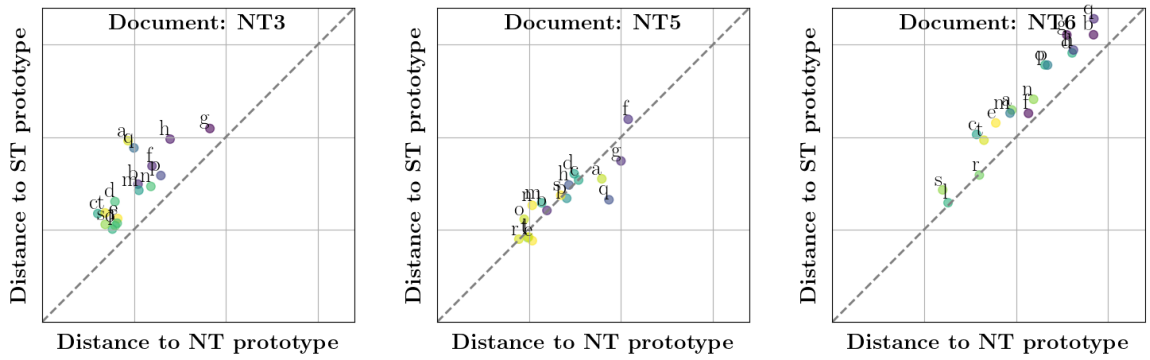


Figure 3.9: **Document comparison graphs.** Visualizing all letters for different documents allows us to identify the document NT5 as an outlier of the Northern Textualis family.

NT prototypes, with a particular difference for ‹a,g,q› ( ). These observations can be summarized in Fig. 3.8c.

3.6 Conclusion

We have presented a document-specific generative approach to document analysis. Inspired by deep unsupervised multi-object segmentation methods, we designed an end to end differentiable approach to accurately model text line images of manuscripts through a set of learned sprites. We outlined that a completely unsupervised approach suffers from the ambiguity of the decomposition problem and the imbalanced character distributions. Therefore, we extended these approaches using weak supervision to obtain robust, high-quality results. These allow us to learn prototypes for the characters of both standard printed documents and of much more complex ones, such as a handwritten ciphered manuscript or ancient charters. Finally, we demonstrated the potential of our approach for a novel application: palaeographic analysis. We extended our approach to a methodology for interpretable qualitative and quantitative palaeography, that can integrate and complement traditional historical approaches.

Chapter 4

Diffusion Models as Data Mining Tools

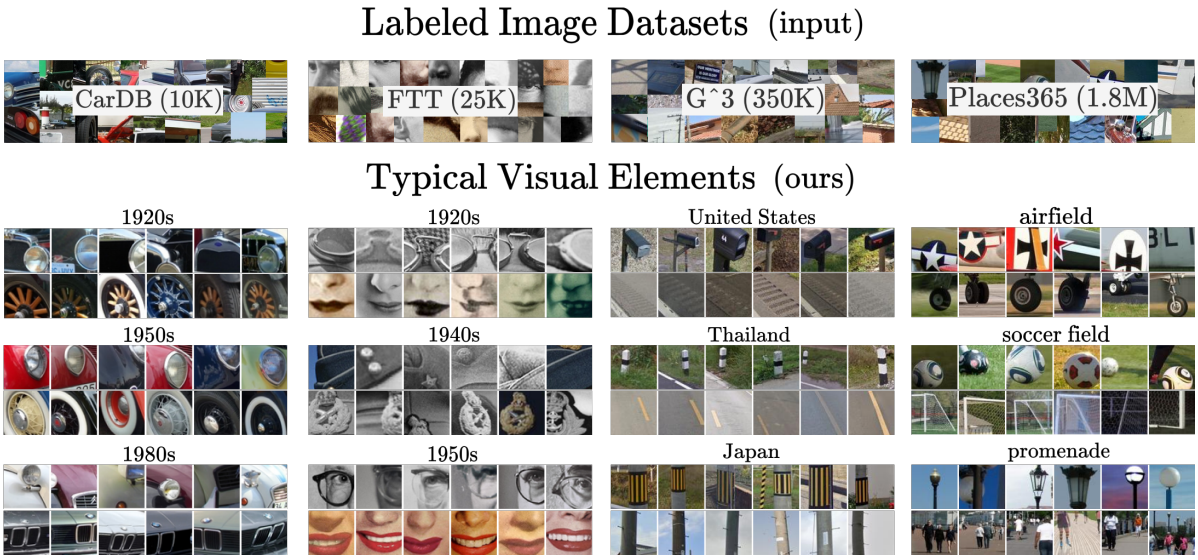


Figure 4.1: **Mining typical visual elements with diffusion models.** We demonstrate how to use diffusion models to mine visual data through a simple pixel-based score and a standard clustering approach. We present high-quality mining results for a diverse range of datasets (from left to right: 10,130 photographs of cars tagged with a creation year between 1920-1999 [Lee et al., 2013], 24,874 portraits from the 19th to the 21st century [Chen et al., 2023], 344,224 StreetView images tagged with country names [Luo et al., 2022], and 1,803,460 images of scenes images associated with descriptive names [Zhou et al., 2017a]). Our results highlight both expected elements and more unforeseen ones.

4.1 Introduction

As we discussed in Sec. 2.2.2 traditional image data mining aims to discover patterns within large visual corpora such as collections of StreetView panoramas [Doersch

et al., 2012; Lee et al., 2015b], historical images of faces [Ginosar et al., 2017; Chen et al., 2023] or photographs of cars [Lee et al., 2013; Dalens et al., 2019]. This would often be achieved through pairwise comparison of candidate patches in order to discover which ones of them are both frequent and discriminative [Singh et al., 2012]. This chapter proposes a novel idea: to turn generative models trained for image synthesis into a scalable method for large scale mining of image datasets. Generative models digest massive amounts of data, which they implicitly store in their weights. Our central insight is that we can use this learned summary of the visual input to identify the most typical image regions. This unconventional use of a diffusion model for studying its training data demonstrates that generative models are potent tools beyond synthesis—for data mining, summary, and understanding.

As we discussed in Sec. 1.3 our target task, mining for informative visual patterns, is challenging. Unlike text, where words act as discrete tokens that we can directly compare, the visual world seldom contains exactly repeating elements. Even common simple visual elements, such as windows, can have different colors and different numbers of panes; they may be seen from various viewpoints, and they may be located at multiple positions as part of different facades. The standard approach to image data mining [Doersch et al., 2012; Lee et al., 2013; Shen et al., 2021a] which we discussed in Sec. 2.2.2, involves learning data-specific similarities with relevant invariances (*e.g.*, such that different-looking windows will be similar) and using them to search for discriminative patterns. However, these techniques are not easily scalable since one must apply them across all pairs of visual elements within all pairs of images in the dataset. In other words, the similarity graph between visual elements scales quadratically with the size of the dataset. In contrast, our proposed analysis-by-synthesis approach does not require pairwise comparisons between different visual elements and thus scales to very large datasets.

The approach we propose takes as input a dataset with image-level tags, such as time [Lee et al., 2013; Chen et al., 2023], geography [Luo et al., 2022], or scene labels [Zhou et al., 2017a]. Our goal is to provide a visual summary of the elements typical of the different tags, such as the common elements that enable us to determine the location of a StreetView panorama. To arrive at this summary, we first finetune a conditional diffusion model on the target dataset. We then use the finetuned model to define a pixel-wise typicality measure by assessing the degree to which the label conditioning impacts the model's reconstruction of an image. We mine visual elements by aggregating typicality on patches, selecting the most typical ones, and clustering them using features extracted from the finetuned model [Tang et al., 2023]. As visu-

alized in Fig. 4.1, this leads to clusters of typical visual elements that summarize the most characteristic patterns associated with the tags available in the input dataset. For example, our face results highlight iconic elements, such as aviator glasses in the 1920s and military hats in the 1940s, and more subtle details, such as period-typical glasses or make-up. Interestingly, our results on StreetView data highlight details that are similar to the ones presented in geographical understanding websites [geodummy, 2023; geohints, 2023; Plonkit, 2023], popularized through the GeoGuessr game [geoguessr, 2023], such as typical parts of utility poles, bollards, or architecture. To our knowledge, no existing visual mining method has demonstrated such high-quality results on diverse datasets.

Contributions. To summarize, we present:

- a typicality score that can be formally derived from a diffusion model, allowing for an efficient extraction of the most typical visual elements of a dataset,
- a pipeline to extract and cluster typical elements in order to create typical summaries of different datasets, including cars [Lee et al., 2013], portraits [Chen et al., 2023], geographical data [Luo et al., 2022], and scenes [Zhou et al., 2017a],
- further applications of our method in locating elements which are typical across location, visualizing the bias of a diffusion model, and localizing abnormalities in chest X-ray images.

4.2 Related Work

Image data mining. Image data mining turned the manual and subjective process of comparing photographs (*e.g.*, [Kotchemidova, 2005]) into algorithmic methods for summarizing image data, such as architectural details [Doersch et al., 2012; Lee et al., 2015b], fashion [Ginosar et al., 2017; Matzen et al., 2017; Chen et al., 2023], industrial design [Jae Lee et al., 2013], and art [Shen et al., 2019, 2021b; Kaoua et al., 2021] by locating visual patterns. As we discussed in Sec. 2.2 this has mainly been achieved using techniques such as discriminative clustering. For example, [Lee et al., 2013] demonstrated how correspondence based mining across time can be achieved in a dataset of objects of similar parts, namely cars, and [Doersch et al., 2012] showed that geographically representative image elements can be automatically discovered from Google StreetView imagery in a discriminative manner. However, these traditional data mining approaches do not scale to large modern datasets. Indeed, they require pairwise comparisons between all the visual elements of each image to the entire

dataset in order to locate nearest neighbors and establish clusters. Notably, the discriminative clustering algorithm of [Doersch et al., 2012] requires training a separate linear SVM detector for each visual element- a computationally prohibitive approach when considering multiple possible visual elements for the purposes of analysis. In contrast, our approach is scalable to very large datasets. Closer to our work, generative models have been trained to analyze the evolution of faces [Chen et al., 2023] and cars [Dalens et al., 2019] across time, and the change in geography across GPS [Feng et al., 2025]. However, these works essentially focus on conditional image translation, and do not try to mine typical elements in their datasets.

Diffusion models. Diffusion models have gained popularity in recent years due to their stability in training and effectiveness in modeling complex multimodal distributions [Sohl-Dickstein et al., 2015; Ho et al., 2020; Dhariwal and Nichol, 2021; Ho et al., 2022; Karras et al., 2022]. These models are capable of generating high-quality imagery conditioned on input signals beyond categorical labels, like text [Rombach et al., 2022; Saharia et al., 2022; Ramesh et al., 2022], and can further incorporate additional modalities [Zhang et al., 2023; Li et al., 2023c]. In addition to generating images from scratch, diffusion models have been used extensively for instruction-driven image-to-image translation [Meng et al., 2022; Hertz et al., 2022; Mokady et al., 2023; Tumanyan et al., 2023; Brooks et al., 2023]. It has also been shown that pre-trained text-to-image diffusion models encode strong priors for natural scenes, allowing their internal features to be used for secondary tasks [Xu et al., 2023; Luo et al., 2023; Tang et al., 2023]. They can easily be adapted for new tasks or to new data distributions through minimal finetuning [Brooks et al., 2023; Zhang et al., 2023; Ruiz et al., 2023].

Beyond mere image synthesis, generative image models, and in particular diffusion models, have been studied as data augmentation engines. While most machine learning approaches treat the data as fixed and improve the learning algorithm, works such as [Jahanian et al., 2022; Chai et al., 2021; Azizi et al., 2023; Tian et al., 2023; Fan et al., 2024] fix the learning algorithm and augment the training data, using generative models to synthesize large amounts of synthetic training data.

In contrast, we present a new way to use generative models, with the goal of gaining insights about their training data.

4.3 Data Mining via Diffusion Models

Given a collection of images with assigned labels, our goal is to extract a small subset of visual elements from these images that best summarize the label inside the context of

the input dataset, or in other words that are highly typical of their label [Murphy, 2004]. The goal of our approach is to use generative probabilistic image synthesis models towards that end. We rely on finetuning a conditional stable-diffusion model [Rombach et al., 2022] trained for image synthesis, using it to extract a summary of the visual world. We start by reviewing diffusion models and the techniques we leverage in Sec. 4.3.1. In Sec. 4.3.2, we introduce our measure of typicality, which allows us to measure how the class label conditioning affects the synthesis of an image by the diffusion model. In Sec. 4.3.3, we describe how we aggregate typicality on patches to mine typical visual elements and cluster them to summarize the training data.

4.3.1 Preliminary

Diffusion models. Diffusion models are generative models trained to transform random noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{1}) \in \mathbb{R}^{H \times W}$ of height H and width W into a target prior distribution of images $p(\boldsymbol{x}) \in \mathbb{R}^{H \times W}$ [Sohl-Dickstein et al., 2015; Ho et al., 2020]. They achieve this by learning to revert a noising process called the *forward process* where noise is linearly interpolated, with a decreasing mixing strength $a_t \in [0, 1]$ to the input image $\boldsymbol{x}_0$ in different resolutions indexed by a fractional timestep index $t \in [0, 1]$:

$$\boldsymbol{x}_t^{\boldsymbol{\epsilon}} = \sqrt{a_t}\boldsymbol{x} + \sqrt{1 - a_t}\boldsymbol{\epsilon}. \tag{4.1}$$

The diffusion model $\epsilon_\theta(\boldsymbol{x}_t^{\boldsymbol{\epsilon}},\ t) \in \mathbb{R}^{H \times W}$ with parameters θ, is trained to predict the input noise ϵ added to that timestep, by minimizing the reconstruction loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{\boldsymbol{\epsilon}, t, \boldsymbol{x} \sim p(x)} \left[\left\| \epsilon_\theta(\boldsymbol{x}_t^{\boldsymbol{\epsilon}},\ t) - \boldsymbol{\epsilon} \right\|^2 \right], \tag{4.2}$$

At test time, the target distribution can be sampled by gradually passing an input noise through an iterative denoising process, known as the *backward process* [Ho et al., 2020; Song et al., 2021], which is based on an inversion of eq. 4.1:

$$\hat{\boldsymbol{x}}_{0,t} = \left(\boldsymbol{x}_t^{\boldsymbol{\epsilon}} - \sqrt{1 - a_t}\epsilon_\theta(\boldsymbol{x}_t^{\boldsymbol{\epsilon}},\ t) \right) / \sqrt{a_t}, \tag{4.3}$$

where $\hat{\boldsymbol{x}}_{0,t}$ is the denoised estimate at that timestep t. For this sampling procedure where noise is iteratively added and removed from $\hat{\boldsymbol{x}}_{0,t}$ from coarse to fine resolutions, various formulations have been proposed including DDPM [Ho et al., 2020] and

its non-markovian counterpart DDIM [Song et al., 2021] whose goal is to factor out stochasticity and speed up sampling.

Conditional Diffusion Models. In order to make diffusion models sample from a conditional distribution $p(\boldsymbol{x}|c)$, [Nichol and Dhariwal, 2021] train $\epsilon_\theta(\boldsymbol{z},\ t)$ on $p(\boldsymbol{x})$, and steer it during sampling towards c, using a trained classifier $d_\theta(c|\boldsymbol{x})$, by simply adding its gradient to ϵ_θ for a given class c according to $\boldsymbol{x}$: $\epsilon_\theta(\boldsymbol{x}_t^\epsilon,\ t) + \lambda \nabla_{\boldsymbol{x}} d_\theta(c|\hat{\boldsymbol{x}}_{0,t})$, with a strength λ. In a later work [Ho and Salimans, 2021] showed that this is equivalent with training a diffusion model $\epsilon_\theta(z,\ t,\ c)$ on $p(\boldsymbol{x}|c)$, with c:

$$\mathcal{L}(\theta) = \mathbb{E}_{\boldsymbol{\epsilon},t,(\boldsymbol{x},c)\sim p(x|c)} \left[\|\epsilon_\theta(\boldsymbol{x}_t^\epsilon,\ t,\ c) - \boldsymbol{\epsilon}\|^2 \right], \tag{4.4}$$

while dropping it for a small percent of the time (10%) in order to learn an unconditional (or "null") distribution $p(\boldsymbol{x}|\varnothing)$, and during sampling replacing $\nabla_{\boldsymbol{x}} d_\theta(c|\hat{\boldsymbol{x}}_{0,t})$ with $\epsilon_\theta(\boldsymbol{x}_t^\epsilon,\ t,\ c) - \epsilon_\theta(\boldsymbol{x}_t^\epsilon,\ t,\ \varnothing)$.

Latent diffusion models. Our work employs a variant of conditional diffusion models, trained with classifier free guidance, known as a *latent* diffusion model (LDM) [Rombach et al., 2022]. Instead of directly modeling the source data distribution $\boldsymbol{x}_0$ $p(\boldsymbol{x}|c)$, LDMs model the distribution of $\boldsymbol{x}_0$ in the latent space of a variational autoencoder $v_\phi(\boldsymbol{x})$ [Kingma and Welling, 2014]. Working in the latent space reduces the complexity of the data distribution. It thus significantly reduces both the number of parameters of the diffusion model and the amount of training samples necessary to learn a good model. As a conditioning they also use pretrained CLIP [Radford et al., 2021] text features $\tau_\phi(c)$, which are multimodal embeddings of text and images, trained on a large dataset of image-text pairs.

$$\mathcal{L}(\theta) = \mathbb{E}_{\boldsymbol{\epsilon},t,(\boldsymbol{x},c)\sim p(v_\phi(\boldsymbol{x})|\tau_\phi(c))} \left[\|\epsilon_\theta(\boldsymbol{x}_t^\epsilon,\ t,\ c) - \boldsymbol{\epsilon}\|^2 \right], \tag{4.5}$$

Diffusion Classifier. As conditional diffusion models are probabilistic models of the form $p(\boldsymbol{x}|y)$ they could be inverted in order to be used as classifier $p(y|x)$. Using the fact that the $-\mathcal{L}(\theta)$ is maximized as a lower bound of the log-likelihood, [Li et al., 2023a] through a standard application of the Bayes rule defined a classifier $p_\theta(c_i|x)$ for a set of labels c_i and an image $\boldsymbol{x}$ as:

$$p_\theta\left(c_i|\boldsymbol{x}\right) = \frac{1}{\sum_j \exp\left\{\mathbb{E}_{\epsilon,t}\left[L_t(\boldsymbol{x},\epsilon,c_i) - L_t(\boldsymbol{x},\epsilon,c_j)\right]\right\}}, \tag{4.6}$$

$$L_t(x,\epsilon,c) \equiv \|\epsilon_\theta(\boldsymbol{x}_t^\epsilon,\ t,\ c) - \boldsymbol{\epsilon}\|^2. \tag{4.7}$$

4.3.2 Typicality

For the purpose of mining we would like to define a typicality measure, that can enable us to sort visual elements coming from images of a specific class by how typical they are of that class. To define that measure we require that it can be computed both **(a)** efficiently, and **(b)** over different regions of the input image, so that we can discover visual structure. Given that $\varnothing$ is used to model the unconditional distribution, we can simply sum the denominator of eq. 4.6 only across the classes c and $\varnothing$ to:

$$p_\theta\left(c|\boldsymbol{x}\right) = \frac{1}{1 + \exp\left\{\mathbb{E}_{\epsilon,t}\left[L_t(\boldsymbol{x},\ \epsilon,\ c) - L_t(\boldsymbol{x},\ \epsilon,\ \varnothing)\right]\right\}}. \tag{4.8}$$

We can quickly observe that for the above expression to increase the denominator needs to maximize. Thus, for the purposes of only rank elements one only needs to compute the expression:

$$\mathbf{T}(\boldsymbol{x}|c) = \mathbb{E}_{\epsilon,t}[L_t(\boldsymbol{x},\ \epsilon,\ \varnothing) - L_t(\boldsymbol{x},\ \epsilon,\ c)], \tag{4.9}$$

where $\mathbf{T}$ is our derived measure of *typicality* of an image x given the ground truth class label conditioning c and the null conditioning $\varnothing$. Intuitively, an image is typical of a conditioning class label (*e.g.*, a country's name or a date) if the diffusion model is better at denoising the input image in the presence of that label than in its absence. However, note that instead of computing this measure across the whole image, one can compute it for any subregion π (down to the individual pixel), using a binary mask:

$$\mathbf{T}^\pi(\boldsymbol{x}|c) = \mathbb{E}_{\epsilon,t}[L_t^\pi(\boldsymbol{x},\ \epsilon,\ \varnothing) - L_t^\pi(\boldsymbol{x},\ \epsilon,\ c)], \tag{4.10}$$

$$L_t^\pi(\boldsymbol{x},\ \epsilon,\ c) \equiv \|\pi \cdot (\epsilon_\theta(\boldsymbol{x}_t^\epsilon,\ t,\ c) - \boldsymbol{\epsilon})\|^2. \tag{4.11}$$

While when computed for the whole image typicality ranks the same as a binary classifier, when computed per pixel it is equivalent to a measure known as pixel-wise mutual information, as we discuss in Appendix B (Sec. A). However, being able to compute typicality per patch is what enables us to discover visual elements.

4.3.3 Mining for Typical Visual Elements

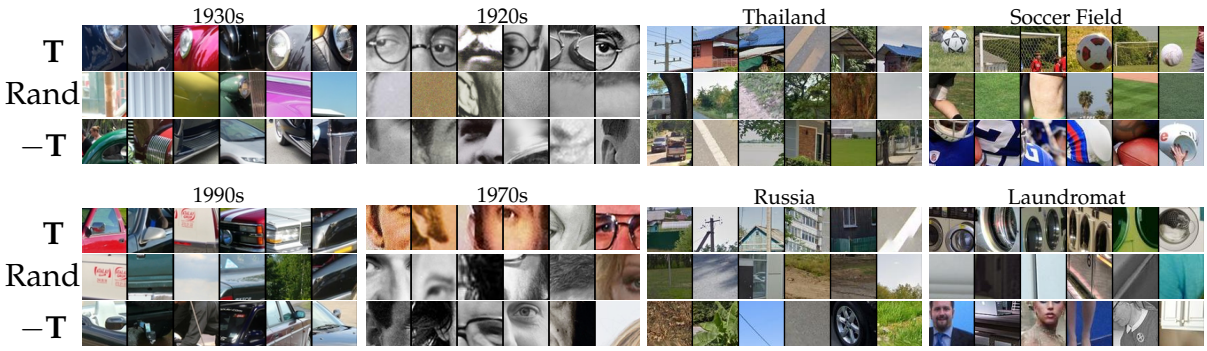
Patch-based analysis. To locate typical elements, we compute our typicality scores over all the patches π of an input image. This can be computed efficiently by computing $(\epsilon_\theta(\boldsymbol{x}_t^\epsilon,\ t,\ c) - \boldsymbol{\epsilon})$ and then using an average pooling operator with a fixed patch size. To extract "mid-level" patches we use a patch size of 50 for images with a base dimension of 256 and 64 for images with a base dimension of 512. Then, to identify the set of most typical visual elements for a dataset we pick the 5 most typical non-overlapping patches in each image according to the patch typicality, and select the 1000 most typical patches over all the dataset. Unlike [Li et al., 2023a], we find that reducing the sampled range of t to $[0.1, 0.7]$ improves the quality of our results, as the tails can contribute uninformative yet typical samples (see Sec. C of the Appendix B).

Clustering visual elements. We cluster the most typical patches using k-means [Lloyd, 1982] with 32 clusters. To cluster elements, we embed them with DIFT [Tang et al., 2023] features, computed at timestep $t = 0.161$ using our finetuned models. For visualization, we rank clusters by the median typicality of their elements in decreasing order and their elements by their distance to the centroid in increasing order.

Conditioning and finetuning. Given that our input dataset is labelled, we convert its labels to sentences: "A car/portrait from the {decade}s." for faces and cars ("A car/portrait." for the null conditioning $\varnothing$), "A Google StreetView image of {country}." for StreetView data ("A Google StreetView image." for the null conditioning $\varnothing$), and "An image of {scene}." for images of the Places dataset [Zhou et al., 2017a] (empty string for the null conditioning $\varnothing$). A latent diffusion model [Rombach et al., 2022] is then finetuned on the target dataset by optimizing the reconstruction loss (Equation 4.5) given the conditioning. We use Stable Diffusion V1.5 [Rombach et al., 2022] as a base model in all our experiments.

4.4 Experiments

We showcase the effectiveness of our approach in summarizing visual data in a wide variety of datasets. First, in Sec. 4.4.1, we introduce the datasets used in our experiments. Second, in Sec. 4.4.2, we evaluate the ranking produced by our typicality measure. Third, in Sec. 4.4.3, we discuss our main result, the mined visual summaries of the analyzed datasets, and compare it with [Doersch et al., 2012]. Finally, we discuss the limitations of our approach in Sec. 4.4.4.



(a) **CarDB** (b) **FTT** (c) **G^3** (d) **Places**

Figure 4.2: **Typical elements are informative of the conditioning label.** We visualize the top-6 patches ranked according to typicality (**T**) with respect to the conditioning class label, negative typicality (−**T**), and randomly (Rand.). The two rows correspond to different classes from each of the four datasets.

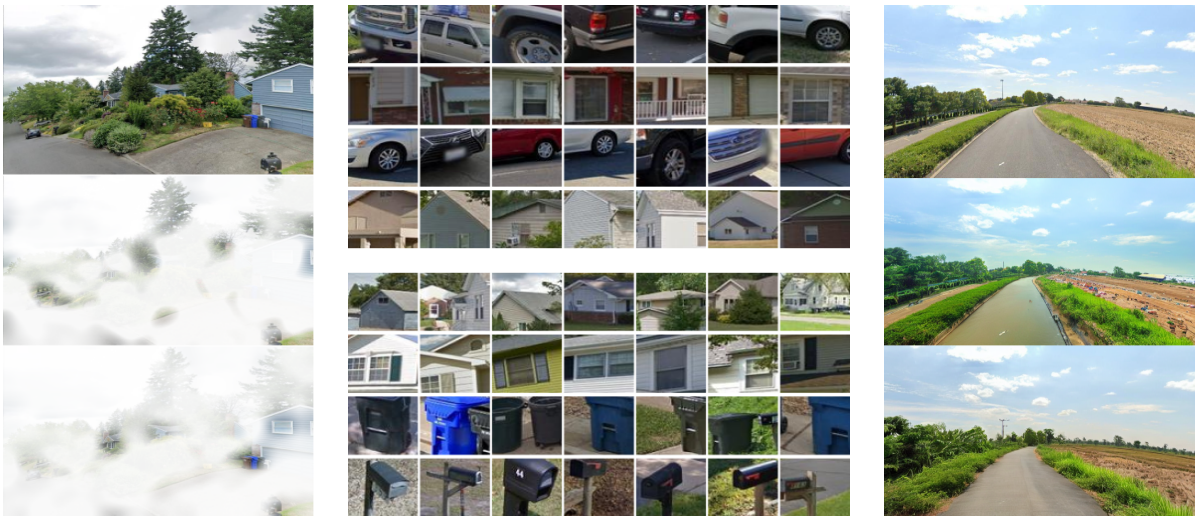
4.4.1 Datasets

We experiment with four diverse datasets. CarDB [Lee et al., 2013] and FTT [Chen et al., 2023] have already been used for image data mining and include a few tens of thousands of images. G^3 [Luo et al., 2022] and Places [Zhou et al., 2017a] are much larger with 344K and 1.8M images respectively, and to our knowledge have never been used for mining purposes.

Cars. The CarDB dataset [Lee et al., 2013] contains 10,130 photos of cars from 1920 to 1999, collected from cardatabase.net. They are labeled with creation years, which we bin into decades for our analysis. This dataset contains cars seen from various viewpoints and in diverse environments. As a result, extracting time-informative elements is challenging. We rescale all images to a height of 256 pixels while preserving their original aspect ratio.

Faces. The Faces Through Time (FTT) Dataset [Chen et al., 2023] contains 24,874 images of notable people from the 19th to 21st century, with roughly 1,900 images per decade, sourced from Wikimedia Commons. All photos come in 256x256 pixels.

Geo. The G^3 [Luo et al., 2022] dataset contains images obtained from crops of StreetView panoramas, diversely sampled worldwide, of which we selected 344,224 images, which we rescaled to 512x756 pixels. This dataset is challenging because of the small details that characterize a scene's appearance and scale. We focus on the 8 countries with the largest number of panoramas (United States, Japan, France, Italy, United Kingdom, Brazil, Russia, and Thailand) and select two countries with much



(a) Typicallity (b) Clusters (c) Parallel Translation

Figure 4.3: **Effect of finetuning**. **(a)** For the same USA image (top), finetuning changes the spatial allocation of typicality before (middle) and after (bottom) finetuning. **(b)** This results in different typical clusters (USA), which, after finetuning (bottom), select for more typical elements like mailboxes. **(c)** Translation (Sec. 4.5.1) of a picture of a road from France (top) to Thailand without finetuning (middle) suffers from data biases in the base model turning the road into a river and erasing utility poles. After finetuning on the G^3 dataset (bottom), the translated image is more consistent with the original.

fewer images (Nigeria and India). We finetune the network using all images from these countries, but we only mine a random subset of 1000 images for each country.

Places. The high-resolution version of the Places dataset [Zhou et al., 2017a] contains 1,803,460 million images from 365 place categories associated with their labels, with a minimum dimension of 512 pixels. For mining, we only use the validation dataset, which contains 100 images per scene category.

4.4.2 Typicality Measure Evaluation

Typicality score for patches. Fig. 4.2 shows the most and least typical patches according to our typicality measure and random patches from the four datasets. We note that the most typical patches are unique to each class and more discriminative than random patches, while the least typical patches are uninformative of the label c.

Effect of finetuning. Unsurprisingly, we found that finetuning the diffusion model on the dataset of interest was critical to the quality of our results. First, on a given image, finetuning changes the spatial distribution of typicality, prioritizing elements more

(a) **1920s** (b) **1930s** (c) **1980s**

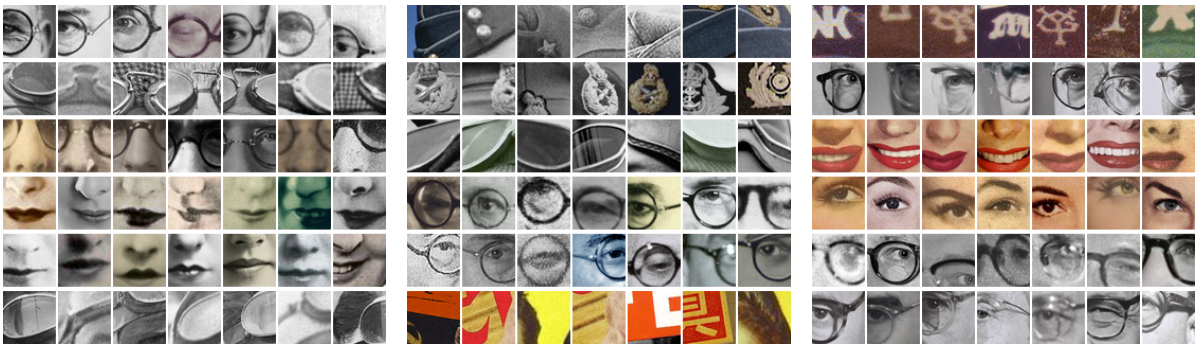
Figure 4.4: **Clusters of CarDB [Lee et al., 2013] visual elements.** Our visual summaries of typical car elements show elements unique to a period and elements that evolve with time. Evolving elements include the shapes of the car's body or headlights, which are parts of the 6 most typical clusters for most periods. More specific elements include running boards in the 1920s (**(a)**, 6th row) or large engine side grills in the 1930s (**(b)**, 3rd, 4th and 6th row). In the 1980s **(c)**, we observe two typical yet very discrete clusters of car design styles, of the curvy French 2CV (1-4 row) juxtaposed to the square American *chevy*-style cars (5-6 rows).

correlated with the training labels (see Fig. 4.3a). Second, in Fig. 4.3b, we show the most typical clusters identified before and after finetuning. The patches selected after finetuning avoid the biases in the training data of the base model and are more specific to the G ^3 dataset, identifying elements such as post-boxes. We also demonstrate this quantitatively in Sec. 4.5.3 for our application to X-ray images. Third, finetuning enables better translation between labels (see Sec. 4.5.1), as can be seen in Fig. 4.3c, allowing vegetation, roads, road tracks, and utility poles to be translated consistently across classes in the parallel dataset, which can be located in Appendix B (Sec. F).

4.4.3 Clusters of Typical Visual Elements

In this section, we analyze the visual summary of each dataset, obtained by clustering the typical visual elements for the different class labels. We demonstrate the mined summaries of Cars, Faces, Geo, and Scenes in Figs. 4.4, 4.5, 4.6, 4.7 respectively. For all datasets we show for selected class labels, the top-6 clusters ranked by median typicality of their elements. Inside each cluster elements are ranked by their distance to the centroid. The resulting clusters are analyzed inside the figure captions for ease of viewing. Complete clusters can be found in the Appendix B (Sec. F).

Comparison to [Doersch et al., 2012]. As the Matlab implementation of [Doersch et al., 2012] is obsolete and hardware-specific, we reimplement their method in Python and release this reimplementation with our code. In Fig. 4.8, we show the results of



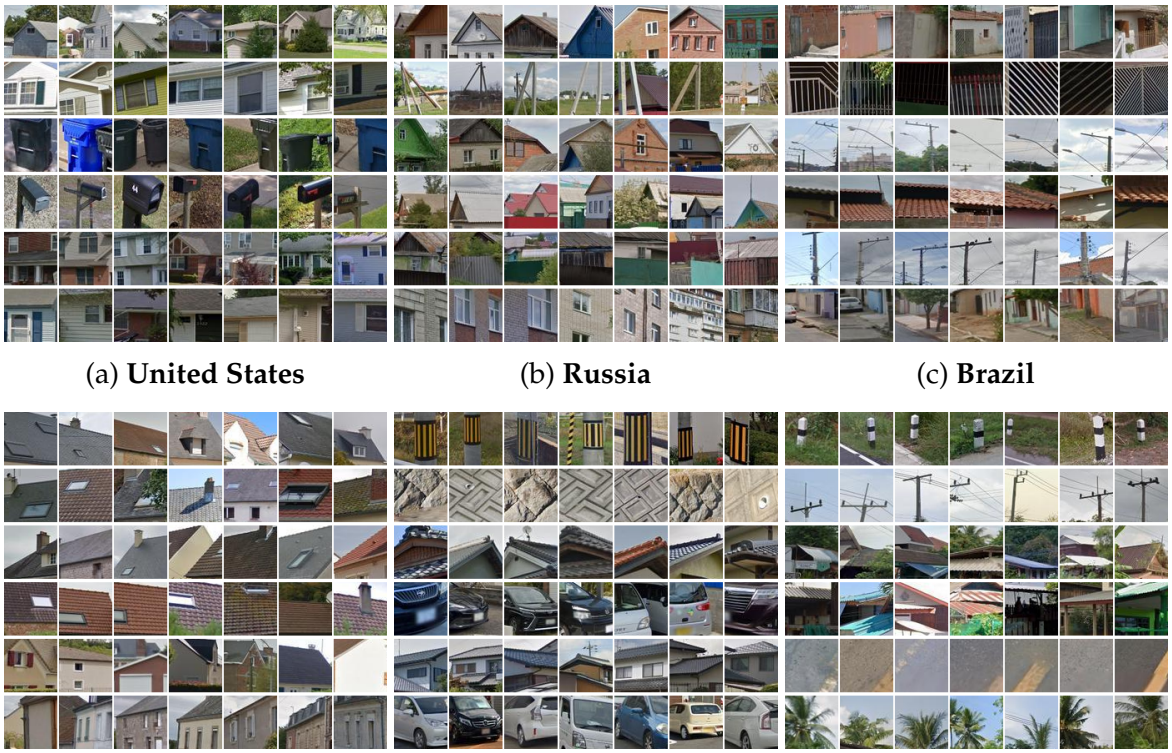
(a) **1920s** (b) **1940s** (c) **1950s**

Figure 4.5: **Clusters of FTT [Chen et al., 2023] visual elements.** Our cluster analysis of faces revealed that eyeglasses of varying designs are indicative of a portrait's decade throughout the history captured by FTT. Observing the 6 most typical clusters for the 1920s **(a)**, the 1940s **(b)**, and the 1950s **(c)**, we see how the shape of glasses is highly informative of each period. We also located fashion items that uniquely trended only in a particular period, such as aviator goggles in the 1920s (2nd row), military caps in the 1940s (1st and 2nd row), and baseball caps in the 1950s (1st row). Consistent with prior analysis [Ginosar et al., 2017], we also found clusters corresponding to smiles and makeup.

this approach when applied directly to the same mining subset of the G^3 dataset used by our approach. Similar to the original paper, we rank the trained detectors by *discriminativeness*, *i.e.*, the percentage of the top-50 final matches inside the positive set [Doersch et al., 2012], and for each we show its top 6 matches. The results produced with [Doersch et al., 2012] method demonstrate more textures, appear much less semantic, and contain much more similar elements than ours. Note that the results in the original [Doersch et al., 2012] paper do not show similar failures, and in particular much less vegetation, simply because the paper used a curated and non-publicly available dataset of images focused on selected cities extracted from Google StreetView.

4.4.4 Limitations

Although our method makes the first step towards utilizing generative models for data mining, it comes with limitations. We visualize our two main failure modes in Fig. 4.9. First, clustering elements using k-means can lead to mixed clusters containing different categories of samples (Fig. 4.9a) or produce repetitively similar clusters. Second, our method identified data artifacts (Fig. 4.9b) that are related to noisy printing or scanning of old photographs or post-processing artifacts of StreetView images, which are highly typical but irrelevant to our purpose. Interestingly, in the case of



(a) **United States** (b) **Russia** (c) **Brazil**

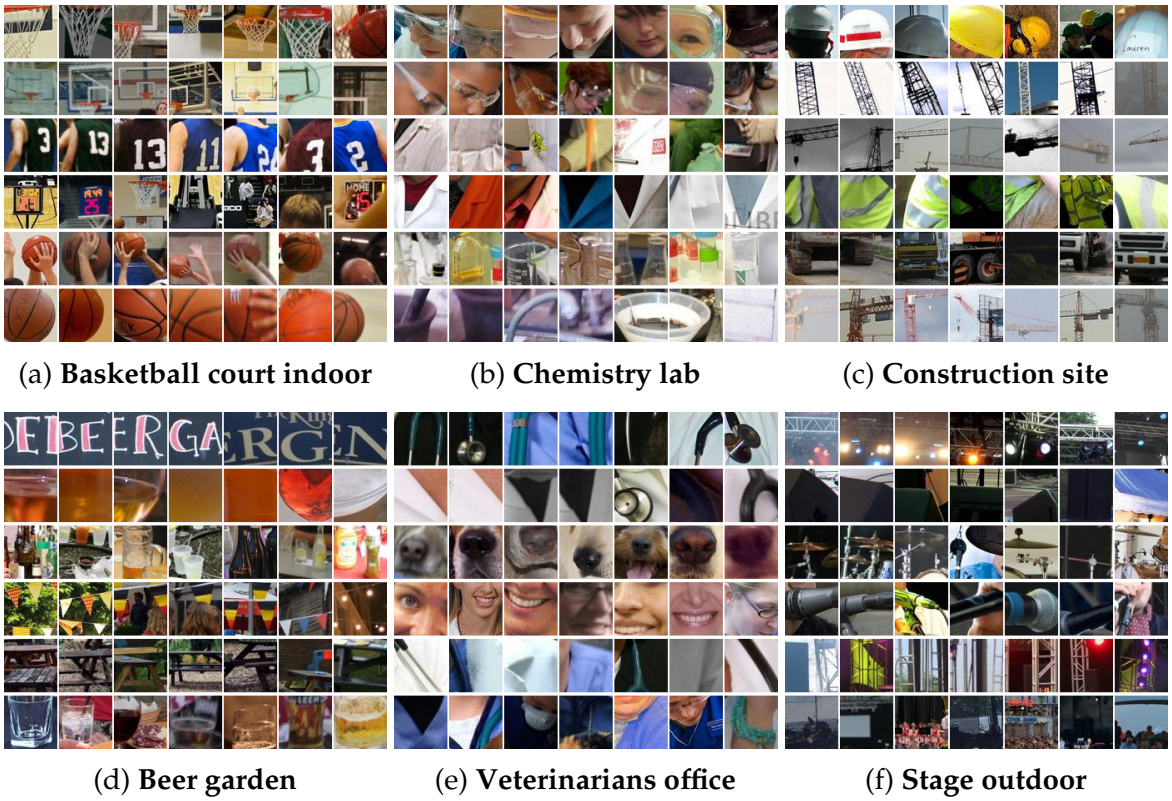
(d) **France** (e) **Japan** (f) **Thailand**

Figure 4.6: **Clusters of G^3 [Luo et al., 2022] visual elements.** Our geographic clusters show a wide diversity of typical elements across different countries. We found architectural elements such as roofs, facades, or windows among the most typical elements in all countries. For example, **(a)** the "double hung" American windows (2nd row), **(d)** French roof windows (1st-4th row), or **(f)** covered pathways in Thailand (4th row). Utility poles are ranked second in Russia and Thailand and 5th in Brazil. We also found typical objects that are unique to a single country, such as **(a)** American garbage cans and post boxes (3rd, 4th row), **(c)** protective guard rails in Brazil (2nd row), **(e)** Japanese electricity warning signs and exterior wall tiles (1st, 2nd row), and **(f)** Thai Bollards (1st row).

StreetView data similar artifacts are suggested in GeoGuessr [geoguessr, 2023] advice websites [geodummy, 2023; geohints, 2023; Plonkit, 2023], as shortcuts for geolocation.

4.5 Applications

Our typicality score allows us to explore three different applications. First, in Sec. 4.5.1, we translate geographical elements across locations and mine typical translations. Then, in Sec. 4.5.2, we show how our method can be used as a qualitative way of understanding bias in the sampled distribution of a diffusion model. Finally, in Sec. 4.5.3,



(a) **Basketball court indoor** (b) **Chemistry lab** (c) **Construction site**

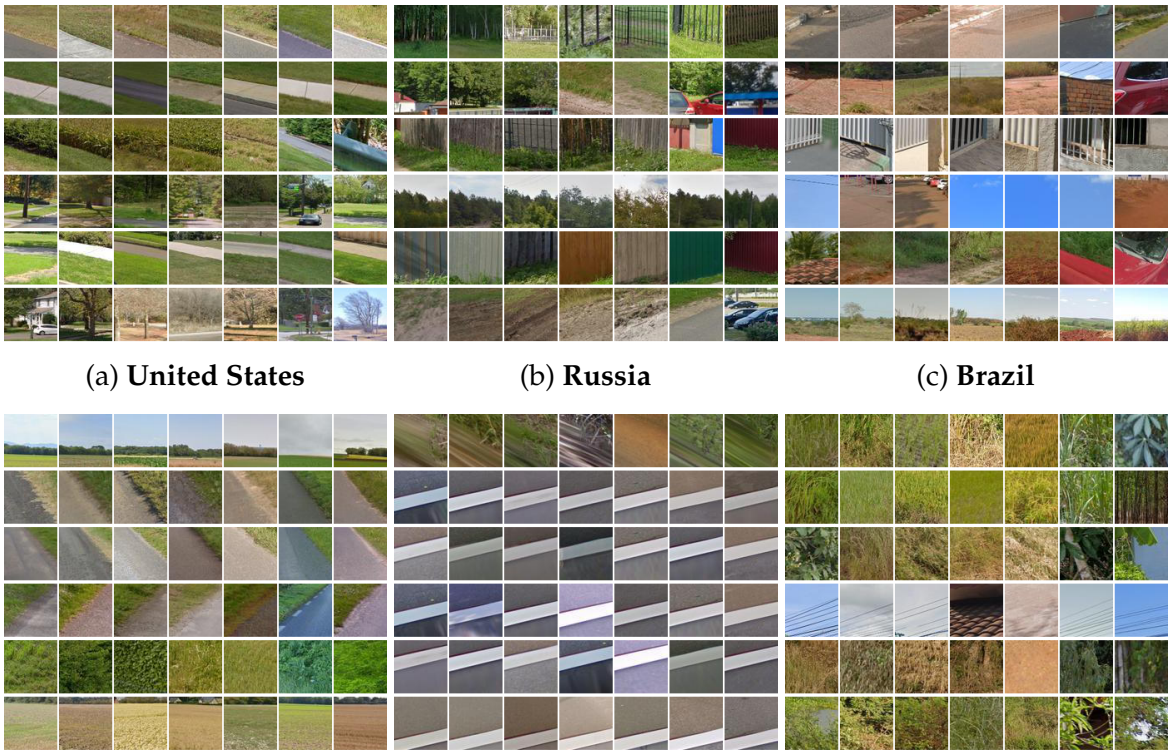
(d) **Beer garden** (e) **Veterinarians office** (f) **Stage outdoor**

Figure 4.7: **Clusters of Places365 [Zhou et al., 2017a] visual elements.** Unlike the other datasets we analyze, each class label correlates with objects of different categories in the scenes dataset, as different scenes contain objects of different categories. Yet, our approach can still summarize a large variety of complex scenes with their unique typical elements. For example, in basketball courts **(a)**, our approach locates the basket (1st row), the backboard (2nd row), the jersey numbers (3rd row), the shot clock (4th row), a shoot (5th row), and the ball (6th row). Our approach can still focus and summarize the most informative elements even in more cluttered scenes like an outdoor stage, chemistry labs, or beer gardens. For example, in the case of "outdoor stage" **(f)**, we see a lot of infrastructural elements, including lights and top rails (1st row), monitor speakers (2nd row), microphones (4th row), and side rails (5th row).

we show how disease localization emerges from typicality when training to generate frontal chest X-rays of patients, of various diseases.

4.5.1 Analyzing Trends of Visual Elements

Having a diffusion model finetuned on a dataset of interest enables further applications that were not possible with previous image mining approaches [Doersch et al., 2012; Lee et al., 2013; Chen et al., 2023; Ginosar et al., 2017]. One new application is the summary of variation of typical visual elements across different classes. As a case study, we use the G^3 dataset to discover and summarize how *co-typical* elements,



(a) **United States** (b) **Russia** (c) **Brazil**

(d) **France** (e) **Japan** (f) **Thailand**

Figure 4.8: **Doersch *et al.*, 2013 [Doersch et al., 2012] results on G^3 [Luo et al., 2022].** See text for details.

such as windows, roofs, or license plates, vary across locations. We start by using our finetuned diffusion model to create a "parallel dataset", by translating all the images in our mining dataset to all location, and then define a measure of co-typicality.

Generating a parallel dataset. We first use Plug and Play [Tumanyan et al., 2023] to translate input images from one location to another, which we denote by $x^{c_0 \to c}$, where c_0 is the initial country and c is the target country. We translate 1000 images for each of the 10 selected countries to all others, resulting in 100K images, which we refer to as our parallel dataset. Performing translation using our finetuned model is critical for keeping scene elements consistent, as seen in Fig. 4.3c. In Appendix B (Sec. E) we show how performing semantic segmentation for each image and its translations to different countries enables measuring statistical trends. For example, we can measure that translations to Thailand or Brazil add many potted plants, and translations to Nigeria add dirt roads and people. This trends can be visually confirmed on our parallel dataset (see Appendix B Sec. E)

(a) Mixed Clusters (b) Mining Dataset Artifacts

Figure 4.9: **Limitations.** The two most common failure modes we observe are: **(a)** issues in clustering, for example, clusters that contain diverse visual content, or multiple clusters that correspond to the same concept; **(b)** typicality highlighting artifacts of the dataset. Discovering artifacts is an expected behavior and can be useful for some applications.

Mining typical transformations across location. To further analyze our parallel dataset, we define a cross-location typicality measure to mine a parallel translation of patches across locations. We define the co-typicality $\bar{\mathbf{T}}$ as the median typicality across location:

$$\bar{\mathbf{T}}(x) = \underset{c \in C}{\text{med}} \left[\mathbf{T}(x^{c_0 \to c}, c) \right], \tag{4.12}$$

where c_0 is the true label of the patch x and the median is computed over all countries in our set of 10 analyzed countries, denoted as C.

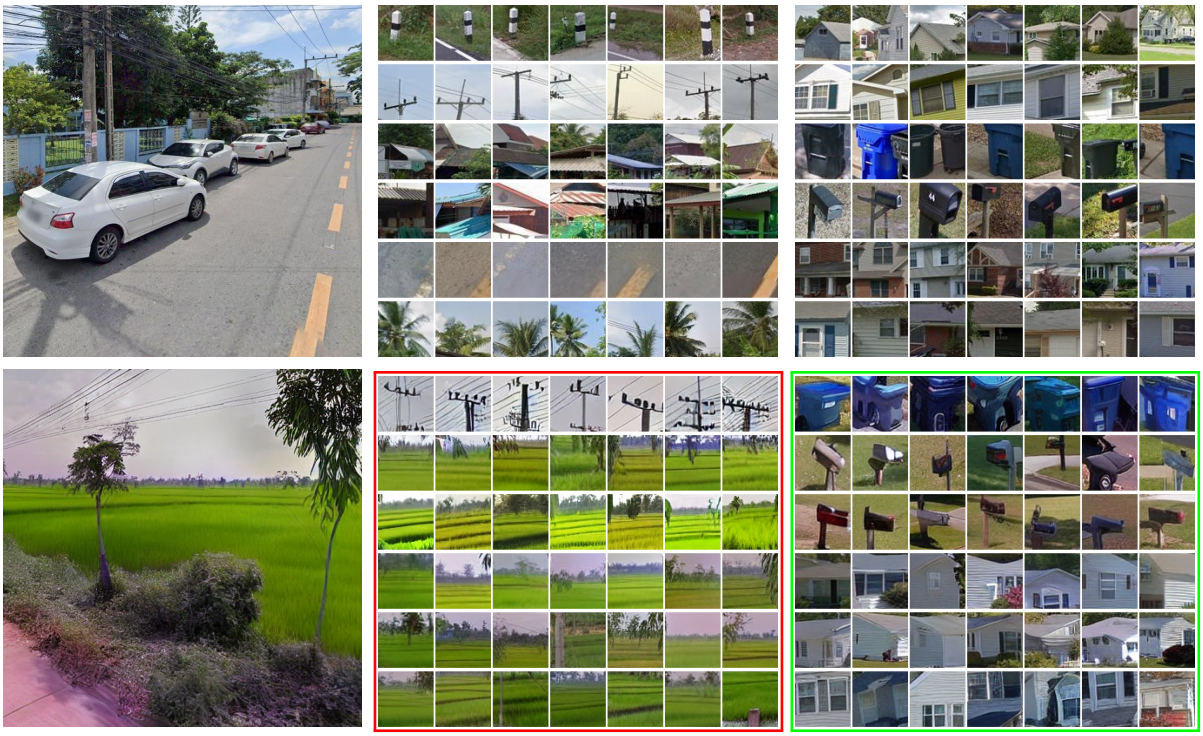
We can now ask: What visual elements are typical of a certain place and whose translation remains typical of another location? Instead of ranking single patches, we now rank a whole sequence of $|C|$ patches translated across locations according to $\bar{\mathbf{T}}$. We represent this sequence by concatenating the DIFT features of each patch [Tang et al., 2023]. To facilitate clustering, we first project the DIFT features of each patch from 1280 dimensions to 32 dimensions using UMAP [McInnes et al., 2018]. To keep the same proportion of typical patches to the number of analyzed images/sequences as in Sec. 4.4.3, we cluster the 10,000 visual elements with the highest co-typicality.

We display our results in Fig. 4.10, where for 6 selected clusters, we show in rows the four translated sequences closest to the cluster mean, highlighting in red the original image in each sequence. On the left column of Fig. 4.10, we show changes in typical architectural elements, such as gables, roofs, and windows. In contrast, on the right we show regulation-related elements, such as road tracks, utility poles, and license plates. Our approach allows us to both locate and visualize how common visual elements would vary from place to place, even though an exact match may not exist in the original data. For example, roofs typically turn dark brown when translated to the UK and black when translated to Japan.

Figure 4.10: **Clustering typical translations of elements across countries.** Ranking translated visual elements according to $\bar{\mathbf{T}}$ and clustering the translated sequences results in groups of elements with similar variations. We show elements from 6 selected clusters out of 32. The source image for each sequence is highlighted in red. See text for details.

4.5.2 Mining Bias in Generation

Our goal in this section is to use mining as a way of interpreting the generative performance of the diffusion model itself. There is a variety of metrics that concern measuring the performance of generative models, including for example Inception Score [Salimans et al., 2016], FID [Heusel et al., 2017], and Precision-Recall [Sajjadi et al., 2018; Kynkäänniemi et al., 2019]. Especially for diffusion models different sampling procedures can reveal different performance of aligning with the training data density for the same model, as demonstrated for two-dimensions in Fig. 1 of [Karras et al., 2025]. This indicates that while the diffusion model may already have knowledge of the true density, sampling may bias it towards certain outputs. In Fig. 4.11 we show how our typicality measure can be used to mine the sampling bias of a diffusion model. We start by sampling synthetic images of an equal amount of 1000 images per class to our real sampling dataset. While there is a clear difference between sampled and real images of Thailand (Fig. 4.11a), it's hard to ground it simply by comparing between



(a) Real vs Synthetic (b) Thailand (c) United States

Figure 4.11: **Mining Sampling Bias. (a)** Example of real image from Thailand (top) and synthetic (bottom). **(b)** Per row the corresponding visual summary for Thailand in the same amount of real and synthetic data. **(c)** The same visual summary but for the United States.

real and synthetic images. Instead, we can compare the mined visual summaries that are extracted using the same model that we used to generate synthetic data (Fig. 4.11b). This can reveal that when sampling Thailand (with c.f.g. of 7.5) the model will place a disproportionately higher amount of rice fields, while the same model is capable of detecting more diverse objects such as bollards, architecture and road-tracks in our real data. This procedure make the bias of the model highly interpretable. Note, that this observation is not common amongst all classes, as can be seen for example in the case of the US (Fig. 4.11c), where the typical visual summaries of real and synthetic images are more similar. Yet, the mined summaries of the real data contain much more distinct elements than the generated ones, which seems related to the diversity-fidelity trade-off of classifier-free-guidance (see Fig. 5 of [Ho and Salimans, 2021]). Examples across all of our ten countries, and different summaries for a range of c.f.g. λ values can be found in Appendix B (Sec. D). Interestingly, this experiment adds another dimension to the concept of *Latent Reading* which I introduced in [Siglidis, 2022], as the practice of understanding cultural data by interacting with a model trained to reproduce them.

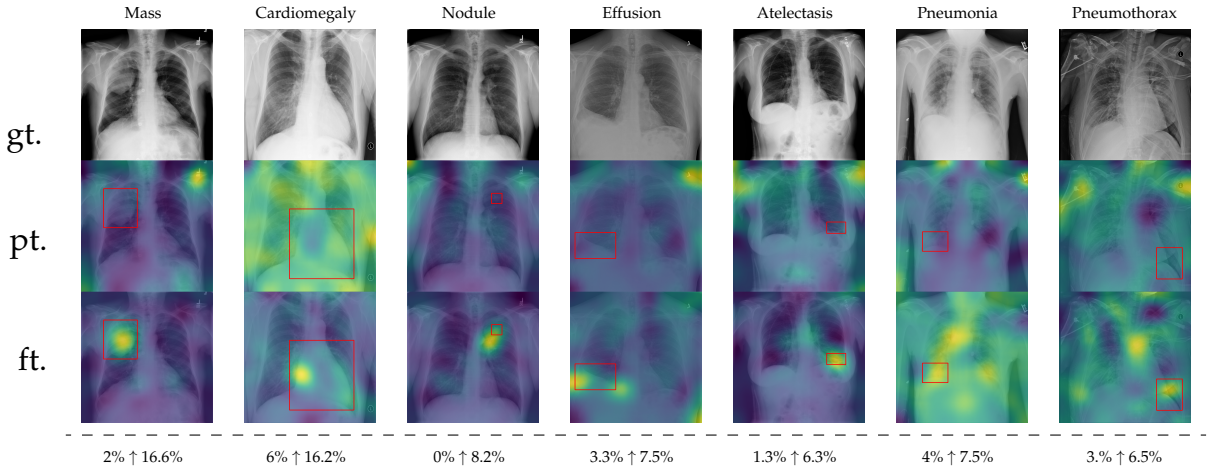


Figure 4.12: **Localizing abnormal areas in medical images.** We visualize typicality when finetuning our model on the CXR8 dataset of thorax diseases [Wang et al., 2017]. After finetuning (ft.), we can see a clear focus of the typicality score on expert annotated areas (red boxes) for each disease, while initial predictions from the pretrained Stable Diffusion V1.5 model (pt.) are mostly noise. Images are ordered by AUC-PR after finetuning [Arun et al., 2021]. With ↑ we delimitate performance before and after finetuning, in the last row.

On one hand, it shows how data is indeed summarized by generative models, which facilitates their analysis. On the other, it shows that the sampling procedure adds bias that doesn't simply truncate non-frequent elements but changes their density. Instead, our method could be used to apply generative models to real data, towards the same goal.

4.5.3 Analysis of Medical Images

In Sec. 4.4.2, we discussed how typicality helps locate relevant patches for an input label. In this section, we test this idea on completely different images: X-rays of patients who may suffer from a combination of various thorax diseases. We finetune Stable Diffusion on the ChestX-ray8 dataset [Wang et al., 2017] containing 108,948 frontal-view X-ray images annotated with 14 single-word disease-name labels. Experts annotated a test set of 879 images with 7 diseases with rectangular regions of interest (ROI) for each disease. For each image, we compute typicality per latent pixel, interpolate the resulting typicality to the input dimension, and blur the resulting typicality map for visualization. In Fig. 4.12, we show the resulting typicality maps together with the ROI annotation before and after finetuning. We can observe that finetuning clearly improves the localization. To quantify this effect on average, we compute the area under the precision recall-curve [Arun et al., 2021] (AUC-PR) of typicality associated

with the annotated ROIs. To do this we binarize typicality across 1000 uniformly (*i.e.*, log-linearly) spaced thresholds and for each count true (typical and inside the ROI) and false (typical and outside the ROI) positives pixels. As reported in Fig. 4.12, we see consistent improvement of this measure when finetuning the network (from 3.2% to 9.6%), ranging from +3.5% for Pneumonothorax (from 3% to 6%) to +14.6% for Mass (from 2% to 16.6%), which are respectively the least and most localized diseases. Similar to Sec. 4.4.3, finetuning uses only image labels without localization supervision.

4.6 Conclusion

This chapter, presented a novel use of diffusion models as visual mining tools. It defined a typicality measure using a pretrained stable diffusion model finetuned for conditional image synthesis. This measure of typicality was used to mine visual summaries of four datasets, tagged by year or location. We then showed that the same typicality measure can be extended in discovering trends when translating visual elements across location, making interpretable summaries of the sampling bias of diffusion models, and even localizing abnormalities in medical data. In summary, this chapter presented a novel approach to image data mining, enabling scaling to datasets significantly more extensive and diverse than those showcased in prior works as demonstrated by our experiments.

Chapter 5

Conclusion

The goal of this section is to first summarize the contributions of this manuscript and then to propose future research directions, for image data mining that are inspired by the work presented in this thesis.

5.1 Summary of Contributions

We have presented two key contributions addressing different types of visual variation, each demonstrating effectiveness in mining visual variation of labeled image collections. Both contributions advance image data mining by combining compositional synthesis with weak supervision. Both discover visual structure in ways that are faithful to the way the trained model represents the input scene. The Learnable Typewriter (Chap. 3), provides explicit synthesis through direct sprite manipulation, while our diffusion-based approach (Chap. 4) leverages implicit synthesis through text conditioning. These two complementary approaches, using both concrete and abstract compositional synthesis methods allow us to mine for two both inter-class and intra-class variation. Our contributions are presented below.

Mining and quantifying morphological variation of characters. We have introduced "The Learnable Typewriter", an interpretable approach that allows capturing and comparing the visual morphology of different character types. Our method, presented in Chap. 3, learns to represent an input collection of text-lines through learnable sprites composed with differentiable transformations. When regularized with an OCR loss it can learn accurate prototypes across a wide range of versatile input documents. Compared to feature-based analysis, our synthesis-based approach allows users to

visually interpret the way that characters are being grouped in respect to the learned prototypes. Using our methodology we can further align and compare them in order to provide a framework for interpretable quantitative analysis of character morphology. We demonstrated the broader applicability of this framework in palaeographic analysis (Sec. 3.5) offering a quantitative validation of established typologies, and an interpretable morphological EDA framework for analyzing historical manuscripts.

Mining typical the structure behind labels. We have presented "Diffusion Models as Data Mining Tools" that provides a novel way to summarize the visual structure that make a label typical inside a training dataset by relying on the synthesis capabilities of diffusion models. Instead of performing pairwise comparisons of input image patches to identify discriminative ones, our proposed approach presented in Chap. 4 leverages diffusion models to introduce a "typicality" score in order to rank and then cluster the most characteristic visual elements. We have demonstrated results across diverse datasets including cars, portraits, geographical data, visual scenes, showing how our approach provides interpretable summaries which remain faithful to the model's perception. Using the diffusion model and our typicality score, we can further create a parallel dataset that allows us to mine visual structure that is typical across different classes, as well as use it to identify the sampling bias of diffusion models, and detect abnormalities in frontal chest X-ray images.

5.2 Future Work

Our contributions to image data mining point to useful research directions which remain unexplored in the current literature. Here, we discuss three key areas for future investigation.

5.2.1 Cross-modal Mining.

Our methods currently rely on annotations in the form of labels or texts related to the input dataset. While such representations are valuable, different modalities could contextualize concepts with higher relevance. For instance, sound information could better describe temporal concepts, while pose information could better represent human behavior. We aim to extend our diffusion approach to mine across any modality for which diffusion can be defined, such as gesture, speech, or text. Ideally, we envision

Figure 5.1: **Data Mining as outlier detection. (a)** A street market in Paris. **(b)** Vendors occluding their license plates with a variety of objects they sell.

an "any-to-any" framework similar to [Bachmann et al., 2024] where users could mine across any modality they select using any other modality as a mining context.

5.2.2 Matryoshka Mining.

Similar to how methodologies for imaging the visual world differ across scales (from telescopy to microscopy), mining approaches may need to be different to properly capture different scales of visual variation. As demonstrated in our work, sprites excel at capturing minute visual facts, while diffusion models are better at grounding conceptual labels onto visual evidence. This raises important questions about developing a unified approach that, like adjusting a microscope's lens, allows for fine-grained categorization. Similar to [Sivic et al., 2008] we would require an approach that learns to contrast a hierarchy of mined visual concepts, that allows us to mine elements at every level of the visual tree. This visual tree could be constructed both in terms of a word hierarchy, for example by contrasting what makes a ragdoll different from a cat, or for exampled be inferred through probabilistic causality [Lopez-Paz et al., 2017].

5.2.3 Data Mining Prior.

The most influential idea behind mining is the discovery of elements that are both frequent and discriminative of their input label [Han et al., 2000; Singh et al., 2012]. In Chap. 4 we showed that a similar set of elements can be computed by finding those which maximize a contrast between the conditional and the unconditional distribution of pixel reconstruction for a given input label. Stated in different terms, this procedure involves finding frequent outliers conditioned to an input context. However, deciding which context to use is really important for the purposes of mining. For example,

when analyzing images of France one may still not want to restrict the mining context in discovering visual structure that are in general, typical to France. For example consider Fig. 5.1. In a historic "marché aux puces" in Paris, vendors would occlude their cars' license plates using scraps from objects they are selling, so that they don't get automatically fined for unnecessary reasons. While their behavior is unique, it is also consistent, which makes it a proper target for a cultural mining application. One approach would be to use a more developed context to differentiate between France and certain areas of France as we discussed above (Sec. 5.2.2). However, similar to our argument about discovering visual structure on Sec. 1.3, inferring their differentiating context is what makes this task challenging. Humans carry a subjective prior of when something is an informative outlier, where the context which grounds it will often be reasoned in retrospect. Figuring out a computational way to learn a data mining prior remains open. One could use a general prior such as the word "category" to filter elements in the scene or use rewards created by users to learn a categorical discovery that is aligned with human preferences. Solving this task can help discover rare consistent outliers across large visual collections, for example species in the context of biodiversity or trends in the context of social media. Ultimately it can provide an automatic way of detecting and evaluating cultural novelty.

5.3 Philosophical Epilogue

In its original goal of summarizing data in a human interpretable way, data-mining reveals a much greater challenge of current AI systems. Even when pretrained on large datasets via generative [Rombach et al., 2022] or self-supervised [Caron et al., 2021] approaches, their integration inside human culture requires datasets with ground truth human annotation to either train or finetune on. Yet, while AI systems can effectively represent and synthesize the knowledge humans have about the world, they are unable to assert their own knowledge on a human level. Various research efforts including data-mining, category discovery (Sec. 2.2.2), mechanistic interpretability (Sec. 2.2.3), and even goal setting in robotics [Ngo et al., 2013] or emergent communication in NLP [Lazaridou and Baroni, 2020], investigate how to improve different components in the automatic production of ground truth. Still, by being partial, these efforts need to be combined in a unified computational approach that closes the cycle between learning and *making* of datasets. Instead of using annotated data to infer annotation on unseen data, research should instead focus in using existing annotation to learn the process of annotating itself (which aligns with a better definition of intelligence [Chollet, 2019]).

Like children becoming adults, or students becoming advisors, AI systems need to graduate from dataset learning to dataset making.

Ethical Statement.

Data Mining can potentially be used for surveillance and war-related applications. As currently technology is not assigned moral responsibilities, I restrict incorporating my work to such processes. In all of my software this use is restricted through a dedicated license. However, as my work is open source and public (as promoted through open-science), I can't exclude the possibility of it being used without my permission towards these ends, in private. I want to clarify that this is neither my responsibility nor my intended use. My work and the scientific work I stand for, is intended in promoting knowledge as means of better orienting and appreciating the world, rather than as a means of control and oppression.

Copyright. For smoothness of presentation I ommitted citations for images used in the introduction. On Fig. 1.1, pictures mainly come from Wikipedia and [Luo et al., 2022; geohints, 2023; Vlachou-Efstathiou et al., 2024]. The font used to depict Textualis in Fig. 1.1b comes from https://www.onlinewebfonts.com/tag/Textualis. On the top row of Fig. 1.2a the images come respectively from left to right from [Vincent, 2007; Marti and Bunke, 2002; Kalleli et al., 2024; Ermengaud, 1400]; on the bottom row from left to right [Johnson et al., 2017; Camps et al., 2022; Luo et al., 2022; Vincent et al., 2024; Kosmyna, 2025; Wang et al., 2017; Cheng et al., 2024].

Bibliography

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. [2023]. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. 29

Adelson, E. H. [2001]. On seeing stuff: the perception of materials by humans and machines. In *Human vision and electronic imaging VI*. SPIE. 17

Aiolli, F., Simi, M., Sona, D., Sperduti, A., Starita, A., and Zaccagnini, G. [1999]. Spi: a system for palaeographic inspections. *AIIA Notizie*. 36, 49

Alba, R., Rubin, G., Boschetti, F., Fischer, F., Clérice, T., and Chagué, A. [2023]. HTRomance, Medieval Italian corpus of ground-truth for Handwritten Text Recognition and Layout Segmentation [dataset]. v1.0.1. 96

Angtian, W., Kortylewski, A., and Yuille, A. [2021]. Nemo: Neural mesh models of contrastive features for robust 3d pose estimation. In *ICLR*. 21

Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., et al. [2021]. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*. 73

Astruc, G., Dufour, N., Siglidis, I., Aronssohn, C., Bouia, N., Fu, S., Loiseau, R., Nguyen, V. N., Raude, C., Vincent, E., Xu, L., Zhou, H., and Landrieu, L. [2024]. Openstreetview-5m: The many roads to global visual geolocation. In *CVPR*. 13

Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., and Fleet, D. J. [2023]. Synthetic data from diffusion models improves imagenet classification. *TMLR*. 58

Bach, F. and Harchaoui, Z. [2007]. DIFFRAC: a discriminative and flexible framework for clustering. In *NeurIPS*. 22

Bachmann, R., Kar, O. F., Mizrahi, D., Garjani, A., Gao, M., Griffiths, D., Hu, J., Dehghan, A., and Zamir, A. [2024]. 4m-21: An any-to-any vision model for tens of tasks and modalities. In *NeurIPS*. 77

Baird, H. S. [1999]. Model-directed document image analysis. In *Proceedings of the Symposium on Document Image Understanding Technology*. 33, 36

Bansal, A., Shrivastava, A., Doersch, C., and Gupta, A. [2015]. Mid-level elements for object detection. *arXiv preprint arXiv:1504.07284*. 23

Baró, A., Chen, J., Fornés, A., and Megyesi, B. [2019]. Towards a Generic Unsupervised Method for Transcription of Encoded Manuscripts. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*. 35, 36, 43

Barriuso, A. and Torralba, A. [2012]. Notes on image annotation. *arXiv preprint arXiv:1210.3448*. 3

Barron, J. T. [2020]. A generalization of otsu's method and minimum error thresholding. In *ECCV*. 89

Barthes, R. [1990]. *The fashion system*. Univ of California Press. 3

Bau, D., Zhu, J.-Y., Strobelt, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., and Torralba, A. [2019]. GAN Dissection: Visualizing and Understanding Generative Adversarial Networks. In *ICLR*. 26

Baudrillard [1999]. Baudrillard interview on his book: "impossible exchange". Accessed: 2024-11-19. 7

Bendale, A. and Boult, T. E. [2016]. Towards open set deep networks. In *CVPR*. 25

Bereska, L. and Gavves, S. [2024]. Mechanistic interpretability for AI safety - a review. *TMLR*. 26

Berg-Kirkpatrick, T., Durrett, G., and Klein, D. [2013]. Unsupervised Transcription of Historical Documents. In *ACL*. 33, 36

Bharadwaj, R., Naseer, M., Khan, S., and Khan, F. S. [2025]. Enhancing novel object detection via cooperative foundational models. In *WACV*. 25

Biederman, I. [1987]. Recognition-by-components: a theory of human image understanding. *Psychological Review*. 3, 10, 20

Blake, A. and Zisserman, A. [1987]. *Visual reconstruction*. MIT press. 3, 7, 17

Bluche, T. and Messina, R. [2017]. Gated convolutional recurrent neural networks for multilingual handwriting recognition. In *ICDAR*. IEEE. 35

Bojanowski, P., Lajugie, R., Bach, F., Laptev, I., Ponce, J., Schmid, C., and Sivic, J. [2014]. Weakly supervised action labeling in videos under ordering constraints. In *ECCV*. Springer. 22

Bolelli, F., Allegretti, S., Baraldi, L., and Grana, C. [2019]. Spaghetti labeling: Directed acyclic graphs for block-based connected components labeling. *TIP*. 90

Bordier, J., Gille Levenson, M., Brisville-Fertin, O., Clérice, T., and Chagué, A. [2023]. HTRomance, Medieval Spain corpus of ground-truth for Handwritten Text Recognition and Layout Segmentation [dataset]. v0.0.6. 96

Bridle, J., Heading, A., and MacKay, D. [1991]. Unsupervised classifiers, mutual information and 'phantom targets. In Moody, J., Hanson, S., and Lippmann, R., editors, *NeurIPS*. Morgan-Kaufmann. 22

Brooks, T., Holynski, A., and Efros, A. A. [2023]. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*. 21, 58

Burgess, C. P., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., and Lerchner, A. [2019]. MONet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*. 34, 36

Camps, J.-B., Vidal-Gorène, C., Stutzmann, D., Vernet, M., and Pinche, A. [2022]. Data diversity in handwritten text recognition: challenge or opportunity? *Digital Humanities*. 12, 33, 35, 41, 42, 50, 81

Cao, L. and Fei-Fei, L. [2007]. Spatially Coherent Latent Topic Model for Concurrent Segmentation and Classification of Objects and Scenes. In *ICCV*. 19

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. [2021]. Emerging properties in self-supervised vision transformers. In *ICCV*. 9, 78

Chai, L., Zhu, J.-Y., Shechtman, E., Isola, P., and Zhang, R. [2021]. Ensembling with deep generative views. In *CVPR*. 58

Chang, H.-S. and Wang, Y.-C. F. [2015]. Optimizing the decomposition for multiple foreground cosegmentation. *CVIU*. 19

Chefer, H., Gur, S., and Wolf, L. [2021]. Transformer interpretability beyond attention visualization. In *CVPR*. 26

Chen, E. M., Sun, J., Khandelwal, A., Lischinski, D., Snavely, N., and Averbuch-Elor, H. [2023]. What's in a decade? transforming faces through time. *Computer Graphics Forum*. 12, 28, 29, 55, 56, 57, 58, 63, 66, 68, 113

Chen, K., Chen, K., Cong, P., Hsu, W. H., and Luo, J. [2015]. Who are the Devils Wearing Prada in New York City? In *ACM*, New York, NY, USA. 24, 29

Chen, K.-T. and Luo, J. [2017]. When Fashion Meets Big Data: Discriminative Mining of Best Selling Clothing Features. In *Proceedings of the 26th International Conference on World Wide Web Companion*. 24

Chen, Z., Dong, S., Yi, K., Li, Y., Ding, M., Torralba, A., Tenenbaum, J. B., and Gan, C. [2022]. Compositional Physical Reasoning of Objects and Events from Videos. In *ICLR*. 20

Cheng, S., Morel, R., Allys, E., Ménard, B., and Mallat, S. [2024]. Scattering spectra models for physics. *PNAS Nexus*. 81

Chernoff, H. [1973]. The use of faces to represent points in k-dimensional space graphically. *Journal of the American statistical Association*. 27

Chiquier, M., Mall, U., and Vondrick, C. [2024]. Evolving Interpretable Visual Classifiers with Large Language Models. In *ECCV*. 29

Chiquier, M., Mall, U., and Vondrick, C. [2025]. Evolving interpretable visual classifiers with large language models. In *ECCV*. Springer. 30

Cho, M., Kwak, S., Schmid, C., and Ponce, J. [2015]. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *CVPR*. 19

Chollet, F. [2019]. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*. 78

Ciula, A. [2005]. Digital palaeography: using the digital representation of medieval script to support palaeographic analysis. *Digital Medievalist*. 36, 49

Ciula, A. [2017]. Digital palaeography: What is digital about it? *Digital Scholarship in the Humanities*, 32(suppl_2):ii89–ii105. 49

Clérice, T., Chagué, A., and Vlachou-Efstathiou, M. [2023]. CREMMA Medii Aevi [dataset]. v0.1.2. 95, 96

Clérice, T. and Pinche, A. [2021]. Choco-Mufin, a tool for controlling characters used in OCR and HTR projects. 95

Coulson, F. T. and Babcock, R. G. [2020]. *The oxford handbook of latin palaeography*. Oxford University Press. 5

Crawford, E. and Pineau, J. [2019]. Spatially invariant unsupervised object detection with convolutional neural networks. *AAAI*. 34, 36

Dalca, A. V., Rakic, M., Guttag, J., and Sabuncu, M. R. [2019]. Learning Conditional Deformable Templates with Convolutional Networks. In *NeurIPS*. 20

Dalens, T., Aubry, M., and Sivic, J. [2019]. Bilinear image translation for temporal analysis of photo collections. *TPAMI*. 56, 58

Dawkins, R. [2000]. *Unweaving the rainbow: Science, delusion and the appetite for wonder*. HMH. 2

Dawkins, R. [2005]. The ancestor's tale a pilgrimage to the dawn of evolution. *The Journal of Clinical Investigation*, pages 299–310. 6

de la Torre, F. and Kanade, T. [2009]. Discriminative cluster analysis. *Theory and Novel Applications of Machine Learning*. 22, 23

de Sousa Neto, A. F., Bezerra, B. L. D., Toselli, A. H., and Lima, E. B. [2020]. Htr-flor: a deep learning system for offline handwritten text recognition. In *SIBGRAPI*. 35

Deng, F., Zhi, Z., Lee, D., and Ahn, S. [2020]. Generative scene graph networks. In *ICLR*. 34, 36

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. [2009]. Imagenet: A large-scale hierarchical image database. In *CVPR*. IEEE. 13, 25

Deprelle, T., Groueix, T., Fisher, M., Kim, V. G., Russell, B. C., and Aubry, M. [2019]. Learning elementary structures for 3D shape generation and matching. In *NeurIPS*. 20

Derolez, A. [2003]. *The palaeography of Gothic manuscript books: From the twelfth to the early sixteenth century*. Cambridge University Press. 5, 12, 49, 51

Dhariwal, P. and Nichol, A. [2021]. Diffusion models beat gans on image synthesis. In *NeurIPS*. 58

Ding, C. and Li, T. [2007]. Adaptive dimension reduction using discriminant analysis and k-means clustering. In *Proceedings of the 24th international conference on Machine learning*, pages 521–528. 23

Doersch, C., Gupta, A., and Efros, A. A. [2013]. Mid-level visual element discovery as discriminative mode seeking. In *NeurIPS*, volume 26. 29

Doersch, C., Singh, S., Gupta, A., Sivic, J., and Efros, A. A. [2012]. What makes paris look like paris? In *SIGGRAPH*. 24, 29, 55, 56, 57, 58, 62, 65, 66, 68, 69, 99, 100, 101, 104

Douglas, R. H. [1985]. Metamagical themas: questing for the essence of mind and pattern. 5

Dravid, A., Gandelsman, Y., Efros, A. A., and Shocher, A. [2023]. Rosetta Neurons: Mining the Common Units in a Model Zoo. In *ICCV*. 26

Dreyer, M., Purelku, E., Vielhaben, J., Samek, W., and Lapuschkin, S. [2024]. Pure: Turning polysemantic neurons into pure features by identifying relevant circuits. In *CVPR*. 26

Dunlap, L., Zhang, Y., Wang, X., Zhong, R., Darrell, T., Steinhardt, J., Gonzalez, J. E., and Yeung-Levy, S. [2024]. Describing Differences in Image Sets with Natural Language. In *CVPR*. 29, 30

Dyson, F. et al. [2004]. A meeting with enrico fermi. *Nature*. 16

Emami, P., He, P., Ranka, S., and Rangarajan, A. [2021]. Efficient iterative amortized inference for learning symmetric and disentangled multi-object representations. In *ICML*. 34, 36

Epstein, D., Jabri, A., Poole, B., Efros, A. A., and Holynski, A. [2023]. Diffusion self-guidance for controllable image generation. In *NeurIPS*. 21

Epstein, D., Park, T., Zhang, R., Shechtman, E., and Efros, A. A. [2022]. BlobGAN: Spatially Disentangled Scene Representations. In *ECCV*. 20, 21

Epstein, D., Poole, B., Mildenhall, B., Efros, A. A., and Holynski, A. [2024]. Disentangled 3d scene generation with layout learning. In *ICML*. 21

Ermengaud, M. [1400]. Breviari d'amor. https://www.bl.uk/manuscripts/FullDisplay.aspx?ref=Harley_MS_4940. Manuscript, Harley MS 4940. 81

Eslami, S. M. A., Heess, N., Weber, T., Tassa, Y., Szepesvari, D., Kavukcuoglu, K., and Hinton, G. E. [2016]. Attend, Infer, Repeat: Fast Scene Understanding with Generative Models. In *NeurIPS*. 34, 36

Faktor, A. and Irani, M. [2012]. "Clustering by Composition" – Unsupervised Discovery of Image Categories. In *ECCV*, Berlin, Heidelberg. Springer. 20

Fan, L., Chen, K., Krishnan, D., Katabi, D., Isola, P., and Tian, Y. [2024]. Scaling laws of synthetic images for model training ... for now. In *CVPR*. 58

Feng, C., Chen, Z., Holynski, A., Efros, A. A., and Owens, A. [2025]. Gps as a control signal for image generation. In *CVPR*. 29, 58

Feng, J., Yang, Y., Xie, Y., Li, Y., Guo, Y., Guo, Y., He, Y., Xiang, L., and Ding, G. [2024]. Debiased novel category discovering and localization. In *AAAI*. 25

Fogel, S., Averbuch-Elor, H., Cohen, S., Mazor, S., and Litman, R. [2020]. ScrabbleGAN: Semi-supervised varying length handwritten text generation. In *CVPR*. 35, 46, 47

Fomenko, V., Elezi, I., Ramanan, D., Leal-Taixé, L., and Osep, A. [2022]. Learning to discover and detect objects. In *NeurIPS*. 25

Fry, H. [2024]. Towards multimodal interpretability: Learning sparse interpretable features in vision transformers. www.lesswrong.com. Accessed: 2025-03-05. 26

Gandelsman, Y., Efros, A. A., and Steinhardt, J. [2024]. Interpreting CLIP's Image Representation via Text-Based Decomposition. In *ICLR*. 29

Garrette, D., Alpert-Abrams, H., Berg-Kirkpatrick, T., and Klein, D. [2015]. Unsupervised code-switching for multilingual historical document transcription. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 36

geodummy [2023]. geodummy. https://geodummy.com/. Accessed: 2023-11-14. 57, 67

geoguessr [2023]. geoguessr. https://www.geoguessr.com/. Accessed: 2023-11-14. 57, 67

geohints [2023]. geohints. https://geohints.com/. Accessed: 2023-11-14. 6, 57, 67, 81

Georgiev, K., Vendrow, J., Salman, H., Park, S. M., and Madry, A. [2023]. The journey, not the destination: How data guides diffusion models. In *Workshop: Challenges in Deployable Generative AI (ICML)*. 27

Gille Levenson, M. [2023]. Towards a general open dataset and model for late medieval Castilian text recognition (HTR/OCR). *Journal of Data Mining and Digital Humanities*. 96

Ginosar, S., Rakelly, K., Sachs, S. M., Yin, B., Lee, C., Krahenbuhl, P., and Efros, A. A. [2017]. A century of portraits: A visual historical record of american high school yearbooks. In *IEEE Transactions on Computational Imaging*. 24, 28, 29, 56, 57, 66, 68

Glaise, A., Clérice, T., Boschetti, F., Fischer, F., and Chagué, A. [2024]. HTRomance, Medieval Latin corpus of ground-truth for Handwritten Text Recognition and Layout Segmentation [dataset]. v0.0.6. 96

Goh, Gabriel and, N. C. a. C. V., Carter, S., Petrov, M., Schubert, L., Radford, A., and Olah, C. [2021]. Multimodal neurons in artificial neural networks. *Distill*. 26

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. [2014]. Generative adversarial nets. In *NeurIPS*. 34

Goyal, K., Dyer, C., Warren, C., G'Sell, M., and Berg-Kirkpatrick, T. [2020]. A probabilistic generative model for typographical analysis of early modern printing. In *ACL*. 36

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. [2006]. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *ICML*. 35, 41

Graves, A. and Schmidhuber, J. [2008]. Offline handwriting recognition with multidimensional recurrent neural networks. In *NeurIPS*. 34, 35

Greff, K., Kaufman, R. L., Kabra, R., Watters, N., Burgess, C., Zoran, D., Matthey, L., Botvinick, M., and Lerchner, A. [2019a]. Multi-object representation learning with iterative variational inference. In *PMLR*. 20

Greff, K., Kaufman, R. L., Kabra, R., Watters, N., Burgess, C., Zoran, D., Matthey, L., Botvinick, M., and Lerchner, A. [2019b]. Multi-object representation learning with iterative variational inference. In *ICML*. 34, 36

Greff, K., Srivastava, R. K., and Schmidhuber, J. [2016]. Binding via Reconstruction Clustering. In *Workshop track, ICLR*. 20

Greff, K., Van Steenkiste, S., and Schmidhuber, J. [2017]. Neural expectation maximization. *NeurIPS*. 20, 34, 36

Gupta, A., Vedaldi, A., and Zisserman, A. [2018]. Learning to read by spelling: Towards unsupervised text recognition. In *Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*. ACM. 35, 36, 42

Haas, L., Alberti, S., and Skreta, M. [2023]. Learning generalized zero-shot learners for open-domain image geolocalization. 100

Han, J., Pei, J., and Yin, Y. [2000]. Mining frequent patterns without candidate generation. *ACM SIGMOD record*. 24, 77

Han, K., Vedaldi, A., and Zisserman, A. [2019]. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*. 25

Hassner, T., Rehbein, M., Stokes, P. A., and Wolf, L. [2015]. Computation and palaeography: potentials and limits. *Kodikologie und Paläographie im digitalen Zeitalter*. 49

He, K., Gkioxari, G., Dollár, P., and Girshick, R. [2017]. Mask r-cnn. In *ICCV*. 7

He, K., Zhang, X., Ren, S., and Sun, J. [2016]. Deep residual learning for image recognition. In *CVPR*. 41

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. [2022]. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*. 58

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. [2017]. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*. 71

Hidayati, S. C., Hua, K.-L., Cheng, W.-H., and Sun, S.-W. [2014]. What are the Fashion Trends in New York? In *ACM*. 24

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. [2017]. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *ICLR*. 21

Ho, J., Jain, A., and Abbeel, P. [2020]. Denoising diffusion probabilistic models. In *NeurIPS*. 11, 58, 59

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., and Salimans, T. [2022]. Cascaded diffusion models for high fidelity image generation. *JMLR*. 58

Ho, J. and Salimans, T. [2021]. Classifier-free diffusion guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Application*. 60, 72, 102

Hochberg, J., Kelly, P., Thomas, T., and Kerns, L. [1997]. Automatic script identification from document images using cluster-based templates. *TPAMI*. 33

Hsu, Y.-C., Lv, Z., and Kira, Z. [2018]. Learning to cluster in order to transfer across domains and tasks. In *ICLR*. 25

Härkönen, E., Aittala, M., Kynkäänniemi, T., Laine, S., Aila, T., and Lehtinen, J. [2022]. Disentangling random and cyclic effects in time-lapse sequences. *ACM Trans. Graph.* 21

Itseez [2015]. Open source computer vision library. https://github.com/itseez/opencv. 90

Jaderberg, M., Simonyan, K., and Zisserman, A. [2015a]. Spatial Transformer Networks. In *NeurIPS*. 40

Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. [2015b]. Spatial transformer networks. In *NeurIPS*. 20, 36

Jae Lee, Y., Efros, A. A., and Hebert, M. [2013]. Style-aware mid-level representation for discovering visual connections in space and time. In *ICCV*. 57

Jahanian, A., Puig, X., Tian, Y., and Isola, P. [2022]. Generative models as a data source for multiview representation learning. In *ICLR*. 58

Jiang, J. and Ahn, S. [2020]. Generative neurosymbolic machines. In *NeurIPS*. 34, 36

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. [2017]. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*. 17, 81

Jones, C., Roulet, V., and Harchaoui, Z. [2022]. Discriminative clustering with representation learning with any ratio of labeled to unlabeled data. *Statistics and Computing*. 22

Joulin, A. [2012]. *Convex optimization for cosegmentation*. These de doctorat, Cachan, Ecole normale supérieure. 19

Joulin, A., Bach, F., and Ponce, J. [2010]. Discriminative clustering for image cosegmentation. In *CVPR*. 18, 19, 23

Joulin, A., Bach, F., and Ponce, J. [2012]. Multi-class cosegmentation. In *CVPR*. 19

Kahle, P., Colutto, S., Hackl, G., and Mühlberger, G. [2017]. Transkribus-a service platform for transcription, recognition and retrieval of historical documents. In *ICDAR*. 34

Kalleli, S., Trigg, S., Albouy, S., Husson, M., and Aubry, M. [2024]. Historical astronomical diagrams decomposition in geometric primitives. In *ICDAR*. Springer. 81

Kamb, M. and Ganguli, S. [2024]. An analytic theory of creativity in convolutional diffusion models. *arXiv preprint arXiv:2412.20292*. 10

Kang, L., Riba, P., Rusiñol, M., Fornés, A., and Villegas, M. [2022]. Pay attention to what you read: Non-recurrent handwritten text-line recognition. *Pattern Recognition*. 35

Kaoua, R., Shen, X., Durr, A., Lazaris, S., Picard, D., and Aubry, M. [2021]. Image collation: Matching illustrations in manuscripts. In *ICDAR*. 57

Karazija, L., Laina, I., and Rupprecht, C. [2021]. ClevrTex: A Texture-Rich Benchmark for Unsupervised Multi-Object Segmentation. In *NeurIPS*. 17, 36

Karras, T., Aittala, M., Aila, T., and Laine, S. [2022]. Elucidating the design space of diffusion-based generative models. In *NeurIPS*. 58

Karras, T., Aittala, M., Kynkäänniemi, T., Lehtinen, J., Aila, T., and Laine, S. [2025]. Guiding a diffusion model with a bad version of itself. In *NeurIPS*. 71

Keim, D. A., Müller, W., and Schumann, H. [2002]. Visual Data Mining. In *Eurographics*. Eurographics Association. 30

Kiapour, M. H., Yamaguchi, K., Berg, A. C., and Berg, T. L. [2014]. Hipster Wars: Discovering Elements of Fashion Styles. In *ECCV*, Cham. 29

Kiessling, B., Tissot, R., Stokes, P., and Ezra, D. S. B. [2019]. escriptorium: an open source platform for historical document analysis. In *ICDAR workshops (ICDARW)*. IEEE. 34

Kingma, D. P. and Welling, M. [2014]. Auto-encoding variational bayes. In *ICLR*. 60

Kirillov, A., He, K., Girshick, R., Rother, C., and Dollár, P. [2019]. Panoptic segmentation. In *CVPR*. 18

Knight, K., Megyesi, B., and Schaefer, C. [2011]. The Copiale Cipher. In *Proceedings of the ACL Workshop on Building and Using Comparable Corpora*. 12, 33, 35, 41, 42, 43, 44, 48, 88

Koh, P. W. and Liang, P. [2017]. Understanding black-box predictions via influence functions. In *ICML*. 27

Kong, X., Liu, O., Li, H., Yogatama, D., and Steeg, G. V. [2024]. Interpretable diffusion via information decomposition. In *ICLR*. 100

Kopec, G. E. and Lomelin, M. [1996]. Document-specific character template estimation. In *Document Recognition III*. 33, 36

Kopec, G. E. and Lomelin, M. [1997]. Supervised template estimation for document image decoding. *TPAMI*. 33, 36

Kopec, G. E., Said, M. R., and Popat, K. [2001]. N-gram language models for document image decoding. In *Document Recognition and Retrieval IX*. 36

Kosiorek, A., Kim, H., Teh, Y. W., and Posner, I. [2018]. Sequential Attend, Infer, Repeat: Generative Modelling of Moving Objects. In *NeurIPS*, volume 31. Curran Associates, Inc. 20

Kosiorek, A. R., Sabour, S., Teh, Y. W., and Hinton, G. E. [2019]. Stacked Capsule Autoencoders. In *NeurIPS*. 21

Kosmyna, N. [2025]. Your brain on chatgpt. 81

Kotchemidova, C. [2005]. Why we say "cheese": Producing the smile in snapshot photography. *Critical Studies in Media Communication*. 57

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. [2019]. Improved precision and recall metric for assessing generative models. In *NeurIPS*. 71

Lazaridou, A. and Baroni, M. [2020]. Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*. 78

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. [1989]. Backpropagation applied to handwritten zip code recognition. *Neural computation*. 34, 35

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. [1998]. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*. 34, 35

Lee, S., Maisonneuve, N., Crandall, D., Efros, A. A., and Sivic, J. [2015a]. Linking Past to Present: Discovering Style in Two Centuries of Architecture. In *2015 IEEE International Conference on Computational Photography (ICCP)*, Houston, TX, USA. IEEE. 25

Lee, S., Maisonneuve, N., Crandall, D. J., Efros, A. A., and Sivic, J. [2015b]. Linking past to present: Discovering style in two centuries of architecture. In *Proceedings of the IEEE International Conference on Computational Photography (ICCP)*. IEEE. 56, 57

Lee, Y. J., Efros, A. A., and Hebert, M. [2013]. Style-aware mid-level representation for discovering visual connections in space and time. In *ICCV*. 12, 24, 25, 29, 55, 56, 57, 63, 65, 68, 112

Leroy, N., Pinche, A., Camps, J.-B., Clérice, T., and Chagué, A. [2023]. HTRomance, Medieval French corpus of ground-truth for Handwritten Text Recognition and Layout Segmentation [dataset]. v0.0.7. 96

Li, A. C., Prabhudesai, M., Duggal, S., Brown, E., and Pathak, D. [2023a]. Your diffusion model is secretly a zero-shot classifier. In *ICCV*. 11, 60, 62, 99

Li, M., Lv, T., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., and Wei, F. [2023b]. Trocr: Transformer-based optical character recognition with pre-trained models. *AAAI*. 34, 35

Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., and Lee, Y. J. [2023c]. GLIGEN: open-set grounded text-to-image generation. In *CVPR*. 58

Li, Y., Liu, L., Shen, C., and van den Hengel, A. [2015]. Mid-level deep pattern mining. In *CVPR*. 25

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramar, J., Dragan, A., Shah, R., and Nanda, N. [2024]. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. ACL. 26

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. [2014]. Microsoft COCO: Common Objects in Context. In *ECCV*. 17

Lloyd, S. [1982]. Least squares quantization in pcm. *IEEE transactions on information theory*. 22, 23, 62

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., and Kipf, T. [2020]. Object-Centric Learning with Slot Attention. In *NeurIPS*. 20

Loiseau, R., Vincent, E., Aubry, M., and Landrieu, L. [2024]. Learnable earth parser: Discovering 3d prototypes in aerial scans. In *CVPR*. 20

Lopez-Paz, D., Nishihara, R., Chintala, S., Scholkopf, B., and Bottou, L. [2017]. Discovering causal signals in images. In *CVPR*. 77

Loshchilov, I. and Hutter, F. [2019]. Decoupled weight decay regularization. In *ICLR*. 41

Luhmann, N. [2013]. Introduction to systems theory. *Polity*. 1

Luo, G., Biamby, G., Darrell, T., Fried, D., and Rohrbach, A. [2022]. G^3: Geolocation via guidebook grounding. *Findings of EMNLP*. 12, 55, 56, 57, 63, 67, 69, 81, 100, 102, 104, 105, 106, 107, 111

Luo, G., Darrell, T., Wang, O., Goldman, D. B., and Holynski, A. [2024]. Readout guidance: Learning control from diffusion features. In *CVPR*. 21

Luo, G., Dunlap, L., Park, D. H., Holynski, A., and Darrell, T. [2023]. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *NeurIPS*. 58

Ma, C., Yang, Y., Ju, C., Zhang, F., Liu, J., Wang, Y., Zhang, Y., and Wang, Y. [2023]. Diffusionseg: Adapting diffusion towards unsupervised object discovery. *arXiv preprint arXiv:2303.09813*. 21

Macy, M. W., Szymanski, B. K., and Hołyst, J. A. [2024]. The ising model celebrates a century of interdisciplinary contributions. *npj Complexity*. 1

Malisiewicz, T. and Efros, A. [2009]. Beyond categories: The visual memex model for reasoning about object relationships. In *NeurIPS*. 4

Marti, U.-V. and Bunke, H. [2002]. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*. 81

Matzen, K., Bala, K., and Snavely, N. [2017]. StreetStyle: Exploring world-wide clothing styles from millions of photos. *arXiv preprint arXiv:1706.01869*. 24, 29, 57

Matzen, K. and Snavely, N. [2015]. BubbLeNet: Foveated Imaging for Visual Discovery. In *ICCV*. 25

McInnes, L., Healy, J., Saul, N., and Grossberger, L. [2018]. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*. 70

Medin, D. L. and Schaffer, M. M. [1978]. Context theory of classification learning. *Psychological review*. 5

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. [2022]. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*. 58

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., and Mullers, K.-R. [1999]. Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop*. IEEE. 23

Mikulinsky, R., Alper, M., Gordin, S., Jiménez, E., Cohen, Y., and Averbuch-Elor, H. [2025]. Protosnap: Prototype alignment for cuneiform signs. In *ICLR*. 35

Mokady, R., Hertz, A., Aberman, K., Pritch, Y., and Cohen-Or, D. [2023]. Null-text inversion for editing real images using guided diffusion models. In *CVPR*. 58

Monnier, T. and Aubry, M. [2020]. docExtractor: An off-the-shelf historical document element extraction. *ICFHR*. 43

Monnier, T., Austin, J., Kanazawa, A., Efros, A. A., and Aubry, M. [2023]. Differentiable Blocks World: Qualitative 3D Decomposition by Rendering Primitives. In *NeurIPS*. 20

Monnier, T., Groueix, T., and Aubry, M. [2020]. Deep Transformation-Invariant Clustering. In *NeurIPS*. 11, 46

Monnier, T., Vincent, E., Ponce, J., and Aubry, M. [2021]. Unsupervised Layered Image Decomposition into Object Prototypes. In *ICCV*. 10, 20, 34, 36, 37, 38, 39, 40, 47, 48, 85, 93, 95

Murphy, G. [2004]. *The big book of concepts*. MIT press. 4, 59

Ngo, H., Luciw, M., Förster, A., and Schmidhuber, J. [2013]. Confidence-based progress-driven self-generated goals for skill acquisition in developmental robots. *Frontiers in psychology*, 4:833. 78

Nichol, A. Q. and Dhariwal, P. [2021]. Improved denoising diffusion probabilistic models. In *ICML*. 60

Nolan, J. C. and Filippini, R. [2010]. Method and apparatus for creating a high-fidelity glyph prototype from low-resolution glyph images. US Patent 7,702,182. 33

Nuhn, M. and Ney, H. [2013]. Decipherment complexity in 1: 1 substitution ciphers. In *ACL*. 91

Oeser, W. [1971]. Das «a» als Grundlage für Schriftvarianten in der gotischen Buchschrift. *Scriptorium*. 53

Ohl, L., Mattei, P.-A., Bouveyron, C., Harchaoui, W., Leclercq, M., Droit, A., and Precioso, F. [2022]. Generalised mutual information for discriminative clustering. In *NeurIPS*, volume 35, pages 3377–3390. 22

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. [2020]. Zoom in: An introduction to circuits. *Distill*. 10, 16, 26

Omiecinski, E. and Ordonez, C. [1998]. Image mining: A new approach for data mining. Technical Report GIT-CC-98-12, College of Computing, Georgia Institute of Technology. 15, 16

Ozguroglu, E., Liu, R., Surś, D., Chen, D., Dave, A., Tokmakov, P., and Vondrick, C. [2024]. pix2gestalt: Amodal segmentation by synthesizing wholes. In *CVPR*. 10, 21

Pach, M., Karthik, S., Bouniot, Q., Belongie, S., and Akata, Z. [2025]. Sparse autoencoders learn monosemantic features in vision-language models. *arXiv preprint arXiv:2504.02821*. 26

Parkes, M. B. [1969]. *English cursive book hands, 1250-1500*. Oxford : Clarendon. 48

Peebles, W., Peebles, J., Zhu, J.-Y., Efros, A. A., and Torralba, A. [2020]. The hessian penalty: A weak prior for unsupervised disentanglement. In *ECCV*. 21, 27

Peleg, S. and Rosenfeld, A. [1979]. Breaking substitution ciphers using a relaxation algorithm. *Communications of the ACM*. 91

Pinche, A. [2023]. Cremma Medieval [dataset]. 96

Pinche, A., Clérice, T., Chagué, A., Camps, J.-B., Vlachou-Efstathiou, M., Levenson, M. G., Brisville-Fertin, O., Boschetti, F., Fischer, F., Gervers, M., et al. [2023]. Catmus-medieval: Consistent approaches to transcribing manuscripts. 95

Plonkit [2023]. plonkit. https://www.plonkit.net/guide. Accessed: 2023-11-14. 57, 67

Pournaki, A., Gaisbauer, F., Banisch, S., and Olbrich, E. [2020]. The twitter explorer: A framework for observing twitter through interactive networks. *Journal of Digital Social Research*. 30, 31

Puigcerver, J. [2017]. Are multidimensional recurrent layers really necessary for handwritten text recognition? In *ICDAR*. 35

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. [2021]. Learning transferable visual models from natural language supervision. In *ICML*. 7, 9, 26, 29, 60, 100, 104

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. [2022]. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*. 58

Rao, S., Mahajan, S., Böhle, M., and Schiele, B. [2024]. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *ECCV*. Springer. 26

Reddy, P., Guerrero, P., and Mitra, N. J. [2022]. Search for concepts: Discovering visual concepts using direct optimization. In *BMVC*. 34

Reeve, C. D. et al. [2004]. *Plato: Republic*. Hackett Publishing. 4

Rematas, K., Fernando, B., Dellaert, F., and Tuytelaars, T. [2015]. Dataset fingerprints: Exploring image collections through data mining. In *CVPR*. 31

Rizve, M. N., Kardan, N., Khan, S., Shahbaz Khan, F., and Shah, M. [2022]. Openldn: Learning to discover novel classes for open-world semi-supervised learning. In *ECCV*. Springer. 25

Roberts, L. G. [1963]. *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology. 17

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. [2022]. High-resolution image synthesis with latent diffusion models. In *CVPR*. 12, 58, 59, 60, 62, 78

Rosch, E. and Lloyd, B. B. [2024]. *Cognition and categorization*. Taylor & Francis. 3, 4

Rosch, E. H. [1973]. Natural categories. *Cognitive psychology*. 4, 5, 30

Rosenblatt, F. [1958]. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*. 15

Rubinstein, M., Joulin, A., Kopf, J., and Liu, C. [2013]. Unsupervised Joint Object Discovery and Segmentation in Internet Images. In *CVPR*, Portland, OR, USA. 19

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. [2023]. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*. 58

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. [2022]. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*. 58

Sajjadi, M. S. M., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. [2018]. Assessing generative models via precision and recall. In *NeurIPS*. 71

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. [2016]. Improved techniques for training gans. In *NeurIPS*. 71

Sapolsky, R. M. [2018]. *Behave: The biology of humans at our best and worst*. Penguin. ix

Sculley, D. and Pasanek, B. M. [2008]. Meaning and mining: the impact of implicit assumptions in data mining for the humanities. *Literary and Linguistic Computing*, 23(4):409–424. 10, 16

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. [2017]. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*. 26

Seuret, M., van der Loop, J., Weichselbaumer, N., Mayr, M., Molnar, J., Hass, T., and Christlein, V. [2023]. Combining ocr models for reading early modern books. In *ICDAR*. 12, 35, 42, 89, 90

Shaham, T. R., Schwettmann, S., Wang, F., Rajaram, A., Hernandez, E., Andreas, J., and Torralba, A. [2024]. A Multimodal Automated Interpretability Agent. In *ICML*. 29

Shen, X., Efros, A. A., and Aubry, M. [2019]. Discovering visual patterns in art collections with spatially-consistent feature learning. In *CVPR*. 18, 57

Shen, X., Efros, A. A., Joulin, A., and Aubry, M. [2021a]. Learning co-segmentation by segment swapping for retrieval and discovery. In *CVPR Image Matching workshop and Transformer workshop, 2022*. 56

Shen, X., Pastrolin, I., Bounou, O., Gidaris, S., Smith, M., Poncet, O., and Aubry, M. [2021b]. Large-scale historical watermark recognition: dataset and a new consistency-based approach. In *ICPR*. 57

Shi, J. and Malik, J. [2000]. Normalized cuts and image segmentation. *TPAMI*. 17, 18

Siglidis, I., Gonthier, N., Gaubil, J., Monnier, T., and Aubry, M. [2024a]. The learnable typewriter: A generative approach to text analysis. In *ICDAR*. Springer. 11, 12, 13

Siglidis, I., Holynski, A., Efros, A. A., Aubry, M., and Ginosar, S. [2024b]. Diffusion models as data mining tools. In *ECCV*. 12, 13

Siglidis, Y. [2022]. Latent reading. In Manouach, I., editor, *Chimeras. Inventory of Synthetic Cognition*, pages 193–195. Onassis Publications. 72

Simard, P. Y., LeCun, Y. A., Denker, J. S., and Victorri, B. [2002]. Transformation invariance in pattern recognition—tangent distance and tangent propagation. In *Neural networks: tricks of the trade*. Springer. 34, 36

Simoff, S., Böhlen, M. H., and Mazeika, A. [2008]. *Visual data mining: theory, techniques and tools for visual analytics*. Springer Science & Business Media. 30

Siméoni, O., Puy, G., Vo, H. V., Roburin, S., Gidaris, S., Bursuc, A., Pérez, P., Marlet, R., and Ponce, J. [2021]. Localizing Objects with Self-Supervised Transformers and no Labels. In *BMVC*. 19

Singh, S., Gupta, A., and Efros, A. A. [2012]. Unsupervised Discovery of Mid-Level Discriminative Patches. In *ECCV*. 6, 9, 23, 24, 25, 56, 77

Sivic, J., Russell, B., Efros, A., Zisserman, A., and Freeman, W. [2005]. Discovering objects and their location in images. In *ICCV*. 18

Sivic, J., Russell, B. C., Zisserman, A., Freeman, W. T., and Efros, A. A. [2008]. Unsupervised discovery of visual object class hierarchies. In *CVPR*. 77

Smirnov, D., Gharbi, M., Fisher, M., Guizilini, V., Efros, A. A., and Solomon, J. [2021]. MarioNette: Self-Supervised Sprite Learning. In *NeurIPS*. 10, 11, 34, 36, 37, 38, 39, 40, 41, 46, 48, 85, 93, 94

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. [2015]. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*. 58, 59

Song, J., Meng, C., and Ermon, S. [2021]. Denoising diffusion implicit models. In *ICLR*. 59, 60

Souibgui, M. A., Fornés, A., Kessentini, Y., and Tudor, C. [2020]. A few-shot learning approach for historical ciphered manuscript recognition. In *CoRR*. 35, 46, 48

Srivatsan, N., Vega, J., Skelton, C., and Berg-Kirkpatrick, T. [2021a]. Neural representation learning for scribal hands of linear b. In *ICDAR 2021 Workshops*. 36

Srivatsan, N., Wu, S., Barron, J., and Berg Kirkpatrick, T. [2021b]. Scalable font reconstruction with dual latent manifolds. In *EMNLP*. 44

Stokes, P. A. [2011]. Describing handwriting, part i-v. Blog Post. Last accessed on 15/03/2024. 48

Studtmann, P. [2024]. Aristotle's Categories. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2024 edition. 4, 30, 49

Stutzmann, D. [2013]. Système graphique et normes sociales : pour une analyse électronique des écritures médiévales. In *Medieval Autograph Manuscripts. Proceedings of the XVIIth Colloquium of the Comité International de Paléographie Latine, held in Ljubljana, 7-10 September 2010*. 49

Stutzmann, D. [2016]. Clustering of medieval scripts through computer image analysis: towards an evaluation protocol. *Digital Medievalist*. 6, 49

Stutzmann, D. [2017a]. Ecmen. 95, 96

Stutzmann, D. [2017b]. Les «manuscrits datés», base de données sur l'écriture. In De Robertis, T. and Giovè Marchioli, N., editors, *Catalogazione, storia della scrittura, storia del libro. I Manoscritti datati d'Italia vent'anni dopo*, pages 155–207. SISMEL - Edizioni del Galluzzo, Firenze. 95

Stutzmann, D. [2018]. Variability as a key factor for understanding medieval scripts: the oriflamms project (anr-12-corp-0010). In Brookes, S., Rehbein, M., and Stokes, P., editors, *Digital Palaeography*, Digital Research in the Arts and Humanities. Routledge. 49

Sun, J. and Ponce, J. [2013]. Learning Discriminative Part Detectors for Image Classification and Cosegmentation. In *ICCV*. 23

Tang, L., Jia, M., Wang, Q., Phoo, C. P., and Hariharan, B. [2023]. Emergent correspondence from image diffusion. In *NeurIPS*. 56, 58, 62, 70, 102

Tian, Y., Fan, L., Isola, P., Chang, H., and Krishnan, D. [2023]. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *NeurIPS*. 58

Troisemaine, C., Lemaire, V., Gosselin, S., Reiffers-Masson, A., Flocon-Cholet, J., and Vaton, S. [2023]. Novel class discovery: an introduction and key concepts. *arXiv preprint arXiv:2302.12028*. 25

Tufte, E. R. and Graves-Morris, P. R. [1983]. *The visual display of quantitative information*. Graphics press Cheshire. 27, 28

Tukey, J. W. [1977]. *Exploratory data analysis*. Addison-Wesley. 2, 30

Tumanyan, N., Geyer, M., Bagon, S., and Dekel, T. [2023]. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*. 58, 69, 102

Uelwer, T., Robine, J., Wagner, S. S., Höftmann, M., Upschulte, E., Konietzny, S., Behrendt, M., and Harmeling, S. [2023]. A survey on self-supervised representation learning. *arXiv preprint arXiv:2308.11455*. 23

Van der Maaten, L. and Hinton, G. [2008]. Visualizing data using t-sne. *JMLR*. 30

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. [2018]. The inaturalist species classification and detection dataset. In *CVPR*. 25

Vapnik, V. [1998]. Statistical learning theory. *John Wiley & Sons*. 15

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. [2017]. Attention is all you need. In *NeurIPS*. 39

Vaze, S., Han, K., Vedaldi, A., and Zisserman, A. [2022]. Generalized category discovery. In *CVPR*, pages 7492–7501. 25

Villa-Vásquez, J.-F. and Pedersoli, M. [2024]. Unsupervised object discovery: A comprehensive survey and unified taxonomy. *arXiv preprint arXiv:2411.00868*. 9

Vincent, E., Saroufim, M., Chemla, J., Ubelmann, Y., Marquis, P., Ponce, J., and Aubry, M. [2024]. Detecting looted archaeological sites from satellite image time series. In *EarthVision Workshop (CVPR)*. 81

Vincent, L. [2007]. Google Book Search: Document Understanding on a Massive Scale. In *ICDAR*. 12, 35, 41, 42, 47, 81, 87

Viola, P. and Jones, M. [2001]. Rapid object detection using a boosted cascade of simple features. In *CVPR*. IEEE. 23

Vlachou-Efstathiou, M., Siglidis, I., Stutzann, D., and Aubry, M. [2024]. An interpretable deep learning approach for morphological script type analysis. In *IWCP*. Springer. 11, 12, 13, 51, 52, 81, 98

Vo, H. V., Bach, F., Cho, M., Han, K., LeCun, Y., Pérez, P., and Ponce, J. [2019]. Unsupervised image matching and object discovery as optimization. In *CVPR*. 19

Vo, H. V., Pérez, P., and Ponce, J. [2020]. Toward unsupervised, multi-object discovery in large-scale image collections. In *ECCV*. 19

Vo, H. V., Sizikova, E., Schmid, C., Pérez, P., and Ponce, J. [2021]. Large-Scale Unsupervised Object Discovery. In *NeurIPS*. 19

Wang, S.-Y., Efros, A. A., Zhu, J.-Y., and Zhang, R. [2023]. Evaluating data attribution for text-to-image models. In *ICCV*. 27

Wang, S.-Y., Hertzmann, A., Efros, A. A., Zhu, J.-Y., and Zhang, R. [2024]. Data attribution for text-to-image models by unlearning synthesized images. In *NeurIPS*. 27

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. [2017]. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*. 73, 81

Wang, Y., Ahsan, U., Li, H., Hagen, M., et al. [2022]. A comprehensive review of modern object segmentation approaches. *Foundations and Trends® in Computer Graphics and Vision*. 9

Wark, M. [2023]. *Raving*. Duke University Press. 4

Wertheimer, M. [1938]. Laws of organization in perceptual forms. In Ellis, W. D., editor, *A source book of Gestalt psychology*. Kegan Paul, Trench, Trubner & Company. 3, 17

Wittgenstein, L. [2009]. *Philosophical investigations*. John Wiley & Sons. 4

Wu, Z., Hu, J., Lu, W., Gilitschenski, I., and Garg, A. [2023]. SlotDiffusion: Object-Centric Generative Modeling with Diffusion Models. In *NeurIPS*. 20

Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., and De Mello, S. [2023]. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*. 58, 102

Xu, L., Neufeld, J., Larson, B., and Schuurmans, D. [2004]. Maximum margin clustering. In *NeurIPS*. 22

Xu, Y. and Nagy, G. [1999]. Prototype extraction and adaptive OCR. *TPAMI*. 33, 36

Yang, Y., Chen, Y., and Soatto, S. [2020]. Learning to manipulate individual objects in an image. In *CVPR*. 34, 36

Yao, S., Hsu, T. M., Zhu, J.-Y., Wu, J., Torralba, A., Freeman, B., and Tenenbaum, J. [2018]. 3d-aware scene manipulation via inverse graphics. *NeurIPS*. 20

Ye, J., Zhao, Z., and Liu, H. [2007a]. Adaptive distance metric learning for clustering. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE. 23

Ye, J., Zhao, Z., and Wu, M. [2007b]. Discriminative k-means for clustering. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *NeurIPS*. Curran Associates, Inc. 22, 23

Ye, V., Li, Z., Tucker, R., Kanazawa, A., and Snavely, N. [2022]. Deformable sprites for unsupervised video decomposition. In *CVPR*. 41

Yuille, A. and Kersten, D. [2006]. Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*. 20

Zeeman, E. C. [1976]. Catastrophe theory. *Scientific American*. 27

Zhang, C., Gupta, A., and Zisserman, A. [2020]. Adaptive Text Recognition through Visual Matching. In *ECCV*. 35, 46, 47

Zhang, L., Rao, A., and Agrawala, M. [2023]. Adding conditional control to text-to-image diffusion models. In *ICCV*. 58

Zhao, Y. [2024]. Alltext nyc. https://www.alltext.nyc/. 7

Zheng, J., Li, W., Hong, J., Petersson, L., and Barnes, N. [2022]. Towards open-set object detection and discovery. In *CVPR*. 25

Zhong, Z., Fini, E., Roy, S., Luo, Z., Ricci, E., and Sebe, N. [2021]. Neighborhood contrastive learning for novel class discovery. In *CVPR*. 25

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. [2015]. Object Detectors Emerge in Deep Scene CNNs. In *ICLR*. 26

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. [2017a]. Places: A 10 million image database for scene recognition. *TPAMI*. 12, 55, 56, 57, 62, 63, 64, 68, 114, 115

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. [2017b]. Scene parsing through ade20k dataset. In *CVPR*. 17

Zhu, J.-Y., Lee, Y. J., and Efros, A. A. [2014]. Averageexplorer: Interactive exploration and alignment of visual data collections. *ACM Transactions on Graphics (TOG)*. 30, 31

Zhu, L., Chen, Y., Freeman, B., and Torralba, A. [2009]. Nonparametric bayesian texture learning and synthesis. *NeurIPS*. 20

Zhu, L., Chen, Y., Torralba, A., Freeman, W., and Yuille, A. [2010]. Part and appearance sharing: Recursive compositional models for multi-view. In *CVPR*. IEEE. 20

Zhu, S.-C., Mumford, D., et al. [2007]. A stochastic grammar of images. *Foundations and Trends® in Computer Graphics and Vision*. 20