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

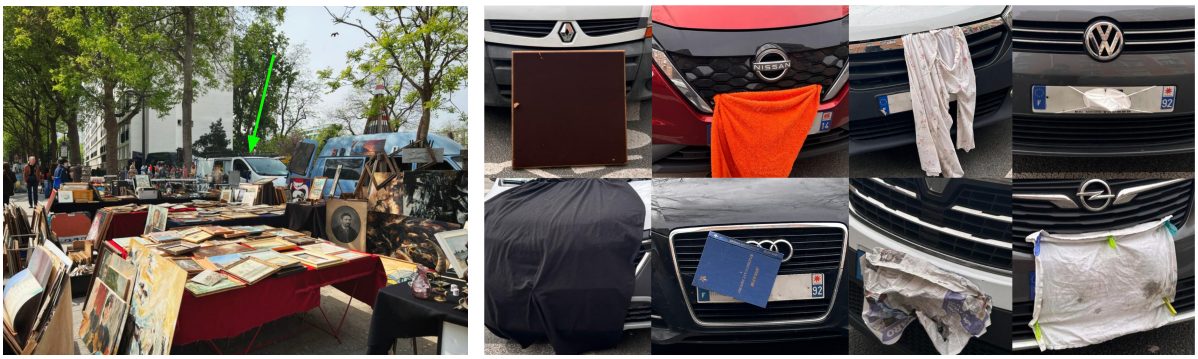

Figure 5.1: **Data Mining as outlier detection. (a)** A street market in Paris. **(b)** Vendors occluding their license plates with a variety of objects they sell.

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].