

DendroMap: Visual Exploration of Large-Scale Image Datasets for Machine Learning with Treemaps

Donald Bertucci, Md Montaser Hamid, Yashwanthi Anand, Anita Ruangrotsakun, Delyar Tabatabai, Melissa Perez, and Minsuk Kahng



Fig. 1. With DendroMap, users can explore large-scale image datasets by overviewing the overall distributions and zooming down into hierarchies of image groups at multiple levels of abstraction. In this example, we visualize images of the CIFAR-100 dataset by hierarchically clustering the image representations obtained from a ResNet50 image classification model. **(B) Treemap View** displays these clusters of images organized as a hierarchical structure by adapting Treemaps. By clicking on a cluster, a user can interactively **(C) Zoom** into that image group, revealing subgroups that replace and fill the available space with animation. The user clicked on a cluster for organism images, which creates distinct subgroups of fish, insects, fruits, and flowers. With **(A) Sidebar View**, the user can dynamically adjust the number of clusters to be displayed and inspect the class-level statistics.

Abstract— In this paper, we present DendroMap, a novel approach to interactively exploring large-scale image datasets for machine learning (ML). ML practitioners often explore image datasets by generating a grid of images or projecting high-dimensional representations of images into 2-D using dimensionality reduction techniques (e.g., t-SNE). However, neither approach effectively scales to large datasets because images are ineffectively organized and interactions are insufficiently supported. To address these challenges, we develop DendroMap by adapting Treemaps, a well-known visualization technique. DendroMap effectively organizes images by extracting hierarchical cluster structures from high-dimensional representations of images. It enables users to make sense of the overall distributions of datasets and interactively zoom into specific areas of interests at multiple levels of abstraction. Our case studies with widely-used image datasets for deep learning demonstrate that users can discover insights about datasets and trained models by examining the diversity of images, identifying underperforming subgroups, and analyzing classification errors. We conducted a user study that evaluates the effectiveness of DendroMap in grouping and searching tasks by comparing it with a gridified version of t-SNE and found that participants preferred DendroMap. DendroMap is available at <https://div-lab.github.io/dendromap/>.

Index Terms—Visualization for machine learning, image data, treemaps, visual analytics, data-centric AI, error analysis

1 INTRODUCTION

The machine learning (ML) community is increasingly aware of the importance of understanding datasets. There is a growing interest in *Data-Centric AI*, as opposed to the model-centric approach [39, 45, 62]. A deep understanding of the datasets can help inform design decisions for building ML models efficiently and appropriately. It motivates important decisions such as collecting more data, changing data labeling policy, debiasing models, and more.

- The authors are with Oregon State University. E-mail: {bertuccd, hamidmd, anandy, ruangroc, tabatase, peremeli, minsuk.kahng}@oregonstate.edu.

This paper has been accepted for the IEEE VIS 2022 Conference and will be published in the IEEE Transactions on Visualization and Computer Graphics.

While images are used extensively in deep learning, fundamental challenges exist in exploring image datasets because they lack attributes like those found in tabular data. A commonly-used approach is to use dimensionality reduction (DR) techniques like t-SNE [57] over multivariate features extracted from the datasets [47]. To enable users to easily see the contents of the images, each data point in the projected space is often replaced with its corresponding image (like in Fig. 2B) [52]. However, for large datasets, there could be overlaps between images, and inefficient use of space (i.e., a lot of white space) makes the size of each image too small for users to inspect.

To scale up DR methods for large image datasets, data points can be re-positioned into a *grid* such that no images overlap and the grid fills the entire screen (Fig. 2C) [23, 30, 46]. It is inspired by a common approach to visualizing image collections—displaying images as a

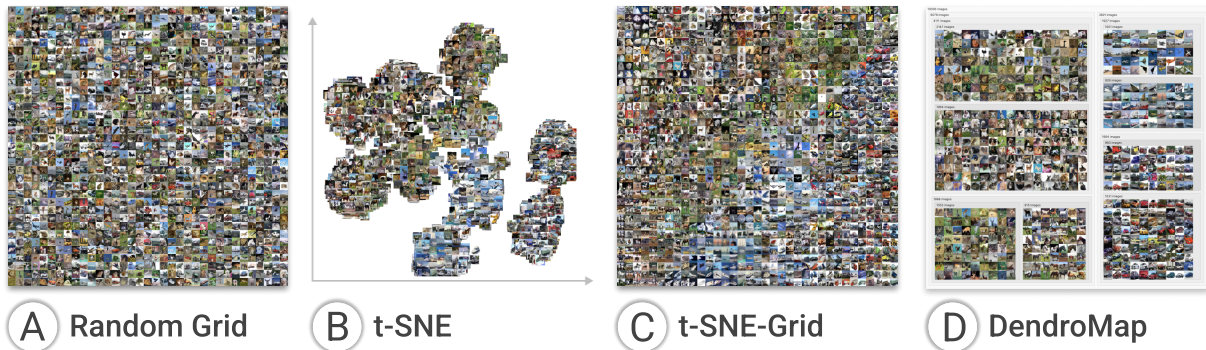


Fig. 2. Commonly used approaches to visualizing image datasets for ML include generating a grid of images (A); projecting the images onto 2-D displays using techniques like t-SNE (B); and using a combination of these two approaches (C). However, they do not scale well to large datasets because images are ineffectively organized and interactions are insufficiently supported. We present DendroMap (D) which effectively organizes image datasets using a modified interactive treemap algorithm.

grid [4, 69]. This is also the case for ML image datasets. Many tutorials for image classification begin by displaying a sample of images as a grid [54]. While the combination of the grid and t-SNE methods effectively use 2-D space, it is still severely limited by the size of the image datasets. Adding interactions may help, but the use of *semantic zooming* is not straightforward for the gridified version of t-SNE. This is because optimization algorithms were applied that distort the original space [28, 46], which means the before-zoom and after-zoom versions may present very different sets of images.

In this paper, we present DendroMap, a novel interactive visualization for exploring large-scale image datasets by adapting treemaps, a well-known visualization technique. DendroMap effectively organizes images using hierarchical clustering algorithms and displays the hierarchical structure with an interactive treemap. A set of image clusters provides an overview of a dataset, and users can interactively zoom into the clusters to investigate sub-clusters in the hierarchy. Fig. 1 illustrates an example. It initially displays eight clusters, each showing a sample of images whose size is proportional to the total number of images in that cluster. Unlike traditional treemaps, the number of image clusters showing can be dynamically changed by the user to customize the level of abstraction. Furthermore, clicking on a cluster will zoom (Fig. 1C) into that cluster to reveal and fill the space with sub-clusters.

DendroMap aims to support a wide range of analytics tasks for ML practitioners. This includes bias and error analysis at the instance and subgroup levels, which have been identified as important in the literature [2, 10, 25]. To build highly accurate and less biased models, it is crucial to have datasets containing a diverse set of images. DendroMap enable users to categorize the types of images present in the datasets and estimate their distributions. Furthermore, DendroMap users can identify underperforming subgroups for error analysis [42, 63, 70].

Contributions. The contributions of this paper are as follows:

- **DendroMap, a novel interactive visualization system for exploring large-scale image datasets used in ML.** DendroMap adapts an interactive zoomable treemap and supports the information seeking mantra “overview first, zoom and filter, then details-on-demand” [51]. Combined with the sidebar as a part of multiple coordinated views, ML practitioners can perform a wide range of tasks for data-centric analysis (e.g., error analysis, bias discovery).
- **An adapted treemap algorithm for hierarchical dendrogram structures of images,** which allows users to dynamically specify the number of clusters to visualize, enabling exploration at multiple levels of granularity. Images are systematically sampled to fill the space for each cluster, providing an overview of the datasets.
- **Live demo on the web**¹ with available code² and use cases for

¹The live demo of DendroMap is available at <https://div-lab.github.io/dendromap/>.

²The code is available at <https://github.com/div-lab/dendromap>.

DendroMap demonstrating users’ dataset exploration, bias discovery, and error analyses.

- **A quantitative user study** designed to compare DendroMap with a gridified version of t-SNE, a space-filling technique used by ML practitioners. Participants performed a wide range of grouping tasks and preferred DendroMap over the baseline.

2 RELATED WORK

2.1 Visualization for Machine Learning

Visualization has helped ML practitioners perform a variety of analytics tasks such as: exploring datasets, analyzing performance results, interpreting and explaining model internals, building models, monitoring training progress, and debugging models [25, 68].

Many existing visualization tools for ML support the tasks of analyzing performance results and exploring datasets at multiple levels of abstraction, ranging from individual instances to entire classes. While ML practitioners often only use summary metrics (e.g., accuracy) or class-level statistics, visualization researchers have argued the importance of instance-level analysis. Early works include ModelTracker [2], Squares [48], and Facets-Dive [60, 61]. These tools represent each instance as a small square using the *unit visualization* technique [43], enabling users to see individual instances in the context of aggregated information. This can work particularly well for image datasets as each square can be replaced with a thumbnail of the actual image content.

While instance-level analysis has benefits, the scale of datasets urges researchers to develop ways to slice and filter datasets, resulting in subgroup-level analysis [21, 25, 29]. This allows users to specify data subsets based on attributes and perform more fine-grained analysis than at the class-level. However, image data creates a fundamental challenge in supporting such analysis because there are no attributes beyond class labels. Therefore, group structures are often created with algorithmic approaches. A common approach is to use a DR technique like t-SNE [57] or UMAP [36], which are often applied to high-dimensional representations obtained from neuron activations [47]. We propose an alternative approach to capturing group structures.

ML researchers have stressed the importance of datasets by coining terms like Data-Centric AI and MLOps [39]. Our work aligns with this trend to ensure that ML datasets are less biased, more fair and inclusive, and contain fewer errors. A recently developed tool named Know Your Data [53] aligns with this goal, providing statistics based on attributes obtained from external APIs (e.g., face recognition, object detection). Our work instead focuses on making sense of raw image datasets by relying on human perception.

2.2 Image Browsing

Zahálka and Worring [69] presented a comprehensive overview of multimedia visualization methods (primarily of images) in their survey. They categorized existing techniques into five types: basic grid, similarity space, similarity-based, spreadsheet, and thread-based. The three

methods commonly used by ML practitioners described in Sect. 1 and Fig. 2 (i.e., random grid, t-SNE, and a grid version of t-SNE) belong to the “basic grid,” “similarity space,” and “similarity-based” categories, respectively. Our proposed treemap-based method can also be placed in the “similarity-based” category.

The idea of using treemaps for image browsing was proposed in PhotoMesa [4]. It consists of two variations of the treemap algorithms: the ordered treemap algorithm ensures the order of images in each treemap block will match the order in file structures (e.g., by timestamp); and the quantum treemap ensures that the widths and heights of the generated rectangles are integer multiples of a given elemental size. Unlike the data commonly used in treemaps, ML datasets have different properties: each dataset has a set of classes, and the images within each class have no order. Because there is no existing hierarchical structure, we extract one using agglomerative clustering algorithms.

An important task in analyzing images or multimedia data is categorizing or exploratory searching. The key difference from tabular datasets is that image datasets are not annotated with structured attributes—images are unstructured. Many common data operations like filtering, grouping, and sorting cannot be easily applied. If we consider low-level tasks by Amar et al. [1], only a few of the 10 tasks can be applied to images [69]. Thus, an important challenge in interactive visualization of image data is automatic extraction of semantic information, interactive exploration of categories, or both [55, 65, 70].

2.3 Similarity-based Visualization Methods

As we discussed in the previous subsection, our proposed work can be considered as a similarity-based approach. We briefly describe both the similarity-space and similarity-based approaches in the ML context.

The t-SNE algorithm is probably the most popular among ML researchers. It is often used to visualize cluster structures learned by deep learning models [11, 47, 57, 58]. While t-SNE often plots each data point as a small circle in a 2-D space, the nature of images provides us with the opportunity to directly plot a small thumbnail instead of a dot. This enables users to see the image contents without interacting with each circle mark (e.g., clicking, hovering). For example, Embedding Projector [52] displays MNIST images in t-SNE plots. However, as the number of images grows, images overlap, making it almost impossible to see them in high-density areas (see Fig. 2B).

Researchers and practitioners have devised methods to address the issue of overlapping images. The images can be rearranged in a grid either by selecting a sample of images among many in each grid or redistributing all images into all the grid spaces in screen using optimization algorithms [28]. Although we have not found research papers to gridify t-SNE or UMAP, there exist several implementations [30, 33, 46], including one by Karpathy [30]. This type of gridifying algorithm has been used in several visual analytics tools for ML for image data [12, 59, 71].

Redistributing data points or images into a rectangular grid has also been studied in non-ML context, such as IsoMatch [17] and rectangular packing [19]. Removing overlaps can be more intelligent by balancing the full use of screen space and intentionally leaving some white-space to reveal cluster structures [22].

2.4 Hierarchical Exploration of Data

To begin our review of hierarchical exploration, we provide a brief background about clustering algorithms [38]. Unlike the k -means algorithm which partitions data points into a fixed number of groups, the hierarchical clustering algorithms iteratively divide data space into smaller space (i.e., divisive) or merge from smaller groups into larger groups (i.e., agglomerative). We use the latter to form a hierarchy (called a *dendrogram*), since divisive does not produce high-quality results for high-dimensional data and is computationally expensive for large data. The agglomerative ones align more closely with useful characteristics of t-SNE: focusing on similar pairs to find cluster structures.

Existing work on visualizing dendrograms include Hierarchical Clustering Explorer (HCE) [49], Stacked Trees which interactively merge parts of the dendrogram [6], and Yang et al. for steering and revising the dendrograms [66]. All these used node-link diagrams to display

dendrograms; however, they are less desirable for image datasets, because the dendrograms require all instances to be positioned along a single line, which means the size of images would become very small if we want to display images in place of the leaves of the dendrogram tree. A space-filling technique like treemaps can resolve this issue.

Hierarchical data exploration has been studied extensively in text domains. Text data is unstructured, so automatic extraction of clusters is important too like images. HierarchicalTopics [15] extracts hierarchical structures of latent topics and enables users to explore and revise them. TopicLens [32] allows users to zoom into certain areas of projected two-dimensional spaces. Marcilio et al. extracts hierarchical structures from high-dimensional representations of deep learning data [35]. Nmap represents data as treemap-style representations, similar to ours [16]. It adjusts initial positions of data items obtained from 2-D projection algorithms by iteratively creating treemap nodes using their modified slicing algorithm. We instead create tree structures using well-known clustering algorithms. Another difference is that our work targeting image data displays image thumbnails within treemap nodes.

3 DESIGN GOALS

To help ML practitioners explore large-scale image datasets, we adapt treemaps with the following design goals:

1. **Overview of Data Distributions.** We aim to assist users in getting an overview of datasets as a beginning step for their analysis of datasets. This includes helping them answer questions like what kinds of images mostly exist in their datasets, whether they are *diverse* enough [27] or biased towards any properties [9].
2. **Exploring at Multiple Levels of Abstraction.** We aim to design our visualization to provide users with abilities to interactively adjust the level of abstraction. While treemaps are effective at supporting *abstract and elaborate* interactions [67], we adapt the original treemap techniques by considering unique properties of the dendrogram structure and the domain of ML for images.
3. **Instance-level Exploration.** As images do not contain attributes, it is important for users to see the individual image contents while exploring datasets. We aim to effectively organize image thumbnails to help users find and inspect individual data points while they navigate over the tree structure.
4. **Subgroup-level Analysis for ML.** Both the literature in multimedia analytics and visual analytics for ML point out the importance of identifying subgroups from datasets [25, 42, 69]. This can be useful for performing a wide range of analytic tasks in ML, such as error analysis and bias discovery [10, 63].

4 DENDROMAP CONSTRUCTION AND INTERACTIONS

This section describes how a dendrogram can be constructed from an image dataset, how DendroMap visualizes the dendrogram, and how supported interactions help achieve our design goals.

4.1 Dendrogram Tree Construction

To create groups of images for hierarchical exploration, we use the well-known hierarchical agglomerative clustering algorithm [38].

The clustering algorithm takes as input high-dimensional representations of images. There are several ways to obtain such representations, such as by extracting high-dimensional embeddings from pre-trained or fine-tuned models, low-dimensional encodings using Autoencoders, or raw image pixels [24, 25, 29, 47]. In our user study in Sect. 6, for the CIFAR-10 dataset, we extracted 1024 dimensional embedding vector representations from the second-to-last hidden (fully-connected) layer in pretrained ResNet50 models that we fine-tuned on CIFAR-10.

Given this input, each image vector is initialized as its own cluster to start, then the most similar image clusters are merged together using Ward linkage with the Euclidean distance metric to form more balanced trees [38]. The agglomerative merging process repeats until the final two clusters merge into one cluster containing all the images in the dataset. The output of the algorithm forms a special tree structure, called *dendrogram*, with leaf nodes corresponding to data instances.

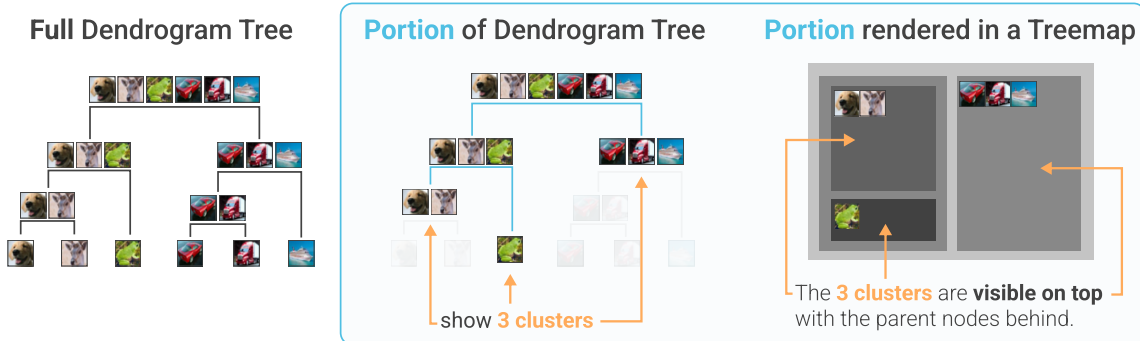


Fig. 3. To scalably visualize the dendrogram tree structure created from agglomerative clustering methods, users can dynamically specify the number of clusters to be rendered in DendroMap. In this example, a portion of the dendrogram is rendered in the treemap view to show three image clusters. Increasing the number of clusters to be shown will result in creating more partitions across the treemap with smooth animations.

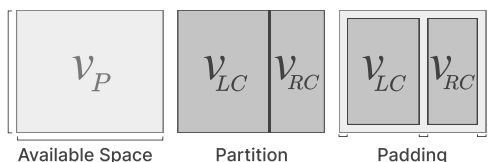


Fig. 4. The slice-dice layout takes the available space given by the parent node v_P and partitions the space into for its two children v_{LC} and v_{RC} . To reveal the v_P 's hierarchy, padding is added to the children boxes.

4.2 DendroMap Visualization

DendroMap visualizes dendrogram structures using a modified treemap algorithm. It traverses the dendrogram and renders each cluster node as a grid of images using the available rectangular space.

Treemap Layout. The dendrogram resembles a binary tree, and all non-leaf nodes have only two child nodes. This allows DendroMap to adapt the traditional slice-dice treemap layout [50]. Normally, slice-dice creates undesirable aspect ratios when laying out many rectangles per level [5]; however, this issue does not occur in ours because the dendrogram will not have more than two children per node, always resulting in just one partition of space.

We modify the slice-dice layout to display a grid of fixed sized images on top and to include padding (to highlight hierarchical structures). To demonstrate one iteration of the modified layout, consider a node v_P that has two children v_{LC} and v_{RC} with 6 and 4 images, respectively. The goal is to fill a 100 by 90 pixel available space depicted in Fig. 4. The algorithm works as follows:

1. **Dice if the available space from the parent v_P is a horizontal rectangle and slice if it is vertical.** In Fig. 4, v_P 's width w_P is 100 pixels and height h_P is 90 pixels, so dicing is chosen.
2. **Compute the ratio to partition the space.** When dicing, the partition ratio is calculated by $ratio := N_{LC}/N_P$, where N_i represents the number of images in v_i . The left and right areas of the partition correspond to each child, v_{LC} and v_{RC} . In Fig. 4, the dice partition ratio is computed as $(6/10) = 0.6$. Meaning 60% of the space is for the v_{LC} and 40% is for v_{RC} .
3. **Adjust the partition to fit images.** Based on the image size, compute the maximum amount of the images that can fit across entire parent's width (or height if slicing) by $fit := \lfloor w_P/w_{image} \rfloor$, where w_P is the width of the available space for v_P and w_{image} is the width of each image. Then the actual partition dimensions can be calculated as $\lfloor fit \times ratio \rfloor$ pixels, resulting in a partition that fits images without cutting them off.
4. **Add padding to show hierarchies.** After laying out the v_{LC} and v_{RC} and assigning them their new dimensions, a fixed padding is added to reveal the parent cluster v_P behind it (like in Fig. 4). We set a fixed padding of 10 pixels in our implementation. Color can encode the remaining height of tree under that node [7].

Adjusting the number of clusters. Traversing the *entire* dendrogram quickly fills the available screen space, making it hard to display

many images. Thanks to the dendrogram's binary tree structure, each iteration of the DendroMap algorithm only lays out two children (one partition), which allows us to render specific number of clusters (i.e., k set by users). By traversing the tree breadth-first and counting the k clusters created so far, the algorithm can stop and show those k clusters. For example in Fig. 3 the dendrogram traversal stops to only render three clusters showing in the treemap.

Organizing images within the clusters. A useful property of dendrograms is that the leaf nodes (i.e., images) are positioned along a line based on the structure of the constructed tree in a way that there is no edge crossing. We use this positional information to organize the list of images for each cluster node. As seen in the Fig. 3 dendrogram, the similar images merge together starting from the bottom, and at each successive merge, they still maintain the position of the leaves in the dendrogram from left to right. The end result is the root node cluster's images are in the same position as the leaf nodes in the dendrogram, which lets similar looking images clump up together and nearby images in a cluster be likely more similar than images located far within the cluster. For example, in Fig. 1 on the right, insect images taken over white background are clustered together with a large node. When there exist a larger number of images to display than the amount of available space, we systematically sample images from the cluster. Specifically, we compute the period by calculating the total number of images in the cluster over the maximum number of images we can possibly show and round down to the nearest integer. We then sample images in the cluster with that period of frequency. For example, if 30 images can be shown in a given space and the cluster has 150 images, we compute the period to sample by $\lfloor 150/30 \rfloor = 5$ and iterate over the cluster $\{x_1, \dots, x_{150}\}$. The end result is an image every 5 iteration to determine 30 images $\{x_1, x_6, x_{11}, \dots, x_{146}\}$. This enables us to show representative samples of a cluster and avoid hiding images that occur later on. We display the total number of images at the top of each cluster node, as well as their classification accuracy if available.

Zooming interactions. For further navigation of the clusters, DendroMap supports zooming. When a cluster node is clicked, DendroMap animates to zoom into the new cluster, which enlarges the selected cluster to fit into the entire space, and creates a set of subclusters within the selected cluster. Our implementation basically follows Bostock's zoomable treemap implementation [8]. In addition, by taking up the entire space with the zoom-in, more images can be shown with more specific hierarchies, leading to more in-depth exploration. This process corresponds to rendering a downstream portion of the dendrogram. At any point, by clicking back on the parent cluster, the reverse process of zooming-out goes back up the tree to reveal the top-level view again. These zooming interactions allow users to quickly explore large image collections at multiple levels of granularity.

4.3 Coordinated Views with the Sidebar

We developed a system for DendroMap by designing coordinated views consisting of the main treemap view and the sidebar. The sidebar contains rendering settings for the treemap display, a class table for

Class Name	Count (Actual)	Count (Predicted)	Accuracy	False Negative Rate	False Positive Rate
boy	91	105	58%	42%	50%
girl	90	71	46%	54%	42%
man	86	98	64%	36%	44%
woman	83	83	63%	37%	37%
baby	72	81	65%	35%	42%

Fig. 5. The class table summarizes class-level statistics of images present in the selected cluster in the treemap view. The user can sort and search for classes, and hover over each entry to quickly locate accurate or error filled clusters highlighted directly on the DendroMap.

class-level error analysis, and a panel for details for a selected image.

DendroMap Settings. The sidebar contains two sliders to change the overview level: one controls the number of clusters visible and the other controls the image size. By default, DendroMap shows eight clusters of medium-sized images to balance the level of detail and overview such that many images can be shown while still separated into distinguishable groups. These sliders allow users to easily change the overview level based on their exploration needs.

When a dataset comes with predictions from a trained model, the sidebar provides two options to highlight misclassified images. One toggle highlights these images using a red border and the other toggle puts the images into focus by making the others translucent. Visually emphasizing misclassified images makes it easier for users to find groups of images that the model consistently misclassifies.

Class Table. The class table is visible if model predictions are present. The table contains information for additional *error analysis* at the class level. The table updates based on the parent cluster’s images (i.e., the root or previously selected cluster; by default, all images). Each row of the table corresponds to a specific class in the dataset (e.g., cat). The next two columns of the table displays the counts of images with a true or predicted class label matching the class specified.

The last three columns of the table provide useful metrics for class-level error analysis: the prediction accuracy (i.e., how often the true and predicted classes matched that row’s class), the false negative rate (i.e., how often the true class matched that row’s class but the predicted class was different), and the false positive rate (i.e., how often the predicted class matched that row’s class but the true class was different). As shown in Fig. 5, each rate is encoded with the opacity of a colored dot.

By hovering over one of these entries in the table, the treemap view highlights the images used to determine that metric by making the other images translucent. This way users can use the class table in tandem with the treemap to isolate and find areas of high error or high accuracy.

Image Details. A user can click on an image in DendroMap to see detailed information: larger view of the image, true class label, predicted class label if it has one, and similar images. The similar images are determined based on distances in the high-dimensional space, which can be used for counterfactual analysis [13, 18].

4.4 Implementation Details

The DendroMap system was built using *Svelte*³, a reactive JavaScript framework that has been increasingly used in the visualization community. The main component, the treemap view, is implemented primarily with *D3.js*⁴ to create SVG elements and to transition the elements for natural animation. The complementary component, the sidebar, is entirely implemented in *Svelte*, and uses *Svelte store* functionality to communicate between the treemap. The dendrogram structure is created from the SciPy⁵ hierarchical clustering implementation with Ward linkage (recommended as default). The output dendrogram is exported as a nested JSON object to be rendered as a treemap on the client side.

³Svelte JavaScript Framework: <https://svelte.dev/>

⁴D3 JavaScript Library: <https://d3js.org/>

⁵SciPy Python Library: <https://scipy.org>

5 USE CASES

In this section, we describe how DendroMap can be used in practice to explore and analyze image datasets through three usage scenarios.

5.1 Examining Bias in Datasets

Consider Priya, a data scientist who lives in the Southeast region of Asia and is evaluating whether ImageNet can be used to train an image classification model that she can deploy in her country. After she loads the DendroMap interface, Priya begins to click around to “zoom” into different portions of the dataset. She first clicks on the rectangle containing the approximately half of the dataset and discovers a cluster containing everyday objects. She notices a cluster of taxi cabs and hovers over the class name “taxicab” in the sidebar’s class table to put just the taxicab photos in focus while the rest become faded. She notices that most are black or yellow, but she knows from personal experience that many taxis are multicolored in her country, so she makes a note to supplement the “taxicab” class with some of those images. Priya “zooms out” by clicking on the outermost rectangle and decides to visit another cluster, this one featuring many images of people interacting with a variety of everyday objects, such as “violin” and “sunscreen”. However, as she clicks on several images to get a better look at each one, she notices that the images tend to include people with lighter skin tones. She makes another note to supplement the dataset with images of people with darker skin tones interacting with the objects corresponding to each of the classes listed in the class table. Given these notes, Priya now would like to train models using this dataset and evaluate them by a set of slices which she made notes (e.g., skin tone), to make sure the models perform consistently over these slices.

5.2 Identifying Underperforming Subgroups

Consider Dave, a ML engineer who is using the CIFAR-100 dataset to evaluate a trained image classification model. He opens DendroMap and sees the default view of eight rectangles or clusters. As Dave inspects the interface, he notices that the group of images with the lowest accuracy score (57 percent) consists mostly of human faces. He sees no obvious pattern at this level of overview in the hierarchical structure, so he clicks on another rectangle to get a closer look. From the class table in the sidebar, he observes that a majority of the images in this group were predicted to be “woman” or “girl”, but most were incorrect. Dave thinks perhaps his classification model has trouble determining which of those two labels is correct. He navigates back up one level by clicking on the outermost rectangle. He selects a different cluster and this time he observes that a majority of the images are predicted as “man” or “boy”, but with similar proportions of incorrect guesses (as shown in Fig. 6). From these two insights, Dave hypothesizes that his model can distinguish male and female faces, but has difficulty determining whether the person is a child or adult. Then he decides to collect additional training data of human faces for four different categories: adult female, adult male, boys, and girls.

5.3 Analyzing Classification Errors

Consider Anna, a ML practitioner who works in a team developing computer vision applications. While she trained a model, she noticed her model consistently had a harder time correctly predicting images from the artifact-related classes so she decided to analyze her model for these classes from the ImageNet dataset, such as “umbrella” and “frying pan”. She opens DendroMap and toggles the “outline misclassified” and “focus misclassified” switches to spotlight the misclassified images, outlined in red, while the others fade. She notices that the red outlined images appear to be scattered without much of a pattern, so she gradually increases the number of clusters until DendroMap splits the images into subgroups of higher or lower accuracy. She stops when it reaches 18 clusters because she notices distinct subgroups of images with high accuracy (over 90 percent). Most of these subgroups focus on particular classes, such as “racket” or “potter’s wheel”. Anna wants to investigate the cause of clusters with much lower prediction accuracy, so she continually clicks on the next visible cluster with the lowest accuracy. She notices a pattern as she keeps drilling down towards the

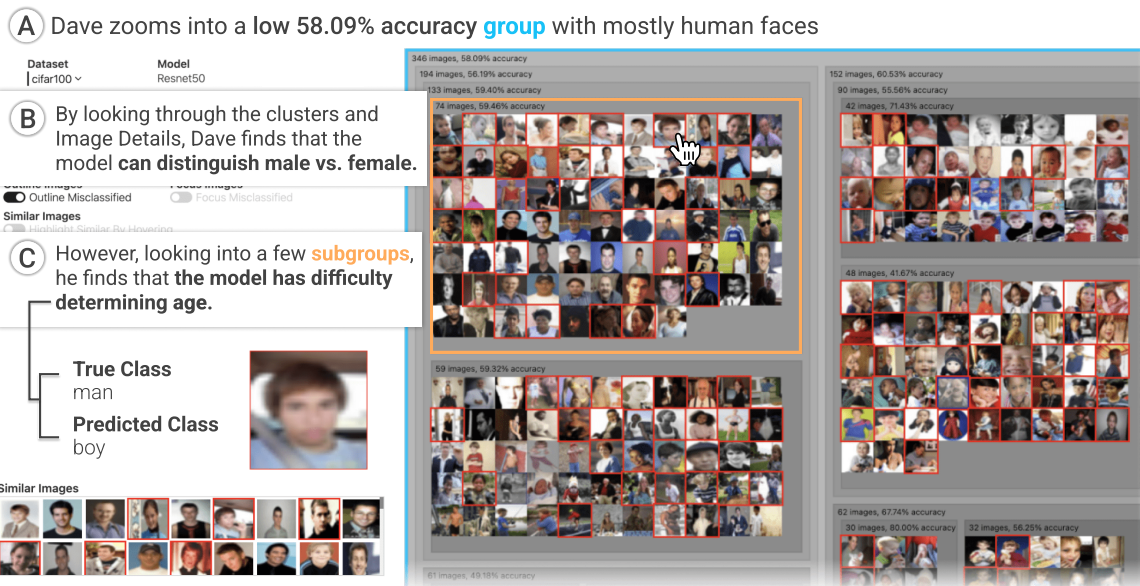


Fig. 6. In our case study, ML practitioner, Dave, investigates the specific classes that his model struggles with using DendroMap.

leaf nodes: the accuracy rate decreases as the images become more cluttered. She clicks on several misclassified images to inspect their true and predicted class labels, and she discovers that the predicted labels are not necessarily inaccurate—it is that the true label and predicted labels are classifying the entire image based on only a portion of it. For example, she clicks on an image of a couple of people sitting on a bench on a sunny day. The true class label for this image is “sunglasses” because one person is wearing sunglasses, whereas the predicted label for the image is “park bench” because the two people are sitting on a bench. These errors can be critical for her team’s applications, so Anna decides to consider object detection models which can locate multiple objects within a single image, instead of image classification models.

6 USER STUDY

To evaluate the effectiveness of DendroMap for exploring large-scale machine learning datasets, we conducted a user study comparing DendroMap and a baseline visualization technique for images, t-SNE-Grid, a gridified version of t-SNE.

6.1 Baseline: t-SNE-Grid

We compare DendroMap with a gridified version of t-SNE, which we call t-SNE-Grid. It re-adjusts the positions obtained from the t-SNE algorithm [57], by filling the available rectangular grid space with the images for effectively using screen space [30].

This process works by first taking the image representations from the dataset and reducing them down to their two-dimensional embeddings using t-SNE (like Fig. 7A). Then, to fill the space, two dimensional grid points are evenly laid out over the space of image embeddings (like Fig. 7B). Finally, each grid point is assigned the closest image

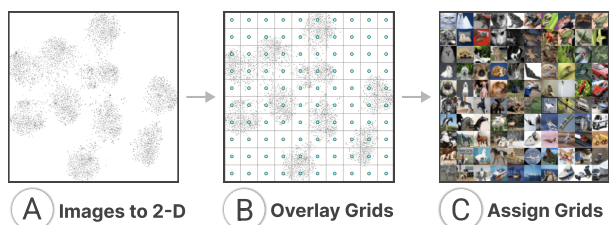


Fig. 7. Steps to generate t-SNE-Grid: From t-SNE embeddings (in A), we first overlay grid points on top of the embeddings (in B; 10×10 in this case). Then in C, we assign each grid with an image that has the smallest distance.

embedding and the corresponding image is displayed on top (like Fig. 7C). The result is a grid of images with the structure from t-SNE.

There may be overlap with what is considered the closest image embedding to each grid point, so to achieve a result where the sum of grid assignment distances is minimized, the Jonker-Volgenant algorithm is used to get the optimal assignments [28]. The optimal grid assignments work by phrasing the problem as a linear assignment problem. For this user study, to enhance the t-SNE-Grid exploration further, we implemented a one-level zoom that recomputes the grid with a smaller number of images based on where the user clicks in t-SNE-Grid. In particular, the top- k closest to the click are recomputed with the Jonker-Volgenant algorithm to display a smaller and more focused grid of images to the user, where k is chosen based on the number of grids to show in the zoomed-in view. For example, to show a 5×5 grid, k is set as 25 to take the 25 closest points and gridify them. We open-sourced the grid assignment implementation and published it as a library⁶.

6.2 Study Setup

6.2.1 Participants.

We recruited 20 participants by using the departmental student mailing lists. Their average age was 26. Five were female and 15 were male. Six were undergraduate and 14 were graduate students. Their degree programs included computer science, robotics, and AI. We recruited only those who have taken at least one AI or ML course. Every participant attended the study in-person and we had one participant per session. Each participant was compensated with a \$20 gift card.

6.2.2 Protocol.

We used a within-subject design such that each participant evaluated both DendroMap and t-SNE-Grid. To let participants work with different images for the two visualizations, we created two variations of the CIFAR-100 dataset (Artifact and Organism subsets which we describe in detail in Sect. 6.2.3). From the two visualizations and two datasets, we created four conditions. Each participant was assigned to one of these four conditions to ensure there was no bias in the order in which a participant used (shown in Table 1).

Every participant completed two sets of tasks, one for each visualization-dataset combination of their respective condition. For each phase, a participant was given a brief tutorial of the visualization, then they were asked to complete seven tasks while thinking aloud. After each phase, the participant filled out a post-questionnaire form. All participants used the same computer setup with a 32-inch monitor.

⁶<https://www.npmjs.com/package/grid-assign-js>

6.2.3 Dataset and Models.

We used the CIFAR-10 and CIFAR-100 datasets [34] for the study. The CIFAR-10 dataset has 10 classes, each containing 6,000 images (5,000 from training set and 1,000 from test set), while the CIFAR-100 dataset has 100 classes, each containing 600 images.

We fine-tuned the ResNet50 [20] architecture that was pretrained on the ImageNet dataset provided by TensorFlow⁷. The CIFAR-10 and CIFAR-100 images were upsampled to fit the input shape of the ResNet50 model (i.e., $224 \times 224 \times 3$). After extracting the image features from the models, we used Average Pooling, followed by three Dense layers (i.e., their sizes are 1024, 512, and the number of classes, respectively). The model was fine-tuned for 20 epochs, achieving a test set accuracy of 92.8% on CIFAR-10 and 76.3% on CIFAR-100. For use in the DendroMap and t-SNE-Grid algorithms, we represented the images in each dataset as high-dimensional vectors from embeddings of one of the last hidden layers in each respective model (i.e., for CIFAR-10, the second-to-last hidden layer, which is 1024-dimensional; for CIFAR-100, the last hidden layer, which is 512-dimensional⁸). The DendroMap and t-SNE-Grid use the same representations as input to their respective algorithms.

We divided the classes of CIFAR-100 into two sets—“Artifacts” and “Organisms”—in order to have two distinct sets of classes for the within-subject design. This helps ensure that results from the first interface only minimally affect those from the second interface. Each set consists of 40 classes (i.e., 4 superclasses, each consisting of 10 classes) [34]. The Artifact set contains classes like chair, television, and bottles, while the Organisms set contains classes like tiger, crocodile, and trout.

6.2.4 Tasks

The participants completed seven tasks which can be divided into two broad categories: grouping and searching. The grouping tasks involved identifying groups of images based on semantically similar properties; the searching tasks involved searching for images based on specific properties. Table 2 provides a summarized description of the tasks.

- In Tasks 1 and 2, participants were asked to categorize images into 3-4 groups based on semantically similar properties. Task 1 was designed to evaluate how users make sense of and categorize images across many (i.e., 40) classes whereas Task 2 focuses on how users make sense of images within a single class. The common objectives of these two tasks include analyzing diversity or any potential bias present in the distribution of the data as well as getting an overview of the data.
- In Task 3, we asked participants to find two large groups, using images from a single class, that have very high classification accuracy and have specific properties. This task was designed to evaluate the scope of subgroup-level error analysis.
- Task 4 is about examining the distribution of images for a single class. This task was designed based on the “characterize distribution” task discussed by Amar et al. [1]. The participants were asked to estimate the approximate proportions of four groups determined based on an attribute (e.g., color of objects).

⁷https://www.tensorflow.org/api_docs/python/tf/keras/applications/resnet50/ResNet50

⁸For CIFAR-10, we chose the layer farther from the output layer, because we wanted to extract lower-level concepts that are less specific to classes for people to explore different types of images within each class [3].

#	Phase 1		Phase 2	
	Visualization	Dataset	Visualization	Dataset
1	t-SNE-Grid	Artifact	DendroMap	Organism
2	DendroMap	Artifact	t-SNE-Grid	Organism
3	t-SNE-Grid	Organism	DendroMap	Artifact
4	DendroMap	Organism	t-SNE-Grid	Artifact

Table 1. Four conditions for counterbalancing the orders of two interfaces in our within-subject design

Task Description

1. **Categorizing images** into groups across 40 classes
2. **Categorizing images** into groups for a single class
3. **Identifying groups** of images with high classification accuracy within a single class
4. Estimating the image count **distribution** over multiple groups within a single class
5. **Searching** for an image with a given text description
6. **Searching** for an image with a given visual description
7. Searching for an **anomalous** image with an incorrect class label

Table 2. Seven tasks designed to evaluate several grouping and searching tasks used in ML analysis

- The following two tasks are conventional searching tasks. In Task 5, participants were asked to find an image that matches a provided text description. In Task 6, participants were asked to find the image that matches the one on the task sheet.
- Lastly, Task 7 was designed to find probable anomalies. Participants were asked to find potential labeling errors among the misclassified images for a single class [40, 64].

Note that every participant worked with the same task list for both DendroMap and t-SNE-Grid, but used a different dataset for each of the visualizations.

6.2.5 Interface Setup

For fairer comparison, the sidebar component from DendroMap was added to the t-SNE-Grid visualization. Additionally, to confirm that certain sidebar components are not overused over the main visualization, the class table, class filtering, and similar images components were removed from the sidebar for both DendroMap and t-SNE-Grid.

6.3 Results

The setup of our user study gives us the scope to analyze data from a multitude of perspectives.

6.3.1 Evaluation of task completion time

Our first set of analyses focused on task completion time. During the study, we recorded the time a participant took to complete each task. We conducted Wilcoxon signed-rank tests, and there is no significant difference between the average time taken by our participants with t-SNE-Grid and that with DendroMap for each of all the tasks.

6.3.2 Evaluation of task responses

We evaluated the responses to the seven tasks using statistical methods.

Task 1. We instructed our participants to identify four groups such that an image can be assigned to only one group (*mutually exclusive*) and most images present in the interface can be assigned to one of the groups (*collectively exhaustive*). To evaluate the quality of groups made by the participants, we conducted three analyses. First, to measure the collectively exhaustive property of the groups, we counted the number of classes covered by at least one of the four groups and divided that number by the total number of classes present in the dataset (i.e., 40). The reason why we counted the number of “classes” instead of “images” is the number of classes can approximate the number of images because each class has an equal number of images. In an ideal scenario, the value would be 1.0. If only a portion of images in a class belongs to a group, we count it as half. With DendroMap, the average value over all participants are higher with a value of 0.82, compared to 0.73 with t-SNE-Grid. A one-sided Wilcoxon signed-rank test indicates that its *p*-value is 0.089. This suggests that on average, participants were able to maintain the “collectively exhaustive” property more with DendroMap than t-SNE-Grid, but we note that the level of significance is not high. Next, to assess the mutual exclusiveness of the groups made by a participant, we counted the number of classes that belong to two or more groups. In an ideal scenario, the value is zero because there is no overlap between the groups. We calculated the average value

to be 0.07 for t -SNE-Grid and 0.13 for DendroMap. The results of the same test indicate that on average participants were able to create more “mutually exclusive” groups with t -SNE-Grid than DendroMap (p -value = 0.062). Lastly, we calculated the entropy score of the probability distribution of the four groups to check how much the groups are equally distributed. We found the average entropy score of DendroMap to be similar to that of t -SNE-Grid (i.e., 1.37 vs. 1.34).

Task 2. Like Task 1, the participants were asked to identify mutually exclusive and collectively exhaustive groups. The main difference for Task 2 is that they worked with images for only one class. To evaluate the quality of groups identified by the participants, we conducted the same three analyses as for Task 1. However, for Task 2, instead of counting the number of classes, we labeled a 10% sample of individual images. In our first analysis of the collectively exhaustive property, the average values for t -SNE-Grid and DendroMap are almost the same with the values of 0.67 and 0.66 respectively. This also happened with the mutual exclusiveness analysis (i.e., 0.10 and 0.13). Our final analysis of the entropy scores is also no exception (i.e., 1.41 and 1.36).

Task 3. This task is also about grouping, but the participants were asked to find two large groups of images with *high classification accuracy*. We conducted two analyses. First, we assessed the average accuracy of the two groups. To find the accuracy of each group, we counted the correctly classified images from the total number of images covered by each group. The average accuracy values of the two groups are 92.2% and 93.2% for t -SNE-Grid and DendroMap, respectively. DendroMap is slightly higher, but there is no significant difference. Second, we measured the size of these groups. The average for t -SNE-Grid is 0.38 and for DendroMap is 0.34, with no significant difference.

Task 4. In this task, the participants estimated the approximate percentage of different cars and birds based on car color (yellow, red, white or silver, or other) or background of birds (e.g., sky), respectively. To evaluate their responses, we counted the number of car and bird images that correspond with the aforementioned criteria and calculated the Kullback-Leibler (KL) divergence score to quantify how much the probability distributions reported by our participants differ from the actual distributions. A score of 0 means the two distributions are the same. Our results show that DendroMap has more counts in between 0.0 and 0.1 than t -SNE-Grid (i.e., 10 vs. 7). This indicates that more participants were closer to the actual distribution when using DendroMap. This is also supported by the medians of the KL divergence scores where the median is 0.10 for DendroMap and 0.17 for t -SNE-Grid.

Tasks 5 & 6. These tasks were about finding specific images. All the participants of our study were successful in finding the correct images using both the t -SNE-Grid and DendroMap.

Task 7. The participants were asked to find labeling errors from misclassified images. Unlike Tasks 5 and 6, multiple correct answers exist. We assessed the images selected by our participants and divided them into three categories: *reasonable, somewhat reasonable, not reasonable*. Based on our assessment of 20 images found by 20 participants, with t -SNE-Grid, 12 are reasonable and 3 are somewhat reasonable; with DendroMap, 15 are reasonable and 3 are somewhat reasonable. This indicates that DendroMap is likely more helpful in finding potential anomalies in image datasets. The images in DendroMap are divided into clusters with distinguishable boundaries, which makes it more convenient to systematically inspect a large number of images than with t -SNE-Grid.

6.3.3 Evaluation of post-questionnaires

Each participant answered 10 questions in two separate post-questionnaire forms: one for DendroMap and one for t -SNE-Grid. They provided ratings on a 7-point Likert scale (7 being strongly agree). The questions and their average ratings are shown in Table 3.

The results indicate that DendroMap received higher ratings than t -SNE-Grid in 8 out of 10 questions. The t -SNE-Grid received a better rating for only the first question regarding the learnability of the visualization. This is reasonable as t -SNE-Grid supports fewer interactions than DendroMap. From the ratings of several important aspects of image visualizations, DendroMap is found to be statistically

Question	t -SNE-Grid	DendroMap
Easy to learn how to use	6.45	6.30
Easy to use	6.00	6.00
Helpful for overview	5.95	6.45 ^o
Helpful for detailed analysis	5.15	6.05 *
Helpful for finding specific images	5.10	5.75 ^o
Helpful to identify image categories	5.70	6.20 ^o
Helpful to discover new insights	5.25	6.00 ^o
Confident when using the tool	5.85	6.05
Enjoyed using the tool	6.10	6.40
Would like to use again	5.80	6.65 *

Table 3. Participants’ average ratings for the two visualizations. DendroMap outscored t -SNE-Grid in 8 out of 10 questions. Bold indicates higher average ratings. * and ^o indicate 95% and 90% statistical significance in one-sided Wilcoxon signed-rank tests, respectively.

significantly more preferable than t -SNE-Grid, such as getting an overview, performing detailed analysis, identifying image categories, and discovering new insights. Moreover, participants on average inclined more towards DendroMap than t -SNE-Grid in mentioning their eagerness to use the tool again.

6.4 Discussion

We observed participants’ usage while they performed the tasks. Based on their usage patterns, we have made a few important findings.

DendroMap provides a more structured workflow. Compared to t -SNE-Grid, it is easier to assess or follow how a user makes certain decisions with DendroMap. In DendroMap, the presence of clusters and the hierarchical relationships within them provide significant semantic information to the participants when they create groups or search images based on certain properties. One participant said: *“The clustering of DendroMap was very intuitive, more so than the grid one where the boundaries between groups were not clearly defined. The ability to click into different levels of clusters was very useful as well.”*

DendroMap helps with extracting more specific properties. Using the semantic information provided by DendroMap, the participants could find more detailed information about different image groups. This is more evident with Task 3 where the participants worked with the images of ships and dogs to find two large groups that have high classification accuracy and specific properties. With DendroMap, the participants mentioned more specific properties compared to t -SNE-Grid. For example, regarding dogs, DendroMap users described their eyes, hair length, and facial structure in addition to generic properties such as size, color, and background. With the t -SNE-Grid, participants mostly described groups using only generic properties.

Image search can be narrowed down more with DendroMap. The hierarchical relationships within the clusters helped the participants narrow their search for a particular image. With DendroMap, they easily found specific clusters with more images similar to the one they were looking for. The sub-clusters present within a cluster then helped them further narrow the search space. On the other hand, with t -SNE-Grid, they had to check a large group of images as there is no structured way of narrowing the search. One participant said: *“With the treemap, the ability to narrow down the search without having to recompute the grid size every time, having some predetermined way of organizing the images, and having the images broken up into clusters made it very easy to scan through the images without getting lost. I was able to quickly filter the exact things I was looking for.”*

Cluster summary provided with DendroMap is helpful. DendroMap provides information about each cluster and sub-cluster, such as the number of images and classification accuracy. The participants found this information useful, especially for Tasks 3 and 4. One participant expressed their liking by saying: *“I like the clusters having details like how many images and the accuracy. Also, the outline of the different clusters having different sizes helped.”*

6.5 Limitations

No empirical study is perfect. We discuss threats to validity.

Different mechanisms for exploration. While DendroMap users can navigate tree-structured data at multiple levels, t -SNE-Grid does not create a hierarchy by default. These differences make DendroMap not-surprisingly do better with deeper levels of hierarchical analysis. Our intent was to compare our method against a popular baseline for ML practitioners, our target population. No matter our intent, the threat that any hierarchical method might show similar improvement over the baseline t -SNE-Grid should still be considered.

Types of images shown. An important potential threat to validity comes down to the image data we used. Depending on the background of the participants, other factors may explain differences in results, such as familiarity of images. In addition, the types of images are potentially an ecological threat to validity. In the real world, datasets may contain more diverse, complicated, and noisier images than what is contained in the CIFAR datasets used in our study. For the purposes of the study, it was necessary to limit the scope for reasonable comparison.

7 EXPERIMENTS: DISTANCE PRESERVATION

Lastly, we evaluate the quality of the cluster structures generated from DendroMap computationally. We quantitatively measure *k-nearest neighbor accuracy*—how well DendroMap preserves the top- k nearest neighbors in the original high-dimensional space.

7.1 Setup

We measure the number of common images in the top- k nearest images between one of the techniques and the original high-dimensional representation of data, while varying k (i.e., the size of nearest neighbor list). This is a common way to evaluate the quality of DR methods [58]. The techniques we compare are: (1) t -SNE, (2) t -SNE-Grid (described in Sect. 6.1), and (3) DendroMap. We performed this experiment over 12 different datasets: CIFAR-10, CIFAR-100, and 10 subsets of CIFAR-10, each from one of the 10 classes. All are trained with ResNet50 (same setup described in Sect. 6.2.3), but for the first two, the high-dimensional representations were taken from the last hidden layer, while those for the 10 subsets were taken from the second-to-last hidden layer.

While we compute Euclidean distances between 2-D points for ranking similar images in t -SNE and t -SNE-Grid which assigns a (x, y) value to each data point, DendroMap needed a different methodology because it additionally encodes hierarchical structures using treemaps. We define a distance between two images \mathbf{x}_i and \mathbf{x}_j in DendroMap by measuring the distance from the node for \mathbf{x}_i in the dendrogram tree to the nearest common ancestor node between \mathbf{x}_i and \mathbf{x}_j . This can be thought of as how many times a user needs to zoom-out from the leaf node for \mathbf{x}_i to reach to the cluster where both \mathbf{x}_i and \mathbf{x}_j belong to.

7.2 Results

Figure 8 shows the results. For each of the 12 plots, the x -axis represents k (in k -nearest neighbor) and the y -axis represents the average number of common images in two top- k image lists. We display up to 300 for 10,000 image datasets and 50 for the 10 class-level CIFAR-10 datasets. As shown in the figure, in all cases, t -SNE outperforms the other two, as we can expect, because t -SNE is designed to optimize this metric. When comparing DendroMap and t -SNE-Grid, DendroMap shares more top- k nearest neighbors with the high-dimensional representations than t -SNE-Grid for all 12 datasets. This indicates that DendroMap preserves the local similarity structures better than t -SNE-Grid.

8 LIMITATIONS AND FUTURE WORK

Computational scalability of clustering. The agglomerative clustering algorithms can be a bottleneck when scaling DendroMap to larger datasets. The naïve algorithms grow by $O(n^3)$ in time but can be brought down to $O(n^2)$ with optimizations [37]. The t -SNE method runs with a time complexity $O(n^2)$ and can use approximation to get to $O(n \log n)$ [56]. Although clustering is less efficient, it only needs to be computed once for interactive use in DendroMap. For the CIFAR-10

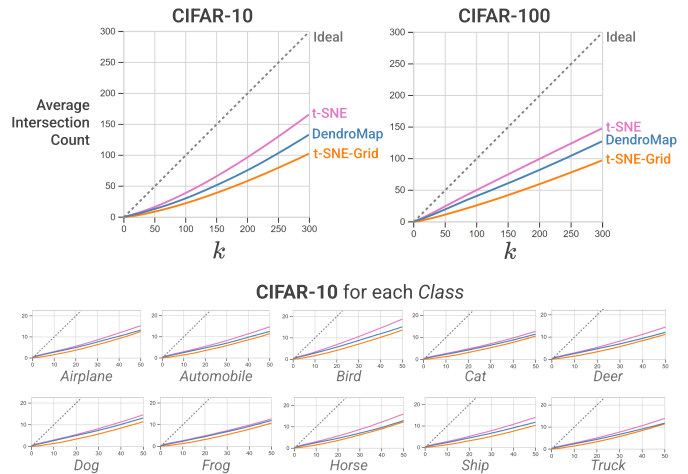


Fig. 8. Average for the number of common k -nearest neighbours between t -SNE, t -SNE-Grid, or DendroMap and high-dimensional representations of images. For all 12 datasets we tested, DendroMap preserves the top- k images better than t -SNE-Grid.

test set with 10,000 images, the clustering algorithm took 36.8 seconds compared to 32.0 seconds for t -SNE⁹ (ran on macOS 12.4, 2.6 GHz 6-Core Intel Core i7 cpu). Future work can investigate more efficient strategies to create hierarchical structures of data.

Comparison with other tree construction methods. In the user study, we compared DendroMap with t -SNE, the most well-known technique (specifically t -SNE-Grid); however, as noted in the limitations in Sect. 6.5, t -SNE does not create an explicitly hierarchical structure. In the future, DendroMap can be compared against a variety of other techniques (e.g., H-SNE [44]) to evaluate the effectiveness of algorithms that produce hierarchical structures.

Interactive refinement of tree structures. While the agglomerative clustering algorithms generate hierarchical structures that allow users to flexibly specify the number of clusters to be displayed, the formed structures may not be ideal for some cases. Visualization researchers have extensively studied interaction methods for steering and refining clustering results [14, 66]. Future research challenges include designing user interactions for refining clustering results in DendroMap.

Using interpretable attributes for tree construction. We used embedding vectors extracted from deep learning models as input to clustering algorithms, but alternative methods may help people better interpret substructures of each cluster in DendroMap. For example, representing each image with human-understandable concepts [31, 71] or additional resources [65] may make each dimension more interpretable. Alternatively, integrating information about each dimension of the embedding vectors into the interface using explainable AI methods can also be helpful [26, 41].

Formalizing interaction operations. Several data manipulation operations can also be provided in DendroMap. For example, sorting images within each node by user-specified criteria (e.g., prediction scores) or splitting and zooming into only a subset of nodes [6, 66]. Formalizing these types of operations would allow for more flexible user exploration. Integrating some ideas presented in the unit visualization literature [43, 48, 61], such as horizontally or vertically separating space based on categorical attributes in Facets-Dive [60, 61], into the treemap context would also be an interesting future direction.

ACKNOWLEDGMENTS

We thank Eric Slyman for their feedback. This work was supported in part by Google Cloud (GCP19980904), NAVER AI Lab, NSF and USDA NIFA (2021-67021-35344), and DARPA (N66001-17-2-4030).

⁹Agglomerative clustering was computed with Ward linkage using SciPy, and t -SNE was computed with the default parameters using scikit-learn.

REFERENCES

- [1] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization (InfoVis)*, pp. 111–117. IEEE, 2005. doi: 10.1109/INFVIS.2005.1532136
- [2] S. Amershi, M. Chickering, S. M. Drucker, B. Lee, P. Simard, and J. Suh. ModelTracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 337–346, 2015. doi: 10.1145/2702123.2702509
- [3] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6541–6549, 2017.
- [4] B. B. Bederson. PhotoMesa: a zoomable image browser using quantum treemaps and bubblemaps. In *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology*, pp. 71–80, 2001. doi: 10.1145/502348.502359
- [5] B. B. Bederson, B. Shneiderman, and M. Wattenberg. Ordered and quantum treemaps: Making effective use of 2D space to display hierarchies. *ACM Transactions on Graphics*, 21(4):833–854, 2002. doi: 10.1145/571647.571649
- [6] G. Bisson and R. Blanch. Improving visualization of large hierarchical clustering. In *2012 16th International Conference on Information Visualisation*, pp. 220–228. IEEE, 2012. doi: 10.1109/IV.2012.45
- [7] M. Bostock. Nested treemap. <https://observablehq.com/@d3/nested-treemap>. Accessed on March 31, 2022., 2019.
- [8] M. Bostock. Zoomable treemap. <https://observablehq.com/@d3/zoomable-treemap>. Accessed on March 31, 2022., 2019.
- [9] J. Buolamwini and T. Gebru. Gender Shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, vol. 81, pp. 77–91. PMLR, 2018.
- [10] Á. A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, and D. H. Chau. FairVis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 46–56. IEEE, 2019. doi: 10.1109/VAST47406.2019.8986948
- [11] A. Chatzimpampas, R. M. Martins, and A. Kerren. t-viSNE: Interactive assessment and interpretation of t-SNE projections. *IEEE Transactions on Visualization and Computer Graphics*, 26(8):2696–2714, 2020. doi: 10.1109/TVCG.2020.2986996
- [12] C. Chen, J. Yuan, Y. Lu, Y. Liu, H. Su, S. Yuan, and S. Liu. OoDAnalyzer: Interactive analysis of out-of-distribution samples. *IEEE Transactions on Visualization and Computer Graphics*, 27(7):3335–3349, 2020. doi: 10.1109/TVCG.2020.2973258
- [13] F. Cheng, Y. Ming, and H. Qu. DECE: Decision explorer with counterfactual explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1438–1447, 2020. doi: 10.1109/TVCG.2020.3030342
- [14] J. Choo, C. Lee, C. K. Reddy, and H. Park. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1992–2001, 2013. doi: 10.1109/TVCG.2013.212
- [15] W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky. HierarchicalTopics: Visually exploring large text collections using topic hierarchies. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2002–2011, 2013. doi: 10.1109/TVCG.2013.162
- [16] F. S. Duarte, F. Sikansi, F. M. Fatore, S. G. Fadel, and F. V. Paulovich. Nmap: A novel neighborhood preservation space-filling algorithm. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2063–2071, 2014. doi: 10.1109/TVCG.2014.2346276
- [17] O. Fried, S. DiVerdi, M. Halber, E. Sizikova, and A. Finkelstein. Isomatch: Creating informative grid layouts. In *Computer Graphics Forum*, vol. 34, pp. 155–166. Wiley Online Library, 2015. doi: 10.1111/cgf.12549
- [18] O. Gomez, S. Holter, J. Yuan, and E. Bertini. ViCE: Visual counterfactual explanations for machine learning models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 531–535, 2020. doi: 10.1145/3377325.3377536
- [19] A. Gomi, R. Miyazaki, T. Itoh, and J. Li. CAT: A hierarchical image browser using a rectangle packing technique. In *2008 12th International Conference Information Visualisation*, pp. 82–87. IEEE, 2008. doi: 10.1109/IV.2008.8
- [20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE Computer Society, 2016.
- [21] W. He, L. Zou, A. K. Shekar, L. Gou, and L. Ren. Where can we help? a visual analytics approach to diagnosing and improving semantic segmentation of movable objects. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1040–1050, 2021. doi: 10.1109/TVCG.2021.3114855
- [22] G. M. Hilaraca, W. E. Marcílio-Jr, D. M. Eler, R. M. Martins, and F. V. Paulovich. Overlap removal of dimensionality reduction scatterplot layouts. *arXiv preprint arXiv:1903.06262*, 2019.
- [23] G. M. Hilaraca and F. V. Paulovich. Distance preserving grid layouts. *arXiv preprint arXiv:1903.06262v1*, 2019.
- [24] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.
- [25] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2674–2693, 2018. doi: 10.1109/TVCG.2018.2843369
- [26] F. Hohman, H. Park, C. Robinson, and D. H. Chau. Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1096–1106, 2019. doi: 10.1109/TVCG.2019.2934659
- [27] J. Hong, K. Lee, J. Xu, and H. Kacorri. Crowdsourcing the perception of machine teaching. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2020. doi: 10.1145/3313831.3376428
- [28] R. Jonker. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38:325–340, 1987. doi: 10.1007/BF02278710
- [29] M. Kahng, P. Y. Andrews, A. Kalro, and D. H. Chau. ActiVis: Visual exploration of industry-scale deep neural network models. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):88–97, 2017. doi: 10.1109/TVCG.2017.2744718
- [30] A. Karpathy. t-sne visualization of cnn codes. <https://cs.stanford.edu/people/karpathy/cnembed/>. Accessed on March 31, 2022., 2014.
- [31] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, vol. 80, pp. 2668–2677. PMLR, 2018.
- [32] M. Kim, K. Kang, D. Park, J. Choo, and N. Elmqvist. Topiclens: Efficient multi-level visual topic exploration of large-scale document collections. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):151–160, 2016. doi: 10.1109/TVCG.2016.2598445
- [33] G. Kogan. Image t-SNE live. <https://m14a.github.io/guides/ImageTSNELive/>. Accessed on March 31, 2022.
- [34] A. Krizhevsky. Learning multiple layers of features from tiny images. 2009.
- [35] W. E. Marcílio-Jr, D. M. Eler, F. V. Paulovich, J. F. Rodrigues-Jr, and A. O. Artero. ExplorerTree: a focus+context exploration approach for 2D embeddings. *Big Data Research*, 25:100239, 2021. doi: 10.1016/j.bdr.2021.100239
- [36] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- [37] D. Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011.
- [38] F. Murtagh. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4):354–359, 1983. doi: 10.1093/comjnl/26.4.354
- [39] A. Ng. A chat with andrew on ml ops: From model-centric to data-centric ai, 2021. <https://www.youtube.com/watch?v=06-AZXmwHjo>.
- [40] C. G. Northcutt, A. Athalye, and J. Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks)*, vol. 1, 2021.
- [41] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev. The building blocks of interpretability. *Distill*, 2018. doi: 10.23915/distill.00010
- [42] M. L. Olson, T.-V. Nguyen, G. Dixit, N. Ratzlaff, W.-K. Wong, and

- M. Kahng. Contrastive identification of covariate shift in image data. In *2021 IEEE Visualization Conference (VIS)*, pp. 36–40. IEEE, 2021. doi: 10.1109/VIS49827.2021.9623289
- [43] D. Park, S. M. Drucker, R. Fernandez, and N. Elmqvist. Atom: A grammar for unit visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 24(12):3032–3043, 2017. doi: 10.1109/TVCG.2017.2785807
- [44] N. Pezzotti, T. Höllt, B. Lelieveldt, E. Eisemann, and A. Vilanova. Hierarchical stochastic neighbor embedding. In *Computer Graphics Forum*, vol. 35, pp. 21–30. Wiley Online Library, 2016. doi: 10.1111/cgf.12878
- [45] N. Polyzotis and M. Zaharia. What can data-centric ai learn from data and ml engineering? *arXiv preprint arXiv:2112.06439*, 2021.
- [46] R. Ratajczak. tsne-grid. <https://github.com/Pandinosaurus/tsne-grid>. Accessed on March 31, 2022.
- [47] P. E. Rauber, S. G. Fadel, A. X. Falcao, and A. C. Telea. Visualizing the hidden activity of artificial neural networks. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):101–110, 2016. doi: 10.1109/TVCG.2016.2598838
- [48] D. Ren, S. Amershi, B. Lee, J. Suh, and J. D. Williams. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):61–70, 2016. doi: 10.1109/TVCG.2016.2598828
- [49] J. Seo and B. Shneiderman. Interactively exploring hierarchical clustering results [gene identification]. *Computer*, 35(7):80–86, 2002. doi: 10.1109/MC.2002.1016905
- [50] B. Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics*, 11(1):92–99, 1992. doi: 10.1145/102377.115768
- [51] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*. IEEE, 1996. doi: 10.1109/VL.1996.545307
- [52] D. Smilkov, N. Thorat, C. Nicholson, E. Reif, F. B. Viégas, and M. Wattenberg. Embedding projector: Interactive visualization and interpretation of embeddings. *arXiv preprint arXiv:1611.05469*, 2016.
- [53] D. Smilkov, N. Thorat, M. Pellat, and L. Peran. Know Your Data: a new tool to explore datasets, 2021. <https://medium.com/people-ai-research/know-your-data-a-new-tool-to-explore-datasets-ba45b7665695>.
- [54] TensorFlow. Basic classification: Classify images of clothing. <https://www.tensorflow.org/tutorials/keras/classification>. Accessed on March 31, 2022.
- [55] P. Van Der Corput and J. J. van Wijk. ICLIC: Interactive categorization of large image collections. In *2016 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 152–159. IEEE, 2016. doi: 10.1109/PACIFICVIS.2016.7465263
- [56] L. van der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15(93):3221–3245, 2014.
- [57] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.
- [58] Y. Wang, H. Huang, C. Rudin, and Y. Shaposhnik. Understanding how dimension reduction tools work: an empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research*, 22(201):1–73, 2021.
- [59] M. Wattenberg, F. Viégas, and I. Johnson. How to use t-SNE effectively. *Distill*, 2016. doi: 10.23915/distill.00002
- [60] J. Wexler. Facets: An open source visualization tool for machine learning training data, 2017. <https://ai.googleblog.com/2017/07/facets-open-source-visualization-tool.html>.
- [61] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, 2019. doi: 10.1109/TVCG.2019.2934619
- [62] S. E. Whang, Y. Roh, H. Song, and J.-G. Lee. Data collection and quality challenges in deep learning: A data-centric ai perspective. *arXiv preprint arXiv:2112.06409*, 2021.
- [63] T. Wu, M. T. Ribeiro, J. Heer, and D. Weld. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1073
- [64] S. Xiang, X. Ye, J. Xia, J. Wu, Y. Chen, and S. Liu. Interactive correction of mislabeled training data. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 57–68. IEEE, 2019. doi: 10.1109/VAST47406.2019.8986943
- [65] X. Xie, X. Cai, J. Zhou, N. Cao, and Y. Wu. A semantic-based method for visualizing large image collections. *IEEE Transactions on Visualization and Computer Graphics*, 25(7):2362–2377, 2018. doi: 10.1109/TVCG.2018.2835485
- [66] W. Yang, X. Wang, J. Lu, W. Dou, and S. Liu. Interactive steering of hierarchical clustering. *IEEE Transactions on Visualization and Computer Graphics*, 27(10):3953–3967, 2020. doi: 10.1109/TVCG.2020.2995100
- [67] J. S. Yi, Y. ah Kang, J. Stasko, and J. A. Jacko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1224–1231, 2007. doi: 10.1109/TVCG.2007.70515
- [68] J. Yuan, C. Chen, W. Yang, M. Liu, J. Xia, and S. Liu. A survey of visual analytics techniques for machine learning. *Computational Visual Media*, 7(1):3–36, 2020. doi: 10.1007/s41095-020-0191-7
- [69] J. Zahálka and M. Worring. Towards interactive, intelligent, and integrated multimedia analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 3–12. IEEE, 2014. doi: 10.1109/VAST.2014.7042476
- [70] J. Zahálka, M. Worring, and J. J. Van Wijk. II-20: Intelligent and pragmatic analytic categorization of image collections. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):422–431, 2020. doi: 10.1109/TVCG.2020.3030383
- [71] Z. Zhao, P. Xu, C. Scheidegger, and L. Ren. Human-in-the-loop extraction of interpretable concepts in deep learning models. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):780–790, 2021. doi: 10.1109/TVCG.2021.3114837