cv:hci @ KIT
Computer Vision for Human-Computer Interaction

Karlsruhe Institute of Technology

Towards Accessible Visualizations with Vision-Language Models

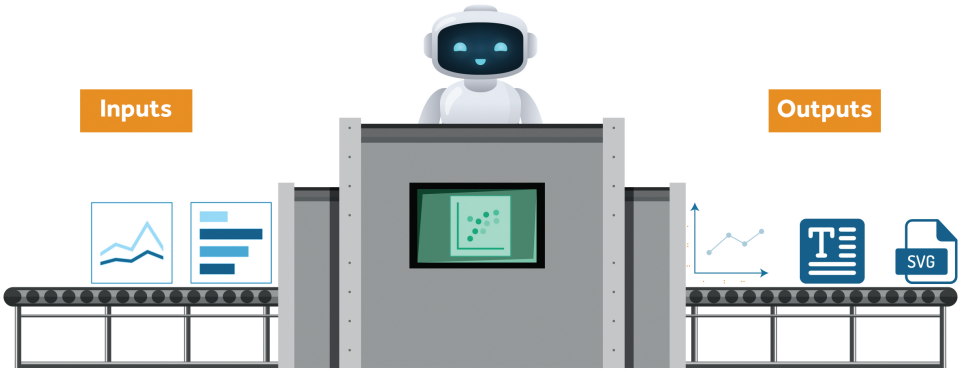
Zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften

von der KIT-Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte
Dissertation

von

Omar Moured
aus Guangdong, Algeria



Tag der mündlichen Prüfung:	12. 12. 2024
Hauptreferent:	Prof. Dr.-Ing. Rainer Stiefelhagen Karlsruher Institut für Technologie
Korreferent:	Prof. Enamul Hoque Prince York University

Width of life is more important than length of life.

— Avicenna

Dedicated to my family and my wife, whose support and inspiration have guided me throughout this journey.

ABSTRACT

Data Visualizations such as charts/plots, and diagrams are multi-dimensional representations commonly used to explore data and communicate insights. They are available in diverse layouts and styles, each tailored to specific analytical needs. For example, with a smartwatch, one may monitor monthly sleeping cycles with a quick glance at a visual plot. Unfortunately, it is estimated that in 2020, approximately 70% of visual content existed in inaccessible modalities for readers with visual impairments. People sharing this content might lack the expertise to make their content accessible or fear the time and labor required to achieve such goals. On the other hand, People with Visual Impairment (PVI) might not be confident in the available assistive tools to enable them to interpret the content independently.

In this thesis, our research focuses on developing visual content analysis systems to assist sighted individuals to make their content accessible and to provide end-to-end access for PVI. More specifically, we investigate how to digitize documents and visuals while ensuring adherence to accessibility guidelines. To this end, we first investigate how to construct and convey layout information from deep learning models to tactile modalities. This approach is based on the idea that diverse categories of documents and variations in visuals can be primarily differentiated by their layout. Similarly, the advantage of tactile materials, compared to mere textual descriptions, lies mainly in their presentation layout. We further explore an application designed to jointly involve inexperienced users and deep learning models to make the conversion process more flexible.

Tactile materials and alternative text are not merely visual drawings or plain texts; they must comply with established standards. Assisting users and training models to author high-quality accessible content involves ensuring both comprehensiveness and adherence to accessibility standards. In a setup where the model is running end-to-end with a blind participant, or a sighted individual is authoring a chart description, we investigate how deep learning can support adherence to these standards and enhance the overall quality.

Given the diverse nature of documents and visuals, we finally ask how to cope with this diversity and adapt models' performance accordingly. Capturing this diversity within the training process is crucial to ensure the robustness of the models in real-life scenarios, such as captured and scanned samples.

Our investigations have unfolded significant insights into accessible digitization, including new benchmarks for connecting vision-language models with assistive technologies. We have developed valuable and intelligent systems from which both inexperienced sighted and blind individuals can benefit. These advancements promise to enhance the usability and accessibility of digital content, making it more inclusive for all users.

ZUSAMMENFASSUNG

Datenvisualisierungen wie Diagramme/Plots und Schaubilder sind mehrdimensionale Darstellungen, die häufig zur Datenexploration und zur Kommunikation von Erkenntnissen verwendet werden. Sie sind in verschiedenen Layouts und Stilen verfügbar, die jeweils auf spezifische analytische Anforderungen zugeschnitten sind. So kann man zum Beispiel mit einer Smartwatch monatliche Schlafzyklen mit einem schnellen Blick auf ein visuelles Diagramm überwachen.

Leider wird geschätzt, dass im Jahr 2020 etwa 70% der visuellen Inhalte in unzugänglichen Modalitäten für sehbehinderte Leser existierten. Personen, die diese Inhalte teilen, verfügen möglicherweise nicht über das Fachwissen, um ihre Inhalte zugänglich zu machen, oder befürchten den Zeit- und Arbeitsaufwand, der erforderlich ist, um solche Ziele zu erreichen. Andererseits sind Personen mit Sehbehinderungen (PVI) möglicherweise nicht überzeugt von den verfügbaren Hilfsmitteln, die es ihnen ermöglichen sollen, den Inhalt eigenständig zu interpretieren.

In dieser Arbeit konzentrieren wir uns auf die Entwicklung von Systemen zur Analyse visueller Inhalte, die sehenden Personen dabei helfen sollen, ihre Inhalte zugänglich zu machen, und gleichzeitig einen End-to-End-Zugang für PVI zu gewährleisten. Insbesondere untersuchen wir, wie Dokumente und visuelle Inhalte digitalisiert werden können, wobei die Einhaltung der Zugänglichkeitsrichtlinien sichergestellt wird. Zu diesem Zweck untersuchen wir zunächst, wie Layoutinformationen von Deep-Learning-Modellen in taktile Modalitäten übertragen werden können. Dieser Ansatz basiert auf der Idee, dass verschiedene Kategorien von Dokumenten und Variationen von Visualisierungen hauptsächlich durch ihr Layout unterschieden werden können. Ebenso liegt der Vorteil von taktilen Materialien im Vergleich zu bloßen Textbeschreibungen hauptsächlich in ihrem Layout. Darüber hinaus erforschen wir eine Anwendung, die unerfahrene Nutzer und Deep-Learning-Modelle gemeinsam einbindet, um den Konvertierungsprozess flexibler zu gestalten.

Taktile Materialien und Alternativtexte sind nicht nur visuelle Zeichnungen oder einfache Texte; sie müssen den etablierten Standards entsprechen. Die Unterstützung von Nutzern und das Training von Modellen zur Erstellung qualitativ hochwertiger, zugänglicher Inhalte erfordert sowohl Vollständigkeit als auch die Einhaltung der Zugänglichkeitsstandards. In einer Situation, in der das Modell End-to-End mit einem blinden Teilnehmer arbeitet oder ein sehender Nutzer eine Diagrammbeschreibung verfasst, untersuchen wir, wie Deep Learning die Einhaltung dieser Standards unterstützen und die Gesamtqualität verbessern kann.

Angesichts der Vielfalt von Dokumenten und Visualisierungen stellt sich schließlich die Frage, wie mit dieser Vielfalt umgegangen und die Leistung der Modelle entsprechend angepasst werden kann. Diese Vielfalt im Trainingsprozess zu erfassen, ist entscheidend, um die Robustheit der Modelle in realen Szenarien, wie z. B. bei erfassten und gescannten Proben, sicherzustellen.

Unsere Untersuchungen haben bedeutende Erkenntnisse zur barrierefreien Digitalisierung hervorgebracht, einschließlich neuer Benchmarks für die Verbindung

von Vision-Language-Modellen mit assistiven Technologien. Wir haben wertvolle und intelligente Systeme entwickelt, von denen sowohl unerfahrene sehende als auch blinde Personen profitieren können. Diese Fortschritte versprechen, die Benutzerfreundlichkeit und Zugänglichkeit digitaler Inhalte zu verbessern und sie für alle Nutzer inklusiver zu machen.

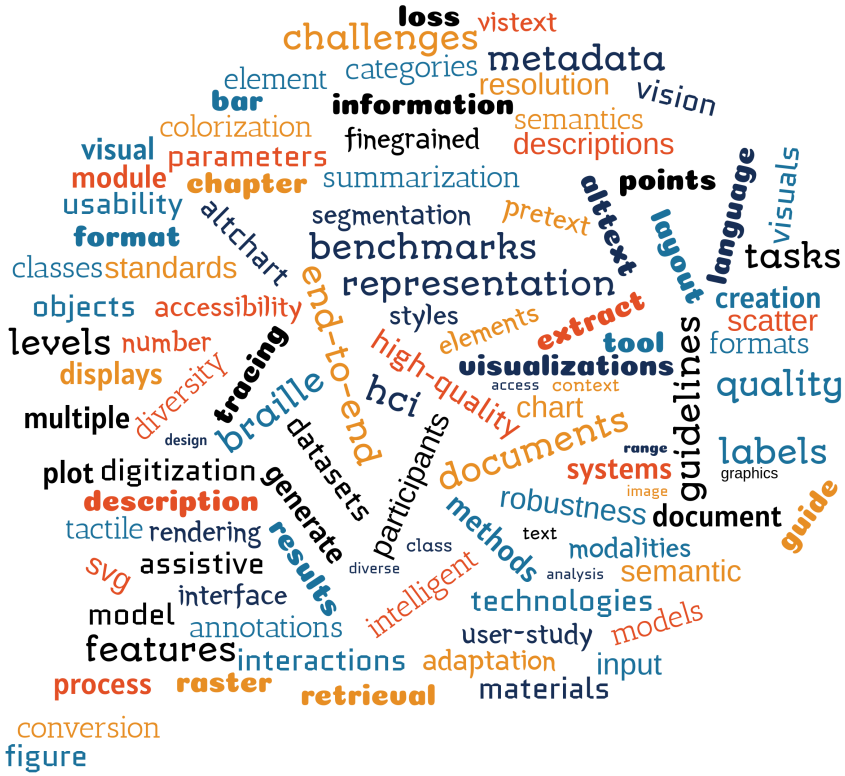


Figure 1: Cloud of frequently used keywords in the thesis.

ACKNOWLEDGMENTS

The research conducted over the past three and a half years at CV:HCI, which has lead to this dissertation, would not have been possible without continuous guidance, encouragement and inspiration of my advisors, mentors, colleagues, friends and family.

First, I want to thank Professor Dr.-Ing. Rainer Stiefelhagen for his invaluable support and guidance throughout my PhD. I am very grateful for the wide range of opportunities that I was given at the lab: conducting research, writing proposals, designing lectures, supervising students, and coordinating the practical course.

I would also like to thank Prof. Enamul Hoque Prince for agreeing to be a reviewer of this thesis, for providing valuable feedback, and for his continuous contributions to this field, which served as a source of inspiration during the initial phase of my thesis.

This work would not have been possible without my dearest friends and colleagues. Thanks to Dr. Jiaming for his support from the beginning, which significantly boosted my performance, and to Yufan, with whom many ideas were developed through our interesting discussions. I must also acknowledge Sara, from whom I learned a lot about conducting user studies and who provided great support.

Special thanks to the accessibility center ACCESS@KIT, which gave me the opportunity to conduct the valuable user studies presented in this dissertation. I am especially grateful to Dr. Karin, Dr. Thorsten, Gerhard, Ann-Christin and Christina. Without their collaboration, this work would not have been evaluated as effectively. I also extend my gratitude to all my fellow co-workers at CV:HCI, including Kunyu, Alexander, Zdravko, Simon, Saquib, David and Corinna for the wonderful time at the lab.

I want to thank my family and friends for their encouragement and for being there for me during difficult times. Most importantly, I thank my parents for their love and support over the past years. They have always encouraged me in my research and in pursuing my PhD studies. Without them, this work would not have been possible.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

This thesis is driven by the end-goal of ensuring equal access rights for all individuals, regardless of their sensory impairments, specifically in the domain of accessible visualizations. To achieve this goal, this thesis presents two types of AI-based algorithms adhering to accessibility guidelines: those that assist sighted people in converting their charts into accessible formats, and those that enable independent, end-to-end access for PVI individuals. Our development seeks to contribute to the field of visualization accessibility by developing and integrating the latest deep learning techniques into assistive technologies, reducing labor effort, and ensuring high-quality content.

1.1 ACCESSIBILITY OF VISUALIZATIONS AND DOCUMENTS

The last few decades have demonstrated that we live in the age of "Big Data." As the volume and complexity of data increase, people require more powerful representation forms that allow them to capture insights at a glance. The challenge becomes "how to interpret this data effectively." As a result, people often create compelling visualization images and use them as supportive materials. These visualizations not only make complex data more understandable but also enhance the communication of insights. However, they are frequently shared as raster images, consisting solely of pixels, which makes them inaccessible to PVI individuals. According to the World Report on Vision 2020 by the WHO [144], approximately 596 million people worldwide suffer from vision impairment. In Europe, there are estimated to be over 30 million PVI, with an average of 1 in 30 Europeans experiencing sight loss [52]. Imagine you open an article about climate change and encounter the following text:

"12, 8, 15, 5, 10, 50, 100, 150, 7-day average temperature"

You might wonder what this is supposed to be. Perhaps it's a mistake, with the author accidentally pasting something from a spreadsheet and forgetting to remove it before publishing the article. What is shown above is a typical output a PVI person might hear from their screen reader when encountering a data visualization on a document or website. While one person sees a compelling chart with statistics about temperature variations over a week, another person hears an incomprehensible string of dates and numbers without context.

Screen reader software makes this inaccessibility especially evident. When it encounters an image, video, or other media, it reads aloud the alternative (alt) text provided in the embedded alt attribute of the <img> tag. Often, visualizations are

simply not detectable by screen readers, rendering them invisible to the user [165]. In other cases, as in the example above, parts of the visualization are recognized by the screen reader but result in an output that is neither comprehensible nor useful. Upon closely examining the example text above, one might speculate that the numbers at the start ("12, 8, 15, 5, 10") are the tick marks of one axis of a chart, and the following numbers ("50, 100, 150") are tick marks of the other axis. The text "7-day average temperature" at the end might be a legend or a label of a data series in the chart. However, without visual context, it is difficult to determine what the visualization is actually conveying.

Tactile materials offer an alternative to screen readers by providing haptic perception, where visual elements are depicted as raised dots with different heights indicating various elements on an embossed sheet, wood, or other materials. Although screen readers provide one-dimensional audio/text information, tactile materials allow PVI individuals to form their own interpretations of images rather than relying on written descriptions. However, as you may infer, these materials require significantly more effort and expertise from sighted individuals to design and produce.

The accessibility of data visualizations for people with disabilities has historically been neglected by research [119]. However, this field is now experiencing increasing interest from the vision and linguistic research communities, addressing the issue from different angles. Some researchers are exploring how to generate meaningful image descriptions using Vision-Language (V-L) models [93, 173], while others are investigating tactile modalities and exploring ways to streamline the creation process [48]. Additionally, another group [95] has examined the current state of the field and identified opportunities and challenges for improvement.

In this thesis, we approach the problem in two distinct yet related directions. From the perspective of sighted individuals, who may be non-experts, we consider intelligent applications that enable them to adapt their content with accessible modalities. The other direction focuses on PVI individuals, where we consider an end-to-end approach that allows them to independently access image content.

1.2 MOTIVATION AND GOALS

The primary motivation of this thesis is to enhance the accessibility of visual content for PVI, thereby ensuring equal access to information. More specifically, it aims to broaden the boundaries of assistive technologies by developing and leveraging state-of-the-art deep learning models. This involves not only building benchmarks and systems but also learning standards and guidelines.

The goal of this thesis is to propose innovative and effective solutions for visual content analysis that adhere to accessibility guidelines and improve the quality and usability of accessible visualizations. In particular, we aim to develop AI-based algorithms that assist sighted individuals to facilitate the conversion of visual content into high-quality alt-text and tactile formats. Additionally, we aim to create end-to-end systems that enable PVI to access and interpret visual content indepen-

dently. In the following chapters, we will explore how we developed intelligent user interfaces based on our AI models to assist in digitizing visualizations and filling accessibility gaps. Through these efforts, we strive to reduce labor efforts, ensure high-quality accessible content, and foster inclusive digital experiences.

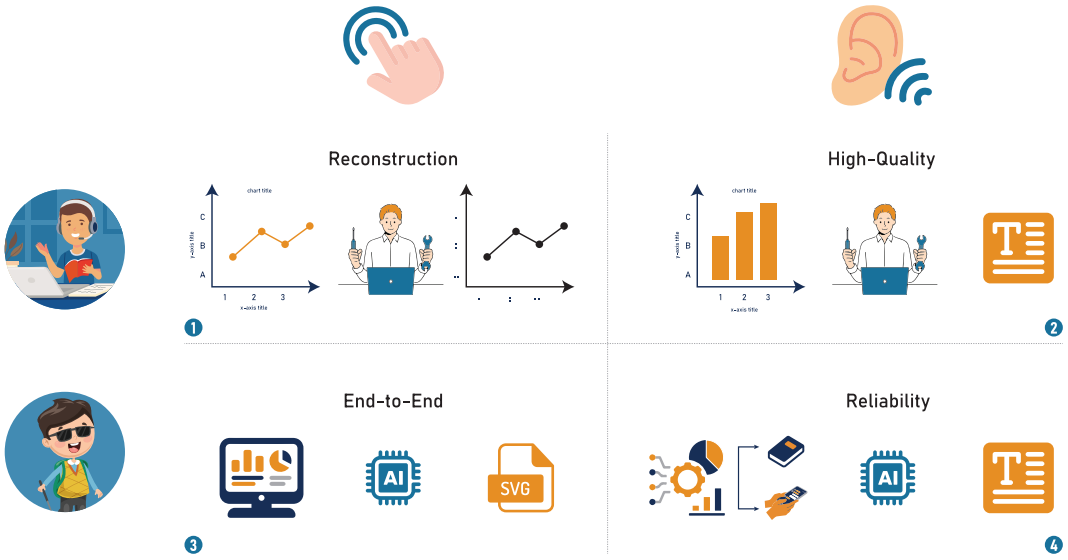


Figure 2: Thesis contributions can be visualized as a 2x2 matrix. Each column represents an accessible modality: tactile and text, respectively. Each row corresponds to the targeted group: sighted assistants and PVI individuals.

1.3 THESIS FOCUS

A typical type of data visualization found in documents is charts. Sharif et al. [165] have found in an empirical study that PVI face significant barriers when accessing these types of data, as they are often either not detectable, incomprehensible, or barely usable. Sharif et al. have called for more research on making charts more accessible for people with disabilities. Given my passion for document analysis and the use of deep learning models for deep document understanding tasks, I have taken this need for more research as an opportunity to focus on the accessibility of charts.

Working closely with the "Center for Digital Accessibility and Assistive Technology" team at KIT [1], which assists PVI in their university education, has allowed my research to contribute to two research directions, as seen in Figure 2. The first direction focuses on assisting sighted individuals in (1) **reconstructing charts** into tactile modalities (Figure 2-1) and (2) authoring **high-quality** alt-text (Figure 2-2). The second direction aims to enable blind individuals to access content independently by providing (3) **end-to-end** tactile materials (Figure 2-3) and improving the (4) **robustness and reliability** of such systems for integration with assistive technologies (2-4). In my thesis, all contributions were primarily driven by interviewing and working closely with relevant individuals and conducting thorough user studies. Although we had a limited number of participants, the long-term

1 https://www.access.kit.edu/english/index.php

discussions during the three-year research period led us to propose novel contributions.

1.4 RESEARCH QUESTIONS

This thesis is one contribution towards answering the call for more research in the specific area of accessible visualizations for PVI individuals. The overarching question this work seeks to address is:

"How can we utilize deep learning models to enhance visualization accessibility for PVI?"

This question has been further divided into three research questions, each addressed by different parts of this work. Together, their results contribute to answering the overarching question. These questions and their resulting contributions are as follows:

RQ1: How can we digitize and depict the layout information of documents and charts for PVI using deep learning models?

The diverse categories of documents and variations in charts are primarily distinguished by their layout. Similarly, tactile materials offer advantages over textual descriptions mainly through their presentation layout. Localizing the content in a document page requires skimming through the layout, and understanding the metadata content of charts requires detailed layout information. The challenges include not only detecting the layout but also understanding the inter-relationships between instances. Previous work has attempted to address this problem using sparse high-level representations such as bounding boxes. In contrast, our work contributes with more dense, low-level pixel-wise paradigms, providing a more detailed, dense paradigms.

RQ2: How can V-L models be trained to comply with accessibility standards to generate high-quality accessible modalities

Building on the first question, whether an AI model is assisting in authoring alt-text or creating tactile materials, it must be aware of accessibility standards, particularly when working with inexperienced users who are prone to errors. This section discusses how V-L models can be trained to adhere to accessibility standards and generate high-quality accessible modalities, including tactile and auditory formats. The focus is on developing models that ensure both compliance and usability, providing support for inexperienced users in creating accessible content.

RQ3: What strategies can models employ to cope with the diverse designs and input formats of visualizations, ensuring robustness and effective handling of real-life scenarios?

The scarcity of well-annotated accessible data remains a significant challenge. Collecting such data requires meticulous quality checks, and labeling from scratch is similarly difficult. Hence, learning to handle real-life samples while ensuring accessibility compliance is a major challenge. This work contributes to understanding how V-L models can learn better representations of complex chart images from a few samples, thereby reducing hallucinations and enriching semantic information.

Moreover, we discuss strategies for benchmarking model robustness to the diverse nature of charts, aiming to better understand real-life performance and integration with assistive technologies.

1.5 THESIS OUTLINE

In this thesis, we analyze state-of-the-art models for visual chart understanding, ranging from simple x-y plots to multi-panel and compose charts. Our approach includes preprocessing educational materials like textbooks and slides, applying layout analysis neural networks to recognize figures, and then analyzing contextual information to generate high-quality summaries and tactile materials. This process aims to improve the accessibility of educational content for PVI.

Chapter 2: Related Work. This chapter provides an overview of the existing literature and previous research related to document and graphical content analysis, highlighting relevant accessibility concerns. First, we discuss several state-of-the-art document and chart understanding models, covering tasks ranging from layout analysis to metadata extraction for content summarization and reconstruction. Next, we present a detailed list of available benchmarks, emphasizing their limited applicability for assistive technologies. Finally, we investigate and present recent chart assistance systems to provide preliminary knowledge and insights into assistive technology for chart understanding.

Chapter 3: Accessible Digitization of Documents and Visuals. To apply conventional deep learning approaches to assistive technologies, we require a post-processing step that converts their outputs into accessible modalities. In this chapter, we analyze various methods for fine-grained localization to extract metadata from different document and chart regions. We then explore how to present these detailed paradigms in assistive technologies, both for sighted individuals using interfaces to convert content into tactile materials and for PVI individuals to access this content haptically. In summary, this chapter discusses the construction and refinement of an "accessible layout" through deep learning models.

Chapter 4: Learning to Comply with Accessibility Standards. Previous deep learning approaches to chart analysis have shown promising performance on various benchmarks; however, the majority are trained without considering compliance with accessibility guidelines and standards. For assistive technologies, whether an AI model is assisting in authoring alt-text or creating tactile materials, it must adhere to accessibility standards, especially when collaborating with inexperienced users. In this chapter, we discuss two case studies: first, how to train and utilize models to guide sighted users in authoring high-quality alt-text; second, exploring new perspectives on how vision-language models can generate tactile materials from images while managing cognitive aspects and adhering to the guidlines.

Chapter 5: Capturing the Diverse Nature of Visualizations. Diverse chart types pose a challenge in employing a single model for all charts, as chart styles and structures can vary significantly even within the same chart class. Although different styles and structures create appealing visualizations, they also theoretically introduce higher dimensionality and many variables. In this chapter, we investigate how assistive technologies can address this problem. We begin by evaluating the robustness of available state-of-the-art models to input diversity in comparison to human performance in a user study. Subsequently, we propose a new training approach utilizing pre-text tasks to equip models with better adaptability to diverse visualizations.

1.6 PUBLISHED CONTRIBUTIONS

The contributions presented in this thesis were published at several computer vision-related venues. The first step in our system chain for chart accessibility is analyzing document layout, localizing the chart, and extracting fine-grained relevant information (Chapter 3).The proposed approach for document layout accessibility is published in [131]. Following the detection of the chart, in the work of [134], we analyze the chart's fine-grained elements, exploring inter-relations and relevant information from the segmented instances. At [132], We demonstrate a use case of our findings in a designed user interface, which streamlines the conventional tactile material creation process. In Chapter 4, we discuss the limited capabilities of current deep learning models to comply with accessibility guidelines. We tackle this problem with a introducing a new datasets and novel neural architectures for both text and tactile modalities: first, by authoring high-quality alt-text through image retrieval as a reference for sighted people [133]; second, by providing a solution for PVI to access charts through end-to-end tactile material generation from raster images [130]. Finally, in Chapter 5, we introduce an approach to enhance the robustness of vision-language models to handle the diverse nature of chart inputs without the need to synthesize additional data, thus avoiding bias towards certain styles. In the paper [135], We tackle this from the vision encoder side by empowering it to better digest and encode the input image into the latent space. A full list of my publications can be found in Section C.

2

BACKGROUND AND RELATED WORK

This thesis addresses the task of chart accessibility for PVI, which requires a high-level understanding of both the textual and visual content from chart images. To this end, we propose several novel deep learning techniques for chart processing and converting them into accessible modalities. Traditionally, this task has been approached using classical computer vision heuristics, which require extensive tuning and laborious work. Additionally, the available benchmarks are often either synthetic or fail to adhere to accessibility guidelines.

In this chapter, we present a broad overview of the latest chart analysis models, benchmarks, and systems, focusing on accessibility issues, particularly in chart-to-text and chart-to-tactile settings. Section 2.1 reviews networks for object localization employed on document and chart images, including both end-to-end approaches and heuristic methods. We explore methods for visual and textual embedding based on vision-only models as well as multi-modal representations using vision-language neural networks. In Section 2.2, we compare popular datasets for pixel-wise recognition in educational materials, discussing their limitations from an accessibility perspective. Finally, in Section 2.3, we discuss several use cases of models and benchmarks for assistive technologies, such as user interfaces to assist sighted individuals in creating tactile materials and authoring alt-text.

2.1 DOCUMENT AND CHART VISUAL ANALYSIS

The fields of computer vision and natural language processing have significantly reshaped research in document and chart analysis. Some researchers define this field as the acquisition of knowledge from documents, often involving extensive handcrafting [8]. Others describe it as the task of extracting suitable symbolic representations from documents, such as text, that computers can subsequently process [82]. However, based on my study on accessibility, I would define it as **a field that aims to make knowledge accessible to people in the symbolic representation that they need**, not necessarily just for computer processing. For example, summarizing a scientific paper helps researchers quickly grasp key findings. Likewise, transforming document visuals into print tactile formats for PVI. Recent years have seen increase interest from AI communities in this field, as seen in Figure 3, Unfortunately, the accessibility field has not kept pace with these advancements and is lagging behind. According to statistical investigations, the number of papers addressing document and chart accessibility concerns has decreased significantly, with only 276 papers in 2023 compared to 545 in 2022. Next, we discuss in detail the layout and robustness analysis concerning accessibility aspects.

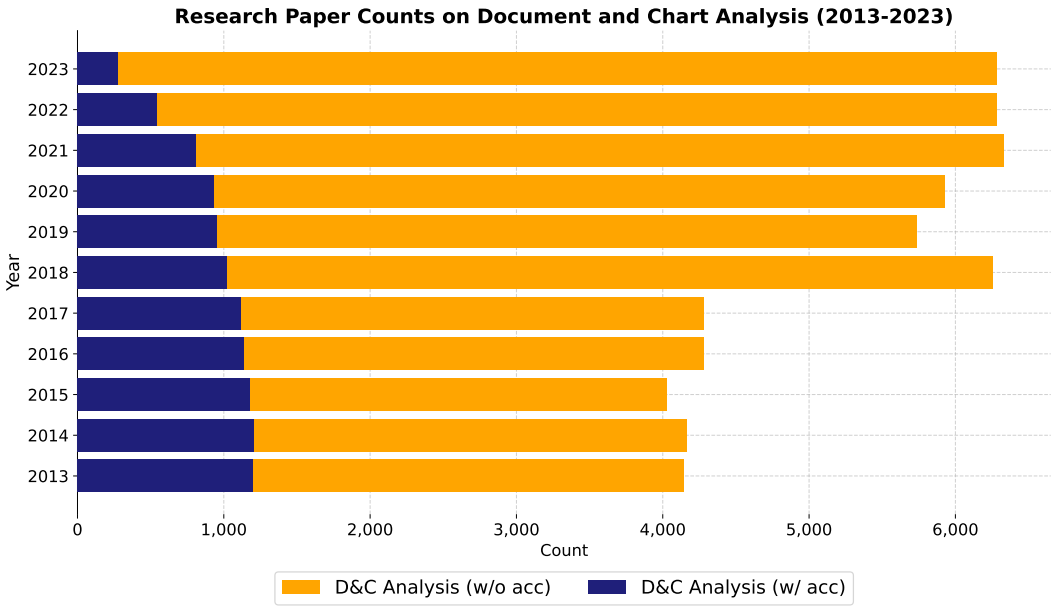


Figure 3: The number of papers addressing document accessibility (acc) concerns over the past 10 years has remained significantly lower and has been consistently decreasing over time.

2.1.1 *Document Layout Analysis*

Document Layout Analysis (DLA) technology involves detecting and identifying internal components within a document image using algorithmic methods. This often includes logical grouping methods that utilize contextual information for layout understanding. Common document categories encompass text, titles, figures, tables, and equations, and these groupings are primarily determined by the document type [61]. To represent the layout, researchers have explored various forms of description, initially using rectangular regions (bounding boxes) [66, 80], and are now advancing towards using segmentation masks [15].

From the early 1990s to the present day, techniques in DLA can be broadly classified into two main groups: heuristics, and deep learning methods.

Hueristics. Falls into mainly two criteria. The first criterion concerns *"how"* the document is analyzed, employing either bottom-up, top-down. Bottom-up techniques begin at the pixel level, progressively grouping these pixels into larger areas, from connected components to meaningful text or non-text regions (e.g., figures). Notable algorithms in this category include RLSA [182], Docstrum [142], and Voronoi diagrams [96]. In contrast, top-down techniques start with the entire document and break it down into basic components, as seen in the X-Y cut algorithm [139]. The second criterion differentiates techniques based on whether they analyze the physical or logical document layout. Physical layout analysis identifies homogeneous regions on the page, while logical layout analysis assigns functional information, or labels, to these regions. Methods are classified according to the downstream tasks they support. For example, Strouthopoulos and Papamarkos [171] used an Artificial Neural Network (ANN) to classify 8×8 document patches as graphics or

halftones. Wu et al. [188] segmented text regions using split-or-merge operations guided by a binary SVM classifier. After segmentation and classification of page objects, post-processing techniques may be applied to generalize results across different layouts [14].

Deep Learning. In the realm of deep learning-based approaches, object detectors have been extensively studied in computer vision. Adapting models like Faster R-CNN [157] and Mask R-CNN [91] has led to notable improvements in object detection performance. For instance, DOLNet [129], which employs a dual backbone ResNext-101 with deformable convolution, achieves impressive results across seven different document benchmarks. Similarly, HiM [22] and VSR [198] are state-of-the-art methods for DocBank [104] and [200], respectively. These methods utilize a Region Proposal Network (RPN) in combination with textual embedding and a graph structure to refine document objects in a multi-modal manner. Since the introduction of the transformer architecture [180], the DLA community has started leveraging transformer-based models for various tasks, including the analysis of scientific articles. LayoutLMv3 [80] is notable for being the first multi-modal architecture that does not rely on pre-trained visual extractors. By integrating visual, textual, and linear embeddings in a transformer-based model, it achieves state-of-the-art performance on benchmarks like PubLayNet [200] and others. Additionally, models such as DocFormer [5] and DiT [102], which belong to the same family, have demonstrated remarkable results, particularly in ICDAR table competitions [63] and RVL-CDIP [71] datasets.

Layout & Accessibility. Among the aforementioned state-of-the-art work, none have truly considered accessibility as a motivation or addressed the needs of PVI. On the other hand, the accessibility community explores this field to find alternative methods that allow PVI readers to *interact* with spatial layout information [18]. For instance, T. Ishihara and H. Takagi [83] proposed an algorithm for analyzing the visual layout of objects in presentation slides. They utilized grouping heuristics called the *parent-child relationship* to find inter-relations between different objects, forming a tree-structured graphical user interface known as *DocExplorer*. Another recent work by L. Lu [185] proposed *SciA11y*, a tool to convert PDF documents into a sequential HTML representation suitable for screen readers. This tool uses the textual layer of the PDF file to segment different entities and is specifically designed for structured scientific documents.

A notable observation is that the AI and accessibility communities have differing focuses. The AI community concentrates on extracting metadata and analyzing context for tasks such as QA, summarization, or indexing. In contrast, the accessibility community aims to improve human-document interactions. Furthermore, the AI community mainly uses digital documents for training and evaluation, while the accessibility community deals with documents found in real-world situations, such as scanned or captured documents. This thesis aims to bridge this gap by incorporating accessibility considerations into AI-driven solutions.

2.1.2 *Chart Understanding*

Chart understanding requires that a model can interpret chart content and execute tasks based on given instructions. This domain includes both low-level recognition tasks like data extraction [107] and high-level tasks such as question-answering (QA) [90, 121, 126], summarization [93, 143], and re-construction [70].

Pipeline Methods. Since charts frequently contain OCR text crucial for data interpretation and often require numerical calculations, chart understanding demands strong text recognition and computational reasoning abilities from the model. Early methods [59, 78, 107] used pipeline approaches, employing off-the-shelf OCR tools or component detectors to convert charts into data tables and other textual forms. These pipelines then utilized language models to perform the specified tasks. However, these approaches were limited by their inability to optimize jointly and suffered from error accumulation.

End-to-end Methods. Recent research [70, 108, 120] has transitioned to end-to-end methods utilizing multimodal large language models (MLLM). These studies adopt the framework of MLLMs [106, 113, 115] and enhance chart understanding through supervised fine-tuning [145] with extensive chart instruction data [70, 122, 124]. While these models show improved performance, their large parameter sizes make them challenging to train or deploy in resource-constrained environments, such as assistive technologies. The latest work so far, TinyChart [196] demonstrates that a 3B MLLM can achieve state-of-the-art performance on several chart understanding tasks by employing the Program-of-Thoughts (PoT) learning strategy [31], which trains the model to generate Python programs for numerical calculations, thereby reducing the burden of learning complex numerical computations.

Table 1: Performance of state-of-the-art models on Chart-to-Text (BLEU4 [146]) and Chart-QA (5% relaxed accuracy) benchmarks.

Model	#Param	Resolution	Chart-to-Text	Chart-QA
Pix2struct [99]	282M	1024×1024	10.30	56.00
Matcha [108]	282M	-	12.20	64.20
UniChart [120]	201M	960×960	12.48	66.24
ChartInstruct [122]	7B	960×960	12.81	61.52
ChartLlama [70]	13B	336×336	14.23	69.66
ChartAst [124]	13B	448×448	15.50	79.90
TinyChart [196]	3B	768×768	17.18	83.60

As shown in Table 1, where the models are listed in chronological order, we observe a trend: earlier shallow models exhibit lower performance despite benefiting from higher resolution images, while newer, larger models achieve better results due to pretraining on large synthetic datasets. Currently, the community is moving toward balancing these two approaches. Smaller models are favored for their efficiency but still benefit from pretraining on available large created benchmarks. This shift is likely driven by the need to integrate these models into human assistive systems, such as document reading tools.

Accessible Charts. People with visual impairments often struggle with bitmap images of charts [13, 169]. To address this, the Web Content Accessibility Guidelines recommend offering a textual description of the chart alongside its graphical representation as alternative text[38]. Yet, such descriptions are rarely created by authors [13]. To reduce the manual effort in this task, some works automate alternative text generation [10, 54, 92]. Unfortunately, such tools have several issues, such as producing irrelevant information and hallucinations [92, 173].

While the alternative text is useful for describing charts, it is not always enough for those with visual impairments [4, 187]. There are three alternative ways to present graphical information for visually impaired people: (1) Tactile graphics, using relief elements for haptic perception [164], (2) Alternative Text, describing graphical content in words [41], or with screen readers [16, 204] and (3) Sonification, mapping raw data values to a diverse range of sounds, varying pitch, frequency, and tone to enable easy distinction between line trends [20, 75, 76].

Tactile graphics can be in different formats: embossed paper, swell paper, thermoform, laser cut, and 3D printed [21]. These formats typically provide better dot resolutions, utilizing various pin height levels to represent more information, and they're cost-effective and portable for individuals with visual impairments [51, 101]. However, they lack the capability for advanced interaction. The second format is digital tactile displays, which can be refreshed and offer additional features such as zooming, interaction buttons, and audio output [131]. Nonetheless, this option tends to be more expensive and offers lower pin resolution.

Having access to a chart's raw data, which includes plot data, titles, axis labels, and descriptions, is highly beneficial. For example, some works use this data to create tactile graphics like Audio-Tactile charts, which combine touch interaction with audio feedback [6, 48]. Similarly, Sonification and alternative text descriptions benefit from raw data access. In terms of raw data extraction methods, they can be categorized as: (1) Manual, requiring human intervention without automated tools, like Data Thief [177], (2) Semi-automatic, combining automatic features with some human intervention [81, 89, 123, 160, 177, 179], and (3) Fully automatic, with no manual intervention [36, 98], though they have limitations in the conversion accuracy.

Given the importance of raw data, vector graphics, such as SVGs, offer a promising alternative representation [64]. SVGs, created using a chart's raw data, have features that make them more accessible than bitmap images [56]. Their structure supports tactile printable graphics creation [97]. For instance, Braille printers can emboss raised dots to mark outlines [97]. Additionally, methods like LineSpace [172], which use 3D printer filament to print SVG file elements, show the potential of SVGs. In conclusion, SVGs can complement traditional alternative text descriptions for charts.

2.1.3 *Robustness Aspects of Visualizations Models*

Existing works in chart analysis focus on clean image data, typically collected digitally from the web or scientific papers, often overlooking real-world issues such as noise and disturbances. As mentioned earlier, a major gap in assistive systems is their aim to help PVI individuals in their daily lives, accessing samples that could be captured or scanned and are available in the wild. Consequently, when these solutions are applied in real-world scenarios, they often exhibit significant performance drawbacks, as we will discuss later.

Robust Visual Architectures. A robust visual architecture is essential for reliable visual analysis. Significant research has been conducted in the areas of object detection [45, 69, 184] and image classification [43, 128, 147]. Modas et al. [128] introduced several primitives that enhance robustness in the field of image classification. The R-YOLO model [184] presents a robust object detector capable of performing under adverse weather conditions. The Fully Attention Networks (FAN) model [201] aims to strengthen robust representations through fully integrated attention mechanisms. Additionally, a Token-aware Average Pooling (TAP) module [68] has been proposed to involve the local neighborhood of tokens in the self-attention process. Despite these advancements, applying existing robust methods directly to domain-specific tasks such as DLA does not yield optimal performance due to unique challenges.

Document Robustness. Document restoration and rectification focus on enhancing document image quality by correcting distortions. DocTr++ [55] investigates unrestricted document image rectification. In [16], the robustness of document image classification against adversarial attacks is examined. Auer et al. [7] present a challenge for robust document layout segmentation. To address this, Zhang et al. [197] develop a WeChat layout analysis system. Robustness evaluation on the RVL-CDIP dataset [71] is conducted for document classification. Tran et al. [176] propose a robust Document Layout Analysis (DLA) system utilizing a multilevel homogeneity structure. Recently, Y. Chen et al. [33] have systematically studied real-world challenges for the first time. They examined extensive perturbation types, encompassing three datasets, five perturbation groups, 12 distinct types, and three severity levels for each type.

> At the time of writing this thesis, there has not yet been a notable study in the chart or visualization domain that measures the robustness and reliability of models against perturbations and noisy samples.

2.2 BENCHMARKS CHART UNDERSTANDING

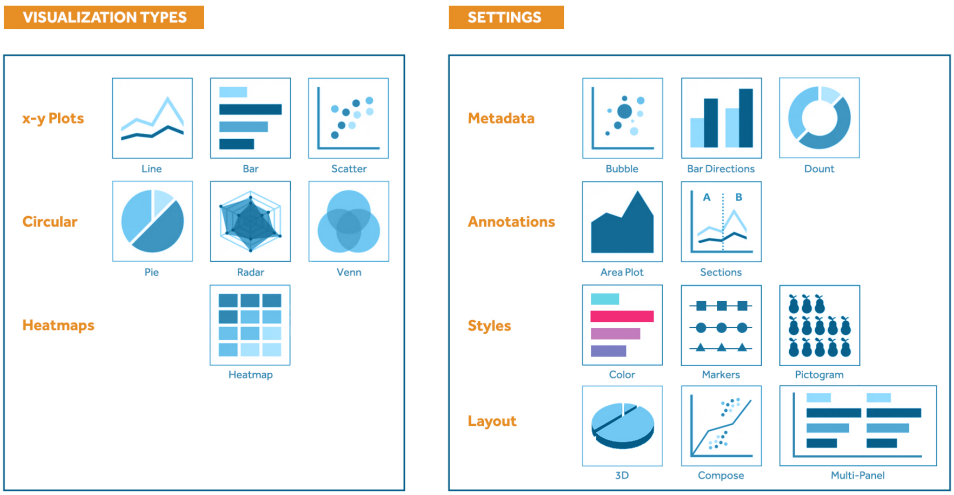


Figure 4: Some of the visualizations worked on in this thesis are categorized by type on the left. The visual settings on the right highlight how the dimensionality and complexity of visualizations increase with different configurations that users may incorporate, including metadata, annotations, styles, and various layouts.

As shown in Figure 4, charts can be categorized into different types based on how the data is presented and the visual settings authors may include, such as bars oriented in various directions, additional plot area annotations, and styling of the data. However, most chart analysis methods have been demonstrated and tested on collections of digital images. Furthermore, it is notable that the majority of methods used for chart understanding, as discussed in Section 2.1.2, rely heavily on supervised learning. Consequently, the amount of labeled data has consistently been a crucial and complex issue in this field. There are two main problems related to the collection and annotation of charts: (1) charts can be considered a form of artwork, resulting in highly diverse structures; thus, many benchmarks either synthesize predefined templates or collect more structured samples from the web and scientific articles; (2) not all available data comes with structured information for automatic annotation, necessitating a choice between manually inspecting a small amount of data with expert assistance or excluding a vast amount of unlabeled charts. These challenges significantly impact the robustness and generalization of proposed model frameworks due to the lack of variability in available benchmarks.

Synthesizing Data. A solution to bridge the gap between expensive annotation procedures and large automatically labeled collections is the generation of synthetic data, which inherently includes annotations (e.g., [202]). While this approach opens new possibilities, it is challenging to generate data that faithfully replicates real data. Although several benchmark datasets have emerged with this concept, and VL models have shown improvements in overall performance [70, 186] by learning fine representations, the use of synthetically generated data raises concerns about model robustness and biases towards certain visualization styles [11, 23].

Downstream Benchmarks. Chart Question-Answering (Chart-QA) addresses queries about charts with some datasets PlotQA [126] and ChartQA [121] focusing on visual and arithmetic reasoning, while others focus on open-ended explanatory question answering (OpenCQA) [91]. Additionally, Chart-to-Text involves creating natural language summaries from charts [143, 203], and Chart-to-Table focuses on converting charts into data tables [36].

Accessibility Benchmarks. There have been very few chart datasets that address accessibility concerns. Although they follow the same downstream tasks as detailed previously, their textual outputs are further curated and processed according to accessibility standards. Notably, HCI Alt Text [35] and VisText [173] have been developed specifically to address chart summarization for PVI. Both datasets are rich in textual semantics. VisText creates synthetic chart images using the Vega-Lite visualization tool, then utilizes crowdsourcing to generate summaries at different levels. In contrast, HCI Alt Text compiles figures from accessibility venues, filtering for those with alternative text. However, this dataset, intended primarily for analysis, comprises only 511 chart images, making it challenging to train effective data-driven methods. To overcome these constraints, we present later our AltChart dataset, which follows a similar methodology to HCI Alt Text but expands the collection to 10,000 chart images, manually annotated with 10 text semantics.

The **chart-to-tactile** task involves reconstructing charts as tactile materials for PVI individuals, requiring not only chart comprehension but also the **optimization** of extracted metadata to balance cognitive load (*amount of data presented*) and tactile resolution (*what to eliminate*). We are the first to address this issue and have proposed the ChartFormer dataset & model to facilitate this task.

2.3 CHART DIGITIZATION SYSTEMS FOR ACCESSIBILITY

Our review in this section builds upon the fundamental elements of earlier studies, which include: (1) accessible charts, (2) chart deconstruction, and (3) chart summarization.

Image to Vector Graphic Conversion. Having determined SVG as a suitable format for an alternative representation, we now describe techniques for converting bitmap images to vector graphics. Tools such as LibreOffice Draw can be used to convert images to a LibreOffice vector graphic format [58], but this method is time-consuming [150]. To address these challenges, Jayant et al. introduced an automated solution that converts bitmap images into Adobe Illustrator vector graphics [87]. However, this tool requires manual training on similar charts and directly translates charts into printable graphics, which is not ideal for tactile charts due to braille embosser constraints [50]. Several guidelines have been introduced to govern the conversion process [60, 151]. Goncu et al. proposed a tool that converts pie and bar chart data into SVG [65], arguing against a one-size-fits-all approach for accessible chart representation. Offering multiple output options can diminish bar-

riers between BVIP and sighted users [65], making the original and tactile charts more alike, potentially facilitating BVI's interpretation of chart data [51].

User Experience Analysis. Usability is vital when designing interactive user interfaces [178]. Recent applications emphasize intuitive design. Tools such as ChartDetective [123] and PlotDigitizer [81] are web-apps that adhere to design best practices, incorporating Visual Information-Seeking Mantra principles [168]. However, the former tool, ChartDetective, accepts only vector graphic charts in PDF format, and neither tool supports accessible output formats.

Enhancing UX also involves offering a magnified view around the mouse pointer, improving application accuracy. While tools like im2graph [179] and WebPlotDigitizer [160] emphasize such explorations.

From 2006 to 2023, across various chart analysis tools [81, 177], manual calibration of chart axes, involving setting four calibration points to map pixel values to the x-y plane, has remained a consistent feature. Despite the potential benefits of automation, its implementation is limited by current algorithms, which only achieve 61.7% accuracy in axis detection [134]. Thus, manual methods, either by prompt-based clicking [160, 177] or drag-and-drop [81], are preferable. Semi-automation also aids in text value entry. Im2Graph [179] and ChartDetective [123] employ OCR to recognize image text [26, 138]. After calibration, various techniques exist for extracting plot data, often resulting in CSV files [81, 89, 123, 160, 177, 179]. WebPlotDigitzer and PlotDigitizer offer semi-automatic extraction, letting users edit detected data markers. Im2graph employs a color-based line detection approach. Data markers, which can mimic curves and match exported values, are commonly used due to their predictable results.

3

ACCESSIBLE DIGITIZATION OF DOCUMENTS AND VISUALS

What sets tactile materials apart is their presentation of layout information. Layouts define how different elements in an image are organized. These different arrangements create various types of data visualizations and documents. For example, a newspaper layout facilitates the quick scanning of multiple articles to find an interesting story, while a presentation slide layout emphasizes visual impact. The spatial structures can significantly enhance educational activities beyond simple reading, such as skimming documents, memorizing information, and comparing texts. Therefore, digitizing the layout is a crucial step towards making visuals and documents accessible for PVI individuals.

This chapter is based on the publications [134] (ICDAR 2023), [131] (PETRA 2023) and [132] (IUI 2024).

3.1 INTRODUCTION

Due to the visual and spatial nature of document layout, PVI individuals rely on assistive technologies such as screen magnifiers, screen readers, and Braille displays to engage with documents and visuals. Screen readers and one-line Braille displays render content solely through audio and audio-Braille feedback, lacking information about the spatial arrangement and order. Consequently, all elements become serialized, stripping away the visual context and leading to a substantial loss of semantics [149].

Tactile materials provide a more spatially aware alternative by allowing PVI individuals to physically interact with the layout of documents. These materials present visual elements as raised textures that can be felt, offering a better understanding of spatial relationships and structures.

While printed tactile materials offer an affordable alternative, they often require a sighted person to extract and reconstruct this information for PVI individuals. Beyond the labor involved, converting a presentation deck or a research paper for PVI individuals is time-consuming and demands expertise. As a result, readers with visual impairments can access layout contents, but only in a linearized form. This means that layout information, such as a document's design or structural arrangement, is mostly absent. The challenge remains in how this layout information can first be digitized and then provided in an equally effective way to them.

This chapter investigates the potential of deep learning models to assist in projecting layout information in accessible modalities. We approach this investigation in two paths: first, we address how AI output formats such as bounding boxes

and segmentation masks can be reformed for tactile accessible content. As a result, various document layout analysis models are trained, and an approach has been developed to present these outputs on a 2D refreshable tactile display. The second path focuses on assisting sighted individuals in digitizing more complex inter-related visual content such as line charts. Based on preliminary interviews and user studies, an intelligent interface was developed to guide non-experts in the process of tactile material creation.

Beginning with addressing the accessibility of document layouts in section 3.2, we specifically address the accessibility of bounding box modalities for PVI individuals. This is the first step for them to locate elements such as figures and captions. This is followed by the accessibility of visuals in section 3.3, which requires more fine-grained details involving the use of segmentation masks and the development of metadata tracing concepts. Finally, in section 3.4, we combine both contributions to prototype an intelligent interface that replaces the conventional, long-time-consuming process of tactile material creation.

3.2 ACCESSIBLE DOCUMENT LAYOUTS

In this section, we present our new tactile layout reader, designed to empower independent access to documents and enhance document navigation through pinpointed audio-tactile explanations. Utilizing a state-of-the-art object detection-driven tactile interface, it generates a high-level abstraction of the document structure. This optimized interface is suitable for both 2D refreshable displays and Braille-embossed documents, offering both audio and tactile representations.

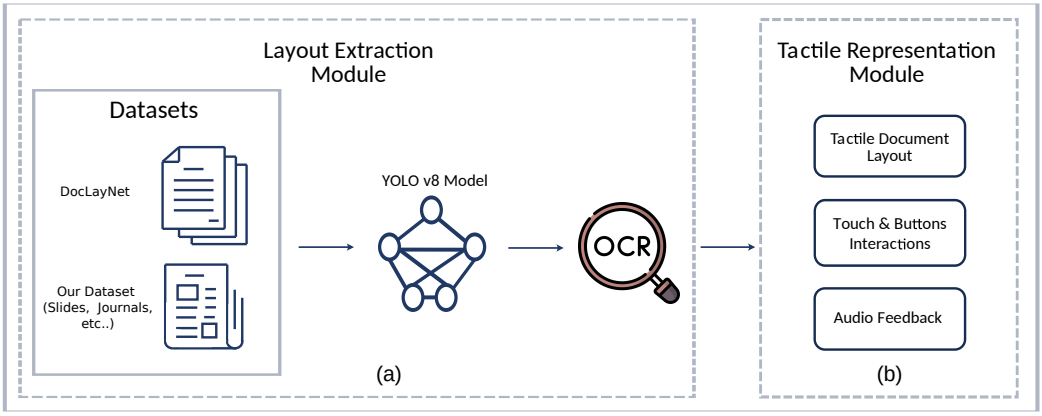


Figure 5: The pipeline of the tactile document interface consists of: (a) the layout extraction module, which utilizes the YOLOv8 [174] detection model and an OCR model to extract metadata from each predicted bounding box; and (b) the tactile representation module, responsible for converting document metadata into a tactile format. This module handles touch and button interactions and provides audio feedback for the auditory representation of text elements.

3.2.1 *Design of Layout4Blind System*

To meet this need, we developed the *Layout4Blind* system. The system comprises two primary modules, as depicted in Figure 5. The first module involves the layout extraction process, utilizing a trained detection model to automatically extract metadata from documents. This metadata includes spatial layout information and textual data. The extracted metadata is stored in *JSON* format, which is then used by the second module—the tactile user interface system. This module presents the extracted metadata on 2D refreshable tactile displays and handles user interactions.

Dataset. To extract the layout effectively, we needed to train models on a comprehensive document layout dataset. Initially, we utilized the DocLayNet dataset [148], which is a substantial collection of real documents encompassing various fields and languages. However, this dataset alone did not sufficiently cover the extensive range of document structures we aimed to address. To enhance the model's robustness against diverse document layouts, we augmented our dataset, *Layout4Blind*, with a small batch of documents featuring more complex layouts, such as multi-column newspapers, magazines, and slides. Table 2 provides the statistics of our dataset. The dataset is categorizes into three main types: Artistic, Educational, and Multi-column, with each type containing specific subcategories. Each category includes an equal count of 100 documents, ensuring a balanced representation across different document structures. Figure 6 illustrates a few samples. The dataset is split into training, validation, and test sets with a ratio of 80:10:10.

Document Type	Subcategories	Image Count	Bounding Box Count
Artistic	Flyer, Poster, Infographic, Brochure	100 images/type	500
Educational	Slides		400
Multi-column	Books, Magazines, Newspapers		600

Table 2: Types of documents, their respective image counts, and bounding box counts in the enhanced dataset

Layout Extraction. To extract the spatial layout information from documents, we utilized object detection models. The object detection task involves identifying and locating objects within an image by drawing bounding boxes around them. This process is crucial for understanding the structure and organization of document elements, such as text blocks, images, and tables. We chose to use models from the YOLO (You Only Look Once) family [174] for our experiments. The YOLO models are well-suited for this task due to their fast inference time on edge devices, which is beneficial for assistive technology applications. They are single-shot detectors, meaning they detect objects in a single pass through the neural network, making them much faster compared to other object detection models that require multiple passes or stages. We conducted multiple experiments with variations of YOLO models to achieve optimal performance in detecting and classifying the various elements within document layouts.

In this experiment, we considered three families of YOLO models—YOLOv3, YOLOv8, and YOLOv10—using the implementation from Ultralytics[1], with multiple variations of each. All models were trained to identify 12 different data categories, as detailed in the first column of Table 4. These categories are further grouped into five meta-classes: title, text, table, mathematical content, and images.

Model	Variation	mAP50	mAP50:95
YOLOv3	Large	82.5	50.6
YOLOv8	Tiny	83.1	51.4
	Large	84.7	53.1
YOLOv10	Tiny	89.2	68.5
	Small	**90.5**	**71.3**

Table 3: Performance of different YOLO model variations in terms of mAP50 and mAP50:95

YOLO-v10 (small) demonstrated superior performance with a mAP50 of 0.905 and 7.2 million parameters, in contrast to YOLOv8 (Large) with a mAP50 of 84.7 and 43.7 million parameters. One can choose according to the trade-off between performance and inference time. During the inference phase of our model, we collected the predicted bounding boxes and confidence scores, then stored the high-confidence predictions (e.g., greater than 0.5) in JSON format.

(a) (b) (c)

Figure 6: Samples from Layout4Blind dataset illustrating its diversity: (a) a multi-column journal article, (b) two magazine pages, and (c) slides from different lectures. These images highlight the variety of document layouts present in our dataset.

Model Training & Evaluation. For this experiment, we trained two YOLOv8 variations: YOLOv8s, which consists of 7.6 million parameters, and YOLOv8x, with 88.8 million parameters. We pre-trained the models with DocLayNet [148] for 30 epochs, followed by fine-tuning with 2 epochs on our custom dataset. YOLOv8s achieved a mean Average Precision (mAP) of 72.5, while YOLOv8x showed superior performance with a mAP of 78.6. For our user study, we utilized the larger model, YOLOv8x.

1 https://github.com/ultralytics/

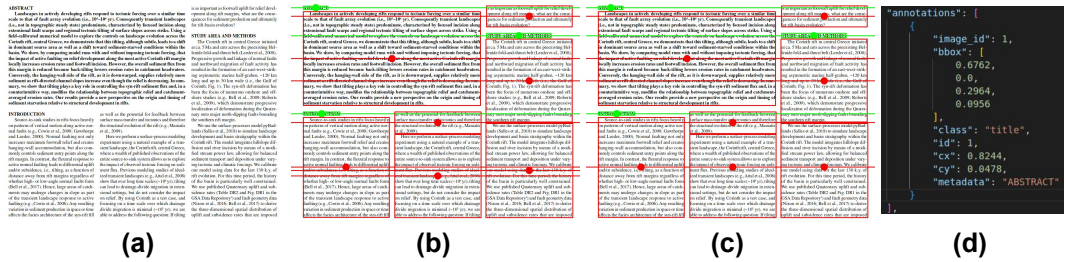


Figure 7: Layout retrieval output: (a) Multi-column document image, (b-c) predicted bounding-box with and without shifting, (d) sample JSON file for a single element.

Bounding Boxes Refinement. Since the bounding box format poses high sparsity for tactile experiences, directly mapping it to tactile modality causes problems for PVI individuals in differentiating between separate instances. That is why, upon the retrieval of bounding boxes for the document layout, we perform post-processing steps to improve the usability of the predicted results. Instead of presenting the four edges of a bounding box, we calculate the center point and assign a tactile letter indicating the object class to it. Although computing the centre of each bounding box provides a more compact format for sensing the layout, it still poses a challenge for multi-column elements. These elements may have slight centre point shifts while they are at the same level, making it difficult to classify and discern their position. Similarly, in the vertical direction, list items, for example, could be better understood if all centre points are well-positioned. This applies to sighted individuals as well, as they tend to align content in a grid-like format. To address these challenges, we propose an alignment algorithm to further post-process the centre points in both the horizontal and vertical directions as depicted in Figure 7 (c), ensuring that the layout is accurately depicted without causing any deformation from the original layout. This ensures that the information is presented in a clear and organized manner, making it easier for users with PVI to interact with the layout.

The output result of the object detection is then stored as a JSON intermediate representation. We followed the COCO dataset [105] format to store the bounding boxes, centre points, and class categories in a normalized format relative to the image resolution for better scalability for different tactile displays. This ensures that the layout is accurately depicted on the tactile display, regardless of the resolution or the zoom level. In addition to localization information, we incorporate audio metadata as text as shown in Figure 7 (d). This is done for text elements such as figure and table captions, titles and paragraphs in digital documents. If the metadata is not present, we use EasyOCR [85] to populate it.

3.2.2 *Tactile Interface*

Our tactile document interface design was guided by the Visual Information Seeking Mantra (VISM) proposed by Shneiderman [167], which emphasizes an approach to presenting data that is most effective for users. The Mantra, summarized as 'Overview first, zoom and filter, then details-on-demand,' provides a

Table 4: Supported classes and their tactile representation.

Detected Sub Classes	Mapped Main Classes	Tactile Representation
Document title, section title, header	Title (t)	⠒
Paragraph, footer, caption, page number, list-item	Text (x)	⠶
Table	Table (b)	⠐
Equation, code	Math (m)	⠿
Figure	Image (i)	⠤

framework for developing user interfaces by describing how information should be visually structured to facilitate exploration and understanding. We followed a user-centered design process, by collaborating with a blind user, to create a tactile document interface that implements the three main concepts defined in the VISM. As shown in Figure 8. The initial component is the "Element Guide", which forms a vertical rectangular region where we draw horizontal lines. These lines indicate the existence of a document element, such as text, images, etc., at that specific level. These lines align with the y-values of the bounding box centers that hold each element. The second component is the "Class Identifier" area, where a Braille character is placed at the center of the bounding box. This character represents the layout class, for example, "x" for a text element. The last part is the "Divider Line," which separates the previously mentioned sections, aiding users in distinguishing between the first two components.

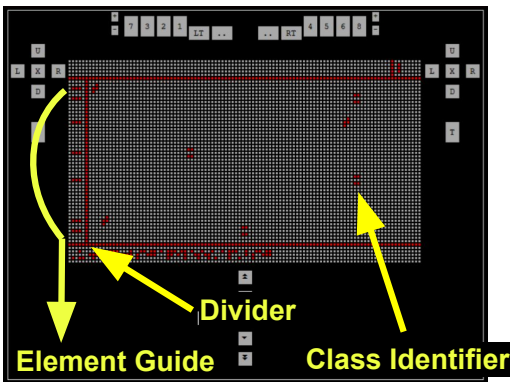


Figure 8: The interface designed for the 2D refreshable tactile display.

Interactions. The interface is designed with fundamental interactions to facilitate document exploration and the transition between the different available views for PVI. The zoom and filter concepts are defined in the VISM as methods for reducing the complexity of the data representation by removing unnecessary information from view and allowing the user to explore certain information in more detail. In our interface, this was achieved by the control buttons. The Braille translation will be displayed on the screen if the selected element is a text. If the element is an image, the caption will be presented. In cases where no caption is available for the

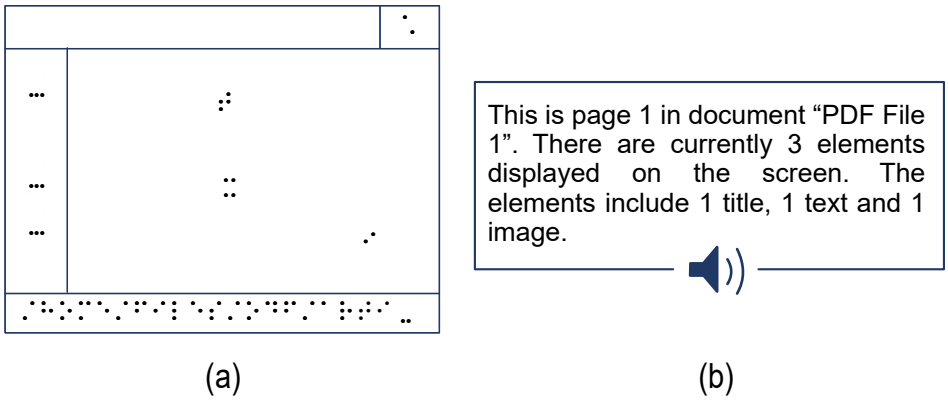


Figure 9: View of the tactile interface interactions.

image, a standard text message is displayed informing the user that this is an image without a caption. The principle of "details-on-demand" of VISM is realized by allowing the user to select a particular element in the document. while in overview mode and receive short acoustic information about it without requiring the user to switch to another view. The acoustic information consists of the initial sentence within the selected element, offering users a concise overview of the element's content.

3.2.3 *User Study*

An exploratory study was conducted with two blind users (both male, P1 and P2) to evaluate our final prototype of the interface. Our participants had prior experience using 2D tactile displays and screen reader software.

For our user study, we used three pages as materials: two pages from a newspaper with multiple columns (Figure 10-a and 10-c) and one page from a lecture slide (Figure 10-b), each presenting a challenging layout for screen readers.

The interface was tested on the Metec Hyperbraille 2D tactile display [125], motivated by its capability to enable an interactive experience and the exploration of diverse modalities. The device features 60 by 104 actuators and a resolution of 10 dpi. In the following section, we discuss the implementation of the two components in detail.

The study consisted of four sessions. The first session was an introductory session where each participant was provided with a training document Figure 10 (a) containing step-by-step instructions for 15–20 minutes on how to use the interface. In the second session, the participants used the new interface to read two documents 10 (b) and (c) on the tactile display. In the third session, they used the NVDA [136] screen reader on a computer to read a different document. Following these sessions, participants performed the following tasks:

- Task 1: Skim the document and give a quick summary of the topic and the main key points.
- Task 2: Answer a specific question about the document.
- Task 3: Explain the structure of the document.

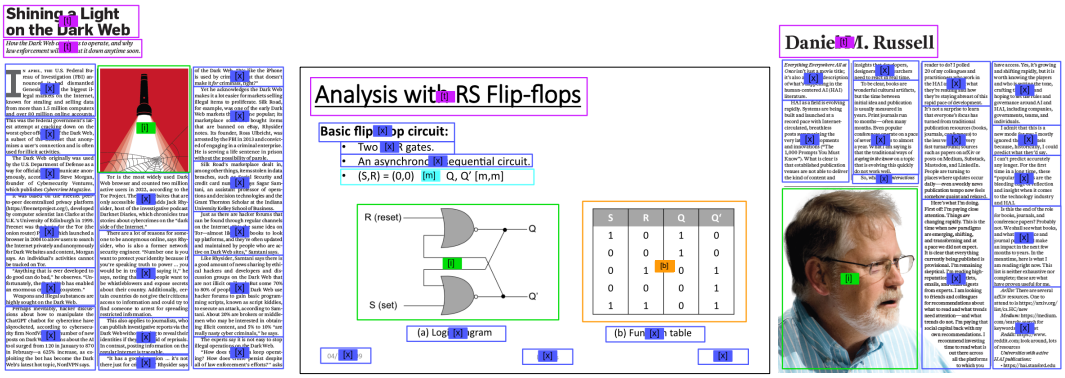


Figure 10: User study samples with the alligned bounding boxes. (a) & (c) are from 3 & 6 columns news paper. (b) is from a lecture slide dick.

3.2.4 *Results & Discussion*

Completion Time. For both tasks 1 and 2, participants took nearly four times longer to complete the tasks using the screen reader compared to using our interface, which averaged 1.8 minutes per task. This significant reduction in completion time highlights the efficiency of presenting the layout for a quick document access.

Task Completion Rate. Both participants successfully skimmed the document and extracted the main ideas for task 1 and correctly answered the questions for task 2. However, this success came at the cost of increased time when using the screen reader. For task 3, participants were unable to accurately describe the document's structure when using the screen reader. They resorted to guessing based on the sequential reading order, lacking information about the horizontal order of elements. This finding underscores the importance of spatial layout information. Upon completion of the study, participants were asked to evaluate their experience with document skimming through the *Layout4Blind* system. Both users strongly agreed that the interface made skimming documents straightforward and intuitive. They did not find the navigation interactions complex, nor did they feel that an extended period was required to learn how to use the system effectively. P1 noted, "The interface allowed me to quickly decide whether the document was useful for me."

However, participants also identified areas where the system could be improved. One participant observed inaccuracies in the predicted classes, such as a title being mistakenly classified as a paragraph. P2 remarked, "The system made it easier to grasp the overall topic and structure, but there were times when I had to rely on guessing because of the misclassified areas." Additionally, the participants suggested enhancements to improve the system's usability. One recommendation was to introduce a button that would allow users to control which classes to focus on, such as keeping only figures or titles visible. They also suggested adding an audio description at the beginning of the interaction to provide an overview of the document layout, including details such as the number of columns, the types of elements present, and the overall type of document (e.g., slide, receipt).

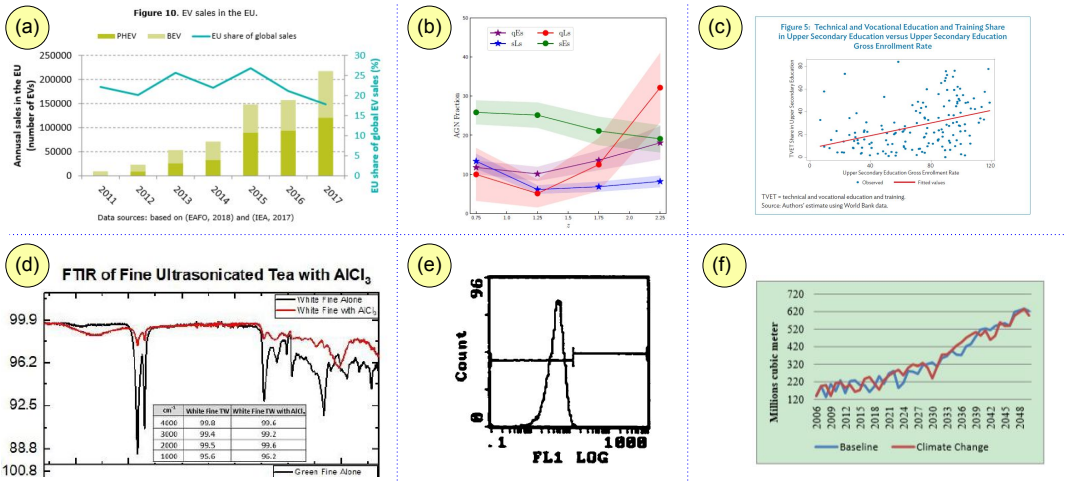


Figure 11: Diverse mathematical graphics covered in our Line Graphics (LG) dataset, including 100 bar charts (a), 320 line graphics (b, d-f) and 100 scatter plots (c). These samples pose significant challenges for existing document analysis methods.

3.3 ACCESSIBLE VISUALIZATION LAYOUTS

With the rapid growth of available data visualizations, there is a need for improved digitization techniques that allow for wider accessibility and reproducibility. While automatic digitization of document layouts and text content has been a long-standing focus of research, the digitization of graphical elements, such as statistical plots, has been underexplored. In this section, we address the task of fine-grained visual understanding of mathematical graphics, presenting a new benchmark and a methodology to trace and extract metadata from line charts.

3.3.1 *Motivation for New Dataset*

During courses, graphs are a vital supplement to lecturers' speech as they effectively summarize complex data or visualize mathematical functions. However, one downside of this medium is the difficulty of automatic information extraction, as graphs contain very fine-grained elements, such as fine lines, small numbers or axes descriptions, while the traditional document analysis frameworks focus on coarse structures within complete pages [27, 37, 193] or slides [72, 73]. The process of separating distinct regions of a plot and assigning them a semantic meaning at a pixel-level, known as graph segmentation, is an important prerequisite step for graph understanding. One application of using pixel-level data to fully automate the process is to generate an imposed document or 2D refreshable tactile display that can be easily interpreted through touch for people with blindness or visual impairment. Hence, end-to-end full automation of plot digitization could be achieved.

Presumably, due to the lack of annotated datasets for fine-grained analysis of plots, the utilization of modern deep semantic segmentation architectures has been rather overlooked in the context of mathematical graphs.

In this section, we introduce the task of fine-grained visual understanding of mathematical graphics and present the Line Graphics (LG) dataset, which includes pixel-wise annotations of 10 different categories. Our dataset covers 520 images of mathematical graphics collected from 450 documents from different disciplines, such as physics, economics, and engineering. Figure 11 provides several examples of statistical plots collected in our dataset. By providing pixel-wise and bounding box annotations, we enable our dataset to support two different computer vision tasks: *instance, semantic segmentation* and *object detection*.

The key findings and contributions of this section can be summarized as:

- We introduce the task of fine-grained visual understanding of mathematical graphics, aimed at reducing manual user input when digitalizing documents.
- We collect the Line Graphics (LG) dataset as a benchmark for semantic segmentation and object detection in line graphics. As well as perform extensive evaluations on 7 state-of-the-art semantic segmentation models, analyzing the impact of factors such as image resolution and category types on the performance.
- Propose a methodology for tracing line plot data using instance semantic segmentation.

LG Dataset. Our dataset contains 520 mathematical graphics manually extracted from 450 documents, comprising a total of 7238 human-annotated instances. The goal is to facilitate the automatic visual understanding of mathematical charts by offering a suitable and challenging benchmark. To ensure that each image presents a unique and challenging point for analysis, the similarity levels between cropped images were kept as low as possible. Documents in the LG dataset were collected from five different disciplines and their top published subcategories to achieve broad coverage, as shown in Figure 12. The collection process involved a manual search using scientific keywords and careful inspection of each document downloaded from sources such as arXiv and Google Scholar, ensuring a consistent and uniform distribution of documents across all categories.

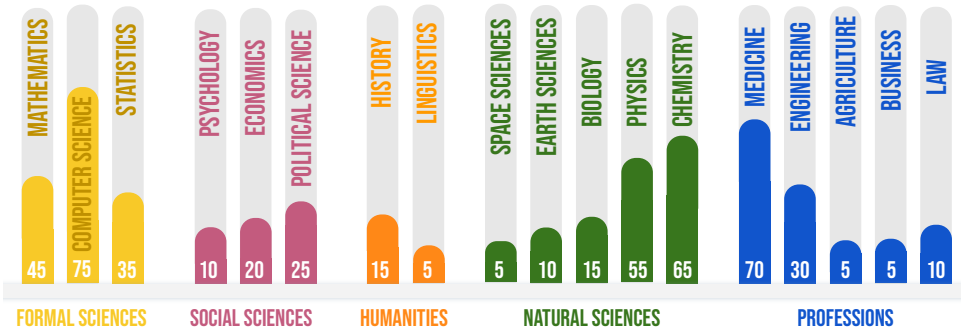


Figure 12: Statistical distribution of documents in our dataset grouped by different disciplines. Our dataset was collected from 18 distinct disciplines from formal-, social and natural sciences as well as humanities and professions.

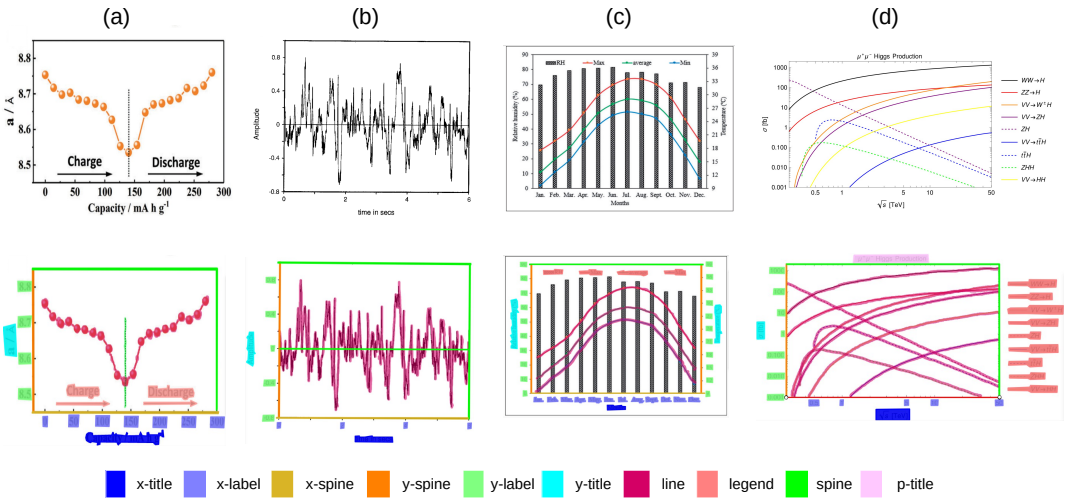


Figure 13: Example annotations of our Line Graphics (LG) dataset. From top to bottom are the challenging line graphics and the ground truth with fine-grained annotations of 10 classes, which are complemented by 5 coarse categories.

To ensure a comprehensive and robust labeling process, we categorized line chart pixels into five coarse and ten fine-grained classes. The primary focus was on creating fine-grained categories that offer a wide range of variations and challenges for further analysis. This was achieved through a thorough review of charts by three annotators with research experience, who identified the most frequent and critical object types encountered in such charts. Based on this review and an inspection of related work, we arrived at ten relevant categories. These categories can be further grouped into three coarse categories: the Title class (e.g., plot title), the Spine class (e.g., "spine" with no label data), and the Label class (e.g., x-axis labels).

Dataset Properties. To facilitate instance-level segmentation, we provide annotations for each instance separately in COCO JSON format [105]. For example, each line has a unique ID in the line mask, which will become handy in later steps like line tracing. Our dataset includes a wide range of instance counts, styles, and locations, without any aforementioned limitations, offering a comprehensive range of variations for all classes, see Figure 13. We have covered a wide array of plot types, including those that feature multiple chart types like bar, scatter, and line charts as well as plots with repeated classes like multiple y or x-axes and ticks. The text content in our dataset is annotated with variations in integer, decimal, and DateTime formats, as well as tilt. Furthermore, we have taken into account different markers, patterns, and sizes for line and spine classes, and added the class "other" to represent the annotated plot area explanatory text, focus points, and arrows. The background variations in our dataset include color (single or multiple), gradient, and RGB images, ensuring a comprehensive representation of real-world scenarios.

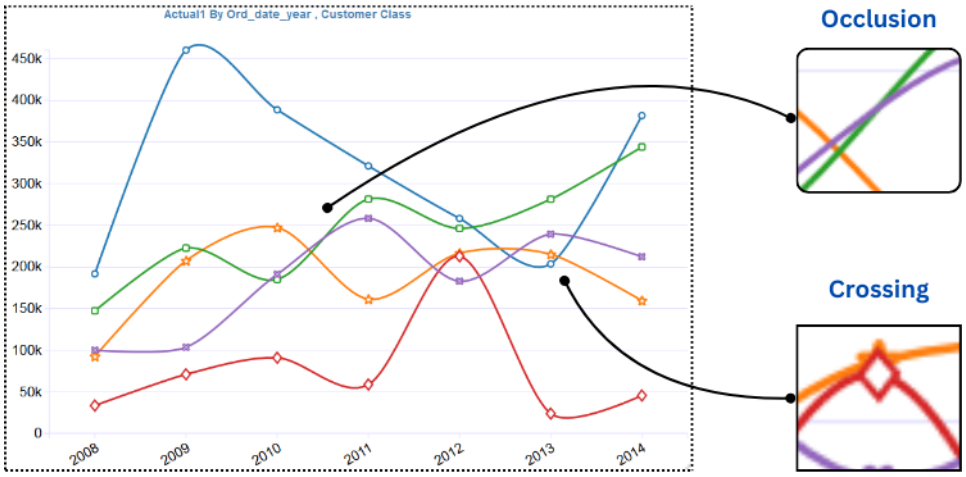


Figure 14: multi-line charts structural complexities.

3.3.2 *Line Tracing*

Line charts, particularly those with multiple lines, present several structural complexities. We identify two primary structural patterns that cause issues: crossings and occlusions. Figure 14 illustrates these examples. In these cases, the visual attributes of the line segments, such as color, style, and markers, are often obscured or blended with adjacent lines. As a result, methods that rely heavily on low-level image features like color, gradients, and texture often fail. Additionally, most keypoint-based line extractors face two significant challenges: a) the inability to predict distinct keypoints for each line at crossings and occlusions, and b) difficulties in the subsequent keypoint grouping step, which struggles to extract features from an already occluded local image patch. Recent approaches have tried to address occlusion by incorporating explicit optimization constraints. However, these methods still rely on low-level features and proximity heuristics, which may limit their robustness. It's important to note that humans can intuitively gather contextual information from surrounding areas to fill in gaps in such cases.

To tackle the challenges of occlusion and crowding, we employ an instance segmentation model based on an encoder-decoder transformer architecture, utilizing a masked-attention pixel decoder that provides the visual context necessary to identify occluded lines. We adopt the architecture and hyperparameter settings from existing works, using SwinTransformer-tiny [116] as the backbone to balance accuracy and inference speed. To ensure a manageable number of line predictions, we limit the number of line queries to 100. The transformer decoder is trained end-to-end with a set prediction objective, where the loss function is a linear combination of classification loss and mask prediction loss, weighted at 1 and 5, respectively. The mask prediction loss itself is a combination of dice loss and cross-entropy loss. All experiments are conducted using the MMDetection framework [28] based on PyTorch.

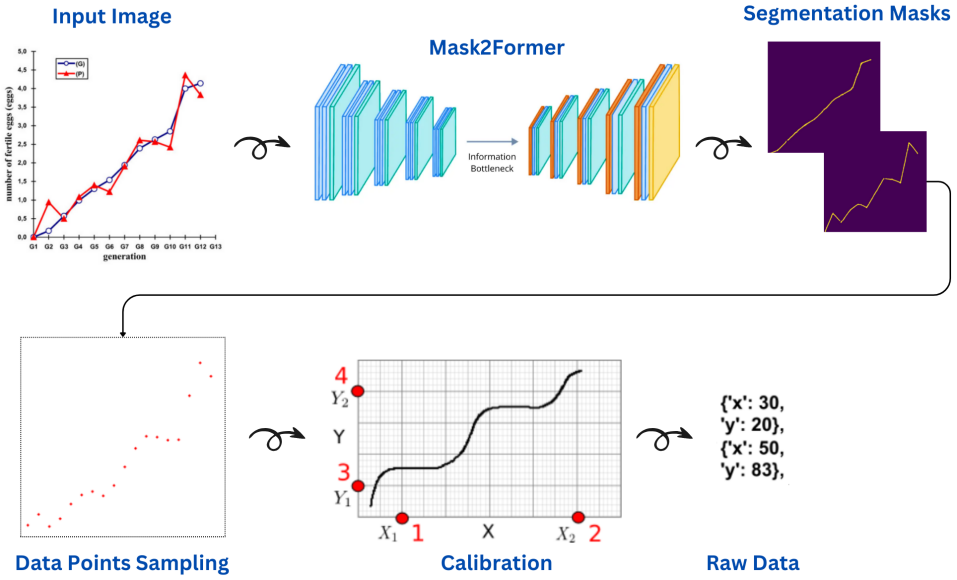


Figure 15: Line tracing system.

From Pixel Domain to Raw Data Domain. Once the predicted line masks for an input image are obtained, we sample foreground points at regular intervals, see Figure 15. Linear interpolation is applied to the initial sampled points to address any gaps or breaks in the predicted line masks. These extracted points in the pixel domain can be scaled to obtain the corresponding raw data values, provided that at least two calibration points are defined on each axis. In the next section, we will manually set sampling rates, data formats, and calibration points from the intelligent interface.

3.3.3 *Experiments*

The models were trained on an A40 GPU with an input resolution of (2048, 1024). Our evaluation metric is Mean Intersection over Union (mIoU). During training, we applied common data augmentation techniques such as random flipping, scaling (ranging from 0.5 to 2), and cropping. The batch size was set to 8 with an initial learning rate of 6e−5, using a poly-learning rate decay policy. The models were trained for 50K iterations. For testing, we employed a single resolution scale to ensure fairness in comparison. To understand the choice of models, we further analyze the properties of the selected models in conjunction with our proposed line graphic segmentation task.

Line Chart Segmentation Experiments. We consider 7 state-of-the-art semantic segmentation models for this task as showin in Table 5. The efficiency-oriented CNN model MobileNetV3 [77] with only 1.14M parameters obtains 56.22% mIoU score on the proposed LG dataset. The high-resolution model HRNet [183] has 57.60% in mIoU and the DeepLabv3+ [29] model has 61.64%, but both have parameter >60M. We found that the PSPNet [199] with pyramid pooling module in the architectural design can achieve better results with 62.04%. In Table 5, the recent Transformer-based models achieve relatively better results than the CNN-based

models. For example, the SegFormer [191] model with pyramid architecture and with 81.97M parameters can obtain 65.59% in mIoU with a +3.55% gain compared to PSPNet. The Swin Transformer [116] with hierarchical design and shifted windows has 66.61% in mIoU, but it has the highest number of parameters. However, the state-of-the-art CNN-based SegNeXt [67] utilizes multi-scale convolutional attention to evoke spatial attention, leading to the highest mIoU score of 67.56% in our LG dataset. Furthermore, the SegNeXt [67] achieve 4 top scores on 5 coarse classes, including *Title, Spine, Label* and *Legend*. Besides, it obtains 6 top scores out of 10 fine classes, which are *xtitle, ytitle, xpsine, yspine, xlabel,* and *legend*. The results show that a stronger architecture for the semantic segmentation task can achieve better results in the proposed LG benchmark, yielding reliable and accessible mathematical graphics.

Table 5: Semantic segmentation results of CNN- and Transformer-based models on the *test* set of LG dataset. **#P**: the number of model parameters in millions; **GFLOPs**: the model complexity calculated in the same image resolution of 512×512; **Per-class IoU** (%): the Intersection over Union (IoU) score for each of coarse and fine classes; **mIoU** (%): the average score across all of 10 fine classes.

Model	**Backbone**	**#P(M)**	**GFLOPs**	**Coarse Per-class IoU**					**mIoU**
				Title	*Spine*	*Label*	*Legend*	*Line*	
MobileNetV3 [77]	MobileNetV3-D8	1.14	4.20	45.06	43.68	68.74	60.86	62.12	56.22
HRNet [183]	HRNet-W48	65.86	93.59	52.48	44.4	67.95	53.34	61.91	57.60
DeepLabv3+ [29]	ResNet-50	62.58	79.15	55.41	46.14	74.72	67.07	32.97	61.46
PSPNet [199]	ResNetV1c	144.07	393.90	57.12	43.77	78.52	67.75	62.30	62.04
SegFormer [191]	MiT-B5	81.97	51.90	58.36	54.13	76.79	67.09	69.67	65.59
Swin [116]	Swin-L	233.65	403.78	62.26	52.85	76.57	68.91	**71.01**	66.61
SegNeXt [67]	MSCAN-L	49.00	570.0	**63.79**	**54.61**	**80.29**	**69.09**	65.07	**67.56**

Model	**Fine Per-class IoU**									
	ptitle	*xtitle*	*ytitle*	*xspine*	*yspine*	*spine*	*xlabel*	*ylabel*	*legend*	*line*
MobileNetV3 [77]	09.03	55.36	70.81	53.47	40.21	37.36	67.83	69.65	60.86	62.12
HRNet [183]	31.30	55.85	70.30	44.68	50.20	38.36	65.42	70.48	53.34	61.91
DeepLabv3+ [29]	30.30	62.21	73.74	49.47	47.57	41.39	73.65	75.79	67.07	32.97
PSPNet [199]	22.92	68.09	80.39	47.09	53.11	31.12	77.28	79.76	67.75	62.30
SegFormer [191]	37.25	61.10	76.75	60.37	55.83	**46.21**	73.35	80.23	67.09	69.67
Swin [116]	**49.21**	59.44	78.15	59.48	55.33	43.74	71.96	**81.54**	68.91	**71.01**
SegNeXt [67]	36.57	**73.95**	**80.85**	**61.77**	**56.20**	45.88	**80.68**	79.91	**69.09**	65.07

Line Tracing Evaluations. Formally, To compare a predicted data series with ground truth, we calculate a similarity score by aggregating the absolute differences between each predicted value and its corresponding ground truth value. In our work, we adopted the same metric used by ChartOCR [118] and the CHART-Info challenge [2]. Specifically, we first compute pairwise similarity scores between each predicted and ground truth line using L2 distance. Then, we perform a bipartite assignment that maximizes the average pairwise score. This is similar to the mean average precision (mAP) calculation used to evaluate Object Detection performance.

2 https://chartinfo.github.io/tasks.html

Line Tracing Experiments. The performance of various systems for line data extraction, evaluated using the similarity metric defined earlier, is shown in Table 6. It can be observed that our LG dataset proves to be the most challenging of all, as it is diverse and composed of real charts from scientific journals. Line Tracing demonstrates state-of-the-art results on LG real and Adobe synthetic datasets. The performance difference illustrated in the table are in fact occur when the the number of lines and the complexity of the chart increases. The Line Tracing model performs significantly better, keeping track of the line even in occlusions and crossing points.

Table 6: Evaluations of various line tracing approaches on AdobeSynth [39] and LG datasets.

Model	Dataset (only line instances)	
-	AdobeSynth19 [39]	LG Dataset
ChartOCR [118]	84.67	55
LineEX [166]	82.52	81.97
Line Tracing (Ours)	**87.51**	**83.1**

3.4 INTELLIGENT INTERFACE FOR CHART ACCESSIBILITY

Converting charts into accessible formats requires considerable effort from sighted individuals. Digitizing charts with metadata extraction is just one aspect of the issue; transforming it into accessible modalities, such as tactile graphics, presents another difficulty. To address these disparities, we developed *Chart4Blind*, an intelligent user interface that converts bitmap image representations of line charts into universally accessible formats. Chart4Blind achieves this transformation by generating Scalable Vector Graphics (SVG), Comma-Separated Values (CSV), and alternative text exports, all comply with established accessibility standards. Through interviews and a formal user study, we demonstrate that even inexperienced sighted users can make charts accessible in an average of 4 minutes using *Chart4Blind*, achieving a System Usability Scale rating of 90%. In comparison to existing approaches, *Chart4Blind* provides a comprehensive solution, generating end-to-end accessible SVGs suitable for assistive technologies such as embossed prints (papers and laser cut), 2D tactile displays, and screen readers.

To address this goal, we started with a series of need-finding interviews to guide our design choices. The tool design emphasizes clear progress indicators during conversion, consistency in the conversion steps, intuitive data input methods, and integration of AI tools for efficient data extraction. To validate our tool, we conducted a usability study with 10 sighted participants aged 22-34 years. The study revealed that the *Chart4Blind* tool is valuable for converting line charts into accessible formats and helps sighted users understand accessibility guidelines with an average System Usability Scale rating of 90%. To validate the accessibility of the produced graphic for PVI in particular, we conducted a follow-up study with 3 blind individuals, asking open-ended questions regarding the output of our interface and collecting valuable feedback for future improvement.

This work introduces Chart4Blind, an intelligent user interface designed to convert bitmap image representations of line charts into universally accessible formats, and has the following major contributions:

- **Chart4Blind System:** Through a series of interviews and feedback, we designed an intuitive tool for converting bitmap line charts into multiple universally accessible formats. It also supports collaborative efforts where multiple people can work together and assist in the conversion process. *Chart4Blind* has a user-friendly interface and can be utilized by individuals with varying levels of experience.
- **Integration of Intelligent Features:** The tool incorporates intelligent deep learning models to ensure a seamless conversion process, particularly OCR and line tracing segmentation models. Users can interact with the model predictions through simplified actions, such as drag-and-drop.
- **Usability Study:** A thorough usability study involving sighted people aged between 22-34 years to validate the effectiveness of the *Chart4Blind* system. Furthermore, another user study was conducted specifically with blind participants to assess the output quality and accessibility. The tool achieved an average System Usability Scale rating of 90%.

3.4.1 *Prototype Development*

To gain a comprehensive understanding of how sighted individuals approach the creation of accessible line charts for PVI, and to identify how future tools can efficiently support their efforts, we conducted a series of exploratory need-finding semi-structured interviews. our need-finding interviews were carried out with the primary objective of uncovering these specific design requirements. Our inquiries primarily aimed to:

- Understand the limitations of common interfaces and user practices for converting charts into accessible formats.
- Pinpoint the features that users find important in such a tool.
- Gain insights into how AI tools are explored in prior research and could facilitate the creation of accessible charts.

We conducted a semi-structured interview with four sighted participants (P1-4, 2 female, and 2 male, age range 25-40 years) who would use the *Chart4Blind* user interface. All sighted individuals exhibited proficiency in working with line charts, familiarity with accessibility guidelines, and practical experience in converting chart materials.

During our need-finding study, we asked participants open-ended questions about their experiences, challenges, and requirements for the chart conversion process. We asked the following questions:

1. Describe your workflow for converting charts into an accessible format.
2. What challenges did you encounter during this process?
3. What specific computer-based tools do you utilize for the conversion process?

We followed up by inquiring about the steps they found most challenging and time-consuming. Depending on the interview, we additionally asked:

1. Could you estimate the time required for chart conversion?
2. Which features of your current tools do you find most useful?

Findings. Despite the diverse tools used, we identified common steps in the conversion process, highlighting an opportunity to streamline this task into a unified tool. However, participants encountered challenges involving inexperienced individuals for assistance, as the existing tools were not designed for this purpose. Extensive tutorials were necessary to prepare individuals for the task. Additionally, the process lacked the potential for parallelization due to limitations in (real-time) data-sharing mechanisms within the current tools. P1 emphasized how these challenges prevented effective crowdsourcing efforts. In response to these findings, we propose developing a user interface that caters to both experts and non-experts, presenting clear and consistent steps (UR-1). Furthermore, the tool facilitates a more parallelizable process to encourage concurrent contributions (UR-2).

Design Principles. Based on our interview findings, we have identified key design principles that guide our *Chart4Blind* system:

- **Clear Progress Indication:** The tool should provide users with a clear view of their current progress during chart conversion, indicating what has been completed and what remains.
- **Maintain Consistent Steps:** The tool should ensure that the steps in the conversion process remain consistent regardless of chart complexity, allowing for a more parallelized process.
- **Intuitive Data Input:** The tool should allow for an intuitive approach to make sure that all necessary information needed for an accessible chart is added. In addition, the tool should also support intuitive interactions such as drag-and-drop for textual content and drop-down lists for axis label formats (e.g., linear, logarithmic).
- **Automated Selection Tools:** To reduce the efforts spent extracting metadata, the tool should integrate an automated solution to ease the task of element selection (e.g., lines, texts). In our case, we choose AI-driven tools for line segmentation, text extraction, and chart description when needed.

- **Ready-to-Use Accessible Exports:** The tool should provide users with accessible output for various modalities such as tactile printing and digital displays used with a screen reader.

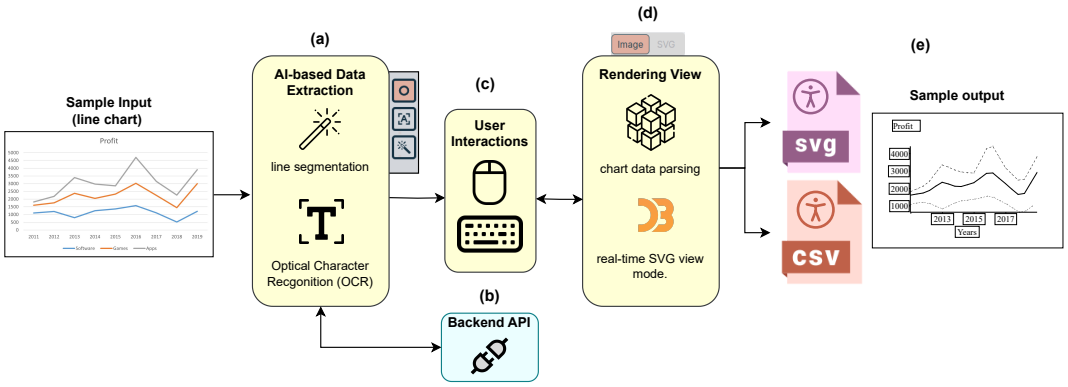


Figure 16: The pipeline of Chart4Blind consists of the input of a bitmap line chart, followed by the Data Extraction Module, which includes an AI-based line segmentation and optical character recognition step, and a manual correction step by a sighted user (a). The Rendering Module updates the information in real-time and ensures an accessible representation (c). The system allows the export of the information to an SVG and a CSV format. The SVG can be accessed with a screen reader or printed as a tactile graphic (d). The metadata can be exported as accessible CSV as well.

Chart4Blind System Flow. Figure 16 shows the pipeline of *Chart4Blind* system. When a user uploads a line chart image, a toolbar appears that is linked to our AI-based *Data Extraction Module*, which is responsible for extracting the chart's textual and visual data. After calibrating four axis points, this intelligent module converts pixel values from the RGB image domain into the original data points domain. For example, if the x-axis represents years, the module can predict the corresponding year for a given pixel point within the plot area. Subsequently, our SVG viewing mode, connected to the *Rendering Module*, generates the SVG view in real-time as users update field information. All session information is stored upon consent acceptance, and a unique Token is assigned to facilitate collaborative work. The final step includes the export of the visual data to a printable tactile graphic format (SVG) and to a textual description (CSV). The following paragraphs describe the two main modules in more detail.

OCR Feature. To categorize the textual content in the line chart, we followed the analysis outlined in [173] and [134]. The chosen categories are shown in Table 7. For each category, a corresponding text field has been added within the metadata tab in the interface. We utilize the Tesseract OCR system [170], operating on the user's browser even when offline. This particular model has demonstrated a good performance in the ICDAR chart text understanding challenge [194]. We employed a template-based approach [1] to auto-fill chart descriptions. This equipped the chart with a summary of the encoded elements and descriptive statistics (e.g., extremes, outliers, etc.), corresponding to Levels 1 and 2, as proven preferable by many BVI individuals [117].

Table 7: Text fields present in the Chart4Blind interface.

Property	Fields
Calibration	Axes calibration points
Chart Information	Plot title
	Axes titles
	Axes labels
Additional Information	Chart description
	Data point description

Rendering Module To address UR-1 and UR-5 findings, we incorporated a real-time view to track conversion progress. Users can switch between modes to visualize textual or line drawings. Our rendering module displays a real-time SVG view using D3.js and exports results at the end of the session. D3.js was chosen for its memory efficiency and rich feature set, surpassing other DOM manipulation methods [12]. It also complies with W3C [181] standards and enables interactive chart creation, beneficial for future audio integration to test screen reader accessibility.

Considering the space and size constraints of printed charts on embossed paper, adherence to several print guidelines for tactile illustrations [47, 137, 141] is essential. Related requirements are summarized as follows:

- Lines should be capable of being distinguished by touch, either by using different thicknesses or different types of symbols such as a dotted or dashed line.
- Lines of < 0.4 mm thick should not be used, as it can be difficult to obtain a sufficient bump on capsule paper.
- Text in tactile illustrations should be written in Braille and oriented horizontally. A margin of at least 3.0 mm should be left around the Braille characters.

Our rendering module offers an additionally accessible visualization mode, allowing users to export SVGs that align with printing guidelines for tactile graphics [141]. This includes a reduced number of axes labels, typically limited to 3 to 5 labels depending on the page size due to the size of Braille script. The (default) digitally accessible SVG mode is more suitable for 2-D tactile displays and screen readers (Figure 17-(b)), enabling the embedding of more visual content with corresponding description tags (e.g. <desc>). Figure 17 illustrates a sample rendering outputs for both digital and print-accessible SVG modes.

Chart4Blind Interface. Informed by insights obtained from need-finding interviews with our participants and guided by our design principles, we developed the Chart4Blind interface. This interface allows users to upload a chart image and facilitates the generation of an accessible version with text descriptions in diverse formats. In the course of our design process, we engaged two participants from the need-finding interviews to gather expert feedback. We carefully considered

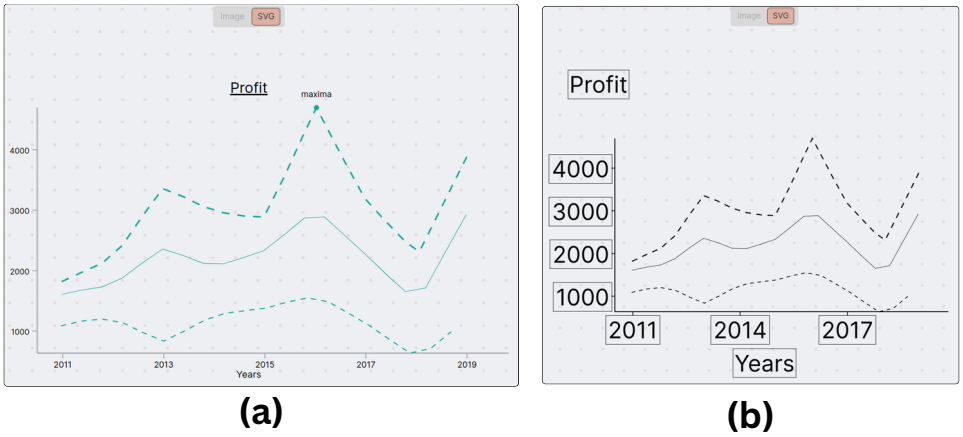


Figure 17: Rendering Module view for a bitmap line chart. (a) displays the digitally accessible SVG view, ideal for screen readers, and refreshable tactile devices. (b) shows the print-accessible SVG view, suitable for print modalities such as embossed papers or laser cut.

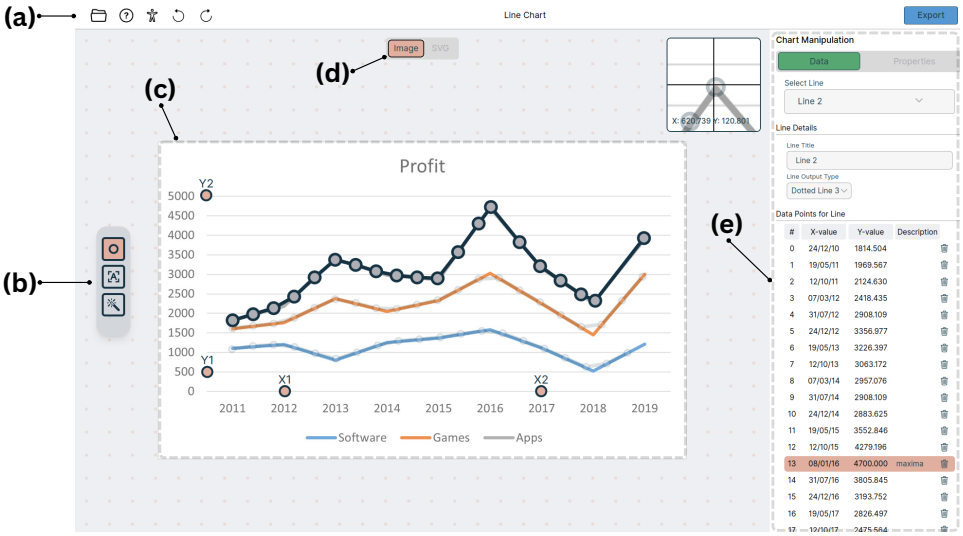


Figure 18: An overview of Chart4Blind interface sections: (a) Home menu for actions like upload, undo, redo, and tutorials. (b) AI toolbar with OCR and segmentation models. (c) Canvas for the uploaded chart, allowing interaction for calibration points and predicted line adjustments. (d) Rendering Module for real-time SVG visualization before export. (e) Metadata section for visualizing extracted line data and seamlessly drag-and-drop of textual content.

this feedback to enhance the seamless conversion process within the interface. The interface components are illustrated in Figure 18.

3.4.2 *User Study with Sighted*

We conducted a user study to see if the current implementation of Chart4Blind fulfils our design principles in terms of supporting user requirements discussed in the previous findings section to semi-automatically support the creation of accessible charts according to current standards [141]. We also conducted a follow-up study measuring the accessibility of our exports involving BVI individuals.

Participant. We invited a total of 10 sighted participants (T1-10, age range 22-34 years). Similar to the previous study, we collaborated with ACCESS@KIT to recruit participants via email lists. They did not receive any compensation. We screened participants for basic knowledge of charts: all participants were familiar with reading line charts and frequently worked with them. While their educational backgrounds varied, none of the participants had prior experience with chart accessibility or the conversion process. The study was part of a series of studies which were approved by the ethical review committee of Karlsruhe Institute of Technology.

Study Design. We prepared three line charts from our test set, for the conversion process and one simple chart for the tutorial session to help users become familiar with the tool. We consider three levels of complexity of charts to experiment with the conversion process, which we define as follows:

- Simple: Few lines with few data points and labels.
- Compound: Two or more lines with different label formats.
- Dense: Complex lines such as long sinusoidal waves or overlapping trends with text annotations.

We randomly selected line charts meeting the established criteria to ensure diversity (see Figure 19). These charts were collected from the recent *LG* dataset [134], featuring real charts from public documents across 5 distinct fields (e.g., social science, natural science, etc.). The charts are provided in PNG format, accompanied by the ground-truth hierarchical segmentation masks for all visual and textual elements.

Each participant completed a total of 3 sessions, progressing from Simpler to Denser charts. For each session, we recorded task completion time (measured in seconds), mouse clicks, SVG, and CSV exports. Heatmaps were generated using the recorded mouse clicks. Line point quality was measured using the Frechet Distance [3], and the SVGs were utilized in the follow-up study, as discussed later.

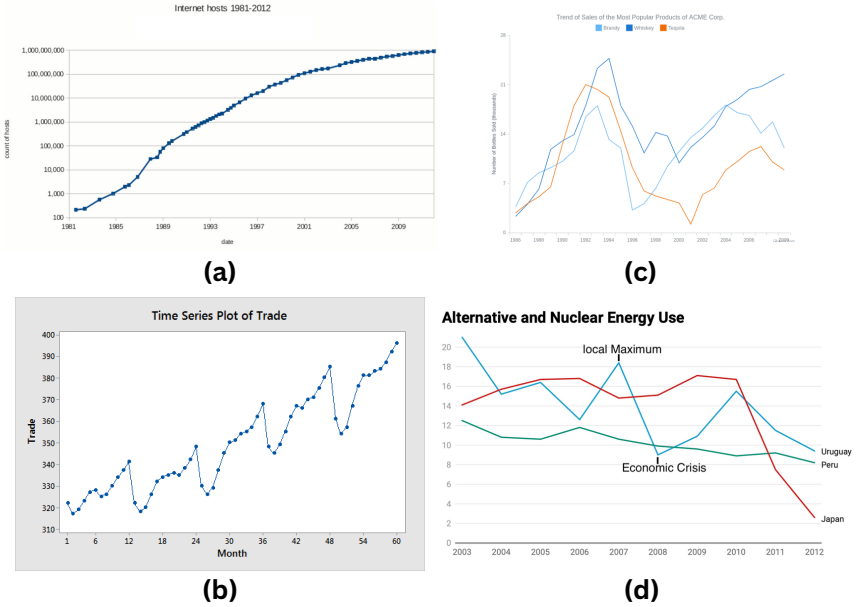


Figure 19: Four line charts with different complexities utilized for the user study: Simple charts (a) and (b) each contain one simple line trend for the tutorial and main session respectively. (c) A compound chart with additional lines overlapped, and visible axes. (d) A dense chart featuring relatively complex trends, point annotations, and less visible axes.

Procedure. After completing the consent form, participants were invited to a face-to-face user study. They went through a tutorial session to familiarize themselves with the Chart4Blind tool. During the tutorial, we presented the Chart4Blind tool using a sample chart (Figure 19 (a)) that was not used in the subsequent sessions. Additionally, participants were guided to an accessibility tutorial to understand the expected results. In order to get comfortable with the tool, we allowed the participant to interact with it for 15-20 minutes.

In the main sessions, each of the 10 participants was provided with three charts (see Figure 19 (c-d)) and asked to upload them and start the conversion process. Participants were informed that they were free to choose their preferred approach, whether utilizing the integrated AI tools or performing manual metadata labelling. They were also informed that they could revisit both the tool and accessibility tutorials at any time without losing progress if they had any questions.

After completing all the trials, the participants were presented with the SUS survey [88], followed by open-ended questions to express their opinions and thoughts. The entire experiment took approximately 90 minutes to complete. We developed an offline experimental system to facilitate the study, setting up Chart4Blind on a local laptop, with the tool accessible through the Chrome browser.

3.4.3 *Results & Discussion*

Task Performance. We measured the average time in seconds spent on each chart as illustrated in Figure 19. For the initial tutorial part, users took, on average, 9min and 5sec (STD: 3min and 45sec), with differences among users—some were

interested in understanding more about accessibility guidelines, while others chose to go through the tutorial via GIF animations only.

For the main session, we observed that participants completed the conversion process in an average of 4min and 36sec (STD: 2min 55sec). We informally asked the most experienced participant about how long it takes on average to create a chart with the tools she used (LibreOffic Draw). She reported a 10 minutes duration.

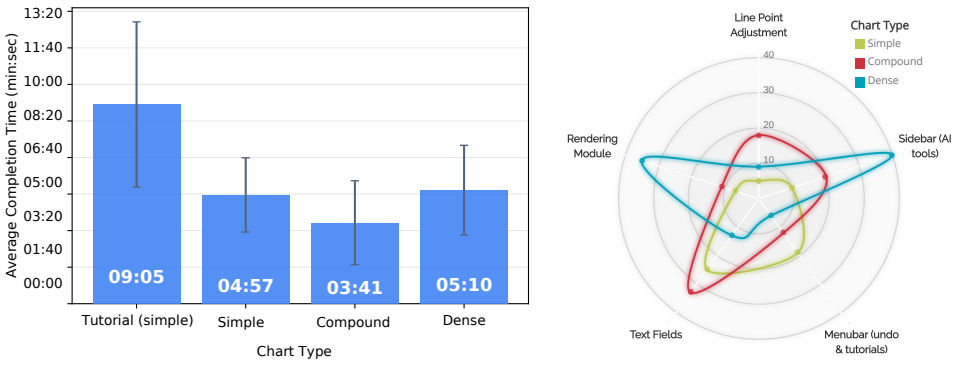


Figure 20: On the left, average conversion task completion time in minutes:seconds. On the right, a radar chart depicting the number of clicks for the top 5 sections interacted with in Chart4Blinds.

Interaction Patterns. To conduct a more comprehensive analysis of the time spent, we closely examined user click patterns and the duration they spent on various parts of the interface. In Figure 19-b, a radar chart demonstrates the average number of clicks on the top 5 interacted components of the UI. The analysis suggests an influence on completion time when incorporating the segmentation model. All users started with initial predictions, later making slight adjustments to the misaligned points, specifically focusing on points situated around the corners of the curved lines (see Figure 19-c). This trend indicates a decline in segmentation performance in these specific areas, suggesting a need for improvement in future work. However, in contrast to annotating the entire line manually, this discrepancy in performance is minimal compared to the overall number of points. In the case of the Dense chart, users edited 10.0% of the overall points.

Further investigation revealed that the time spent varied based on the elements within the charts. For instance, in charts with more textual components like the Compound chart, users tended to spend more time engaging with the text fields. On the other side, in the Dense chart, users engaged more with the rendering module, likely due to a few misaligned points caused by the curved lines. As users gained more experience and knowledge of accessibility guidelines, we observed a decrease in interactions with the menubar in each subsequent session.

Post-study Feedback. Sighted user satisfaction was assessed through a SUS survey [88] conducted after the interview. The study yielded an average score of 90. All participants either agreed or strongly agreed that the Chart4Blind tool is valuable for converting line charts into accessible formats and helps to understand the accessibility guidelines. Additionally, they agreed that they would recommend this tool to other colleagues.

We conducted a follow-up discussion to gather participant insights, covering the following aspects:

- Describe your experience with the tool.
- What did you find most intuitive and why?
- What parts do you find difficult or confusing and why?
- What suggestions or functions you would like to see or improve in the tool?

All participants expressed a positive experience with the tutorials, finding them informative enough to create an accessible chart. They felt confident following the conversion process through the rendering view, even without prior experience. T8 remarked, "I didn't feel the difference in converting Dense and Simple charts; they both took similar effort," indicating that the tool was seamless to utilize regardless of chart complexity.

6 participants found the line segmentation feature valuable as it significantly reduced the effort required to trace lines manually. A few participants also valued the OCR and drag-and-drop features equally alongside the segmentation model.

Interviews further revealed that while Chart4Blind generally met their needs, some participants suggested the following improvements: Three participants mentioned that automating the calibration step would be helpful, expressing an occasional lack of confidence when placing calibration points. T5 explained, "This chart [Dense] has a transparent axis with light gridlines; maybe other charts have no axis. I can't locate efficiently even with the zoom window."

Our sighted participants found the template-based summaries good but expressed a preference for more contextual summarizations reflecting domain knowledge presented in the chart. T10 suggested further integration of natural language models to enhance chart summarization.

3.4.4 *Exploratory Evaluation of Output Accessibility with Blind Users*

As an exploratory effort to evaluate output accessibility with assistive technologies, we conducted a remote study. We recruited three visually impaired individuals via the ACCESS@KIT center email list. We randomly selected three converted charts from our test set, and sent them to the participants, including embossed prints via mail and SVG exports via email. For the full user study materials refer to Appendix A.1. Participants were instructed to review the charts using a screen reader first and then explore the corresponding embossed print. We utilized the SoSciSurvey tool [100] to create an online questionnaire accessible with screen readers. The complete version of the survey is available in the Appendix A.1. Participants were guided to answer two questions:

- Could you describe your experience in using a screen reader to access the provided SVG? Were both the SVG and the accompanying chart description digitally accessible to you?
- Please discuss your interaction with the provided embossed print. Were you able to access both the visual and textual elements effectively?

All participants are familiar with the usage of screen readers, and with reading charts on tactile prints provided by ACCESS@KIT due to their studies in STEM subjects.

A few notable points were analyzed: two participants found the template summarization informative, while one preferred more comprehensive descriptions that also summarize the overall trends and patterns of the data. Furthermore, participants reported a few missing elements: plot title in chart (a), as well as missing legends in charts (b) and (c), highlighting potential errors made by sighted users when converting the graphics to an accessible format. Regarding the tactile prints, one participant mentioned that a few labels were very close to the border of the text bounding box, making it slightly more difficult to interpret.

In response to these observations, we made updates to the SVG rendering attributes to maintain a larger distance between the border and the inner text. Additionally, we are actively working on implementing a status system that notifies the user if any information is still missing before the export.

3.5 CHAPTER CONCLUSION

In this chapter, we introduced an approach for PVI individuals to access 2D document layouts, allowing them to locate text, images, equations, and various other elements. For this application, we developed new tactile interactions that provide both haptic and audio feedback, as well as control buttons for additional details. We evaluated the interface through user studies to ensure its effectiveness.

We then integrated a system for analyzing the layout of document visualizations using segmentation models and proposed the first benchmark for this task. Based on the hierarchical segmentations obtained, we developed a line tracing system specifically for line charts. We compared different networks and settings to achieve state-of-the-art performance in this task.

Towards the end of this chapter, we integrated our tracing approach and designed an intelligent user interface that streamlines the conventional process of creating tactile materials. This tool allows sighted participants to quickly extract, visualize, and export line chart images into accessible audio-tactile formats. We also conducted a formal user study to evaluate this tool.

In summary, this chapter represents a significant step towards enabling PVI individuals to localize visualizations and providing sighted individuals with tools to extract underlying data from images. The subsequent chapters will explore how these multi-stage approaches can be implemented using vision-language models, ensuring compliance with accessibility standards while maintaining robustness and efficiency. Here is a more detailed summary of the contributions of each section in this chapter:

Contribution 1: We are the first to propose a multi-modal (audio-tactile) interface that allows PVI individuals to access two-dimensional document layouts.

Contribution 2: For the first time, we explore semantic segmentation for visualization metadata extraction. We also developed a practical line tracing system specifically for line charts. A new benchmark was proposed, and comprehensive experiments were conducted.

Contribution 3: We propose *Chart4Blind*, an intelligent user interface that streamlines the chart-to-tactile conversion process for sighted users. A domain expert assessed the tool and found it especially valuable for complex tasks, where it has the potential to reduce labor effort.

4

LEARNING TO COMPLY WITH ACCESSIBILITY STANDARDS

Whether an AI model is assisting in authoring alt-text or creating tactile materials, it must be aware of accessibility standards, especially when collaborating with inexperienced users prone to errors. For instance, simply extracting data points from a line chart does not necessarily imply that the chart is now accessible. The extracted data needs to be reformed to align with the cognitive load and sensory capabilities of PVI individuals—one may not want to hear thousands of line values. This chapter discusses how models can be tailored to comply with accessibility standards and generate high-quality accessible modalities, both tactile and auditory.

This chapter is based on the publications [133] & [130] at (ICCHP 2024).

4.1 INTRODUCTION

Vision-language models continue to improve in generating chart descriptions and reconstructing visual content. However, from an accessibility perspective, the critical issue is not just in generating descriptions but in ensuring that these descriptions are truly accessible to PVI users by complying with established accessibility guidelines and standards. Despite advancements in VLMs, there has been limited effort focused on aligning these descriptions with the required standards. Moreover, even in the few attempts made, there are still frequent issues with quality and consistency, which affect their usability and effectiveness for PVI people. In this chapter, we present the *Alt4Blind* interface to aid sighted users in writing high-quality alt-text, and later, we discuss the development of the *ChartFormer* model, an end-to-end image-to-tactile conversion model that complies with accessibility standards.

4.2 COMPLYING TO STANDARDS THROUGH IMAGE RETRIEVAL

Alt text is a metadata field associated with an image that "serves the equivalent purpose" as the image, according to WCAG 2.0's A level (required) criteria [1]. The W3C categorizes images into seven classes, each requiring specific content for the associated alt text [2]. While decorative images should not receive alt text, other classes, such as "informative" images, should include a "short description conveying the essential information presented by the image." At an intermediary level, "func-

1 https://www.w3.org/TR/WCAG20/
2 https://www.w3.org/WAI/tutorials/images/

tional" images (e.g., images serving as buttons) should only convey the functionality of the icon. Other groups, such as Document Visualizations, classify images based on content type rather than purpose, including diagrams, graphs, photos, and art [25]. This group offers further considerations for alt text, including how to incorporate context, tone, language, and maintaining a neutral, factual description [24].

Alt text can originate from two primary sources: manually created by humans or (semi-)automatically generated by AI systems. Human-generated descriptions are often accurate but typically require expertise. Recent studies have highlighted a significant alt-text deficiency in publications, as shown by an analysis of 3,386 author-written alt-texts from HCI publications, where only 50% addressed extrema or outliers, and just 31% included details on major trends or comparisons conveyed by the graph [35]. Conversely, automatically generated captions are quicker to obtain and require no expertise but may suffer from information inaccuracy. Therefore, crafting high-quality alt text, particularly for charts, is a complex task.

Understanding the appropriate content to include in alt text is challenging [117]; indeed, the W3C provides a decision tree to help authors determine the necessary information [3]. However, existing alt-text interfaces do not sufficiently address the complexities of image classes impacting alt text content. To improve accessibility, several commercial platforms have incorporated automatic alt-text generation. For instance, PowerPoint recently updated its automatic captioning system for images in slides [159]. Social media platforms, such as Facebook, have developed automatic alt text features that provide a list of tags for items identified in an image, prioritizing people, followed by objects, and then elements of the setting [189]. These features represent significant advancements in alt text coverage on their respective platforms. However, questions regarding author interaction with and review of automatic alt text are increasingly important given its growing use.

Although these automatic tools enable the captioning of considerably more images on these platforms, their accuracy is not guaranteed. Sighted users are often presented with a suggested alt-text prediction and may accept it without further review due to a lack of knowledge. This lack of awareness regarding content quality raises concerns about accuracy. Hence, it is crucial to understand authors' experiences in creating alt text.

In this section, we investigate strategies to enhance the quality of alt text and foster greater engagement among authors in its creation. A significant focus of our investigation is on promoting positive behaviors that contribute to user awareness of accessibility. Additionally, our approach includes the development of a web tool designed to streamline the process, augmented by a novel deep retrieval model.

4.2.1 *Guiding Users to Author High-Quality Alt-text: Methodology & Dataset*

To address these issues, we initially conducted interview sessions with six participants possessing varying levels of accessibility expertise, ranging from none to advanced (working in the field of accessibility). We conducted exploratory interviews with these participants (experts: P1-P3, non-experts: P4-P6) to identify effective methods for authoring high-quality alt text.

3 https://www.w3.org/WAI/tutorials/images/decision-tree/

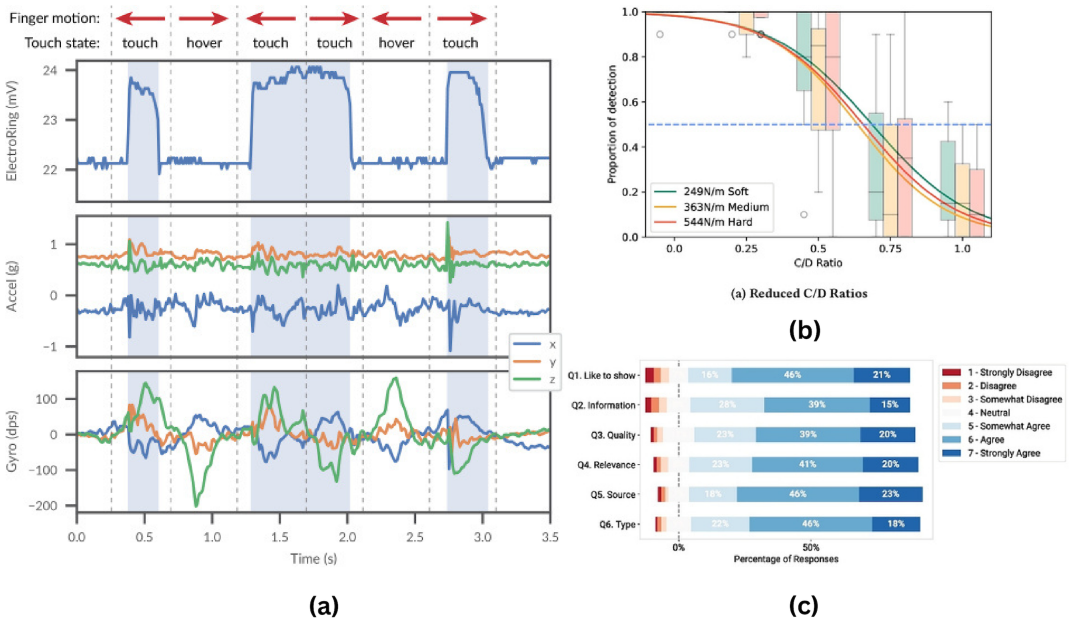


Figure 21: Sample Charts from the Alt4Blind dataset: (a) vertical panel line charts, (b) composite line-error bar chart, and (c) horizontal bar chart.

The interviews were structured as open-ended discussions. To facilitate the conversation, we followed these steps:

1. We introduced a basic line chart featuring three distinct lines and provided the W3C guidelines for writing image descriptions (illustrated in the upper left of Figure 23-2.
2. Participants were then instructed to write a description for a blind person.
3. Following this, we inquired about the challenges faced and their perspective on essential features for a web tool designed for this task.

The interviews took place during one-hour face-to-face sessions, during which we took written notes.

Interviews. Following the interviews, participants with limited experience in alt text creation highlighted their unfamiliarity with the concept, noting that alt text is often embedded within hidden tags. They emphasized the necessity of reviewing several examples before attempting to write alt text themselves. Moreover, they indicated that reading the guidelines alone was insufficient for producing comprehensive alt text, as the guidelines offer general, abstract recommendations that do not necessarily address specific image contexts. In contrast, expert participants, drawing on their extensive experience, were able to produce informative summaries. When questioned about the source of their expertise, they also attributed it to the review of multiple samples. This clearly suggests that examining high-quality examples in conjunction with reading guidelines would significantly enhance the user experience.

Technically speaking, this is an image retrieval task where the system could retrieve and rank samples that fit the user's specific case. It is akin to someone capturing an image of a dress and searching for similar or exact matches online

for shopping purposes. Based on these insights, we developed our subsequent chart retrieval dataset and model.

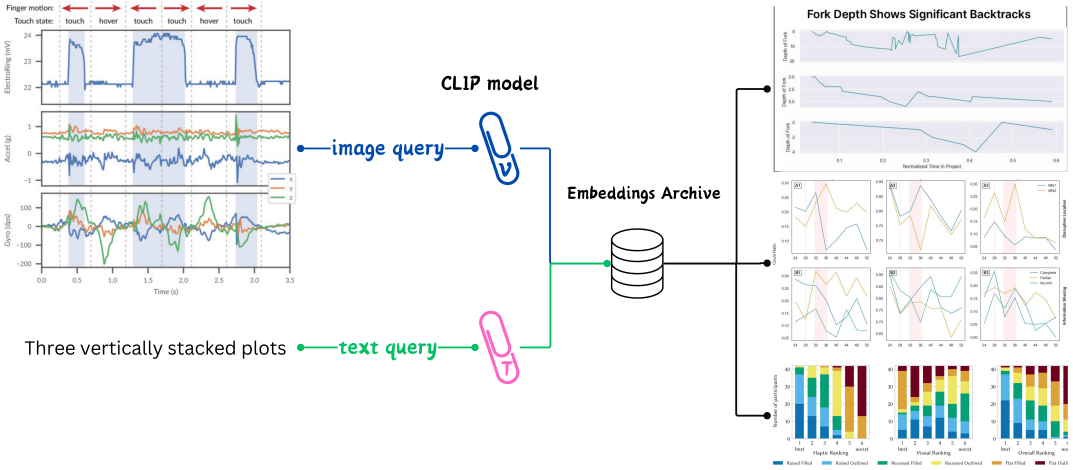


Figure 22: Our retrieval system leverages both the text and image encoder modules of the fine-tuned CLIP model. This ensures similarity at both visual and contextual levels.

Chart Retrieval Dataset. To provide high-quality reference images for use in this study, we first collected 25,000 images with alt text from publicly available HCI-related conferences over the past decade, thereby expanding the HCI Alt-Text dataset [35]. Subsequently, we filtered out images that were not charts, significantly short alt text, or lacked semantic content, resulting in a dataset of 5,000 chart images. This represents a tenfold increase over the previous dataset. These selected images, characterized by rich semantics, will serve as high-quality references for participants using our intelligent interface. Our dataset encompasses a diverse range of chart types, including common forms such as Line, Bar, Area, Pie, and Scatter charts, as well as unique visualizations not previously included, such as multivariate and panel charts (see Figure 21). We employed the 4-level semantic model proposed by Lundgard et al. [117] to assess alt text quality. Initial scoring was performed using ChatGPT-4, evaluating the number of semantic elements and levels present in each alt text. The highest quality samples were then manually reviewed for further validation.

4.2.2 *Model & Interface Design*

Image retrieval involves searching for images similar to a given query image or text, with a focus on ensuring the top results include similar chart semantics. To this end, we employed CLIP ViT-B/32 [152] as our baseline model. The CLIP architecture integrates both a text and an image encoder, each generating a 512-dimensional feature vector. The CLIP model utilizes an unsupervised contrastive pre-training approach to cluster samples in latent space. We fine-tuned the model on our dataset with a batch size of 16 for 10 epochs. During the inference phase (Figure 22), we input the image, extract the encoders' representations, and compare them to samples from the dataset. The top three candidates with the highest

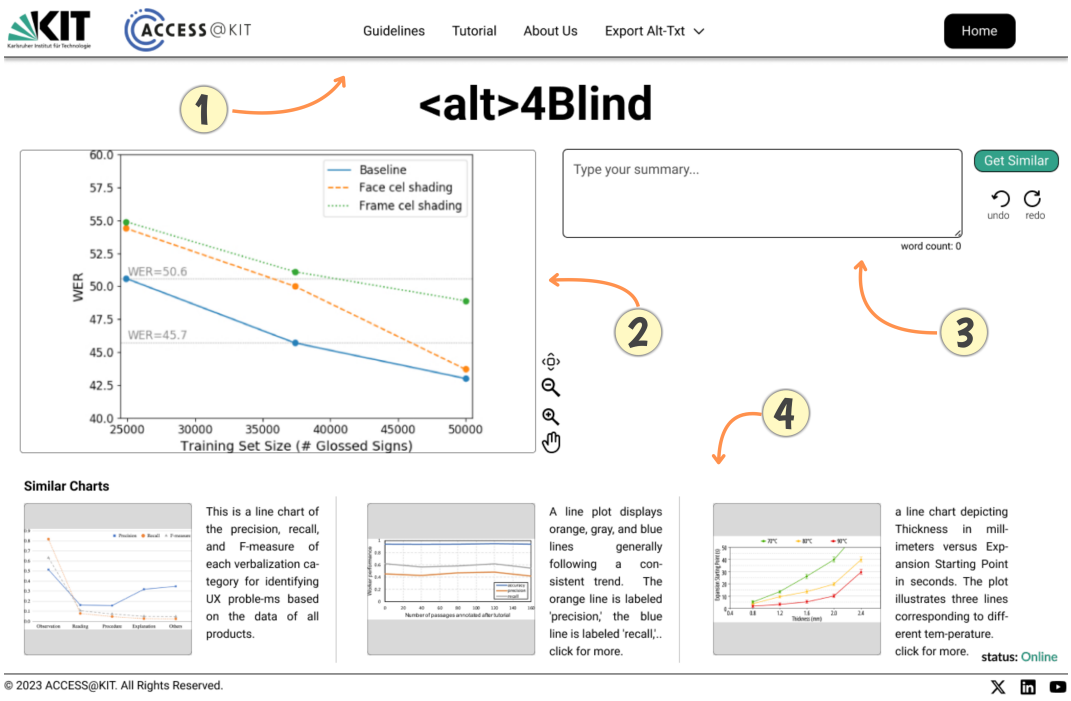


Figure 23: Alt4Blind UI: (1) Menu bar offering access to guidelines and a tutorial. (2) Space for uploaded images featuring a function bar (zoom, move, fit). (3) Text field for user input, accompanied by a button to update the retrieved image. (4) Retrieved charts based on the uploaded image, can be further enhanced with text query.

cosine similarity scores are then retrieved. Our model achieved 92% in Precision at 3 (P@3) and 85% in Recall at 3 (R@3), demonstrating high capabilities in displaying similar chart images within the top three results. The model can also be queried with text input, which is useful when the user types their own summary and seeks similar suggestions.

Alt4Blind Interface. To provide a better streamlined experience for users, our prototype leverages the capabilities of React JS[4], a robust JavaScript library for building dynamic user interfaces. The interface includes a user-friendly landing page where users can drag-and-drop chart images and access a tutorial designed to guide first-time users through the process.

Users can refer to the guidelines provided in a pop-up window, ensuring they are well-informed about best practices for alt-text creation. Once an image is uploaded, the backend model retrieves three images that are contextually and visually similar, as previously described. These similar images are displayed on the main page, allowing users to interact with each reference. Users can enlarge the images to view the full alt-text, see the reference paper, and copy or learn from the given summary.

The interface also includes a "Get Similar" button, located to the right of the text field, enabling users to refresh the list of suggested candidates based on the

4 https://react.dev/

summary they input. This feature ensures that users can continually refine their search to find the most relevant examples. As shown in Figure 23.

Finally, The interface have the option to export the alt-text as an HTML tag, SVG tag, or plain text, depending on their preference. This functionality allows for easy integration of the alt-text into various document formats.

4.2.3 *User Study*

To evaluate the efficacy and usability of the Alt4Blind interface, we conducted a comprehensive user study with six participants (P1-P6), encompassing both experts and non-experts in accessibility. The participants were presented with Panel and Multivariate charts, which were unfamiliar to all. The chart samples are depicted in Figure 21-(a) and (b). The participants were instructed to follow a structured protocol:

1. Review the guidelines and participate in the tutorial chart session first.
2. Use Alt4Blind to upload the chart and create alternative text for the image provided.
3. Describe their experience in detail.

During the sessions, we recorded user interaction behaviors using manual notation. Each session lasted approximately one hour.

4.2.4 *Results & Discussion*

All participants completed the guideline review and tutorial session without difficulty, gaining a basic understanding of the process required for creating alt-text. Participants P4-P6, who were less experienced, found the feature allowing the copying of sentences from similar charts especially useful in crafting their descriptions. They noted that this feature significantly improved their confidence and the quality of their alt-text. Expert participants (P1-P3) as well appreciated the ability to review high-quality examples. P1 suggested increasing the number of similar charts displayed from three to a larger number to provide more options. P3 suggested a feature to replace irrelevant images among the selected similar charts, suggesting an enhancement in the relevance and customization of the AI retrieved results. Overall, participants found the interface intuitive and user-friendly.

Limitations. Our tool has proven effective in enabling both inexperienced and experienced users to author high-quality alt texts. However, it could be further enhanced with additional intelligent functionalities, such as LLM-based feedback and descriptions, which are currently under development. While the current implementation aids users in creating alt text, we believe integrating captions into the tool is crucial, as captions and alt texts often complement each other. The existing user interface does not allow users to control the similar chart section, a feature multiple participants requested to improve their interaction experience. Future iterations should offer users the ability to omit, replace, or view additional charts.

Although our initial investigations are promising, they were conducted with a limited number of users and lacked comprehensive control measures. Future studies should involve a larger and more diverse group of participants, including individuals with blindness and visual impairments.

4.3 END-TO-END TACTILE MATERIAL CREATION

One of the key elements of the educational process is to provide students with appropriate scientific materials that facilitate the acquisition of knowledge and skills. These materials often engage multiple senses, enhancing understanding and retention. For example, visual aids such as diagrams, schematics, and charts help students grasp difficult concepts, with sight playing a crucial role in familiarizing students with these resources. However, PVI individuals face significant challenges in their cognitive processes due to the lack of visual input. This required the development of methods to present information that sighted people can access without difficulty.

While the problem of textual materials for PVI individuals was addressed early on through the use of the Braille alphabet, which allows the representation of 63 characters using raised dots, adapting drawings and graphs to the needs of the blind is much more complex. Technological solutions are required to make these materials accessible. Screen readers with speech synthesizers can provide audio descriptions of information presented in graphics. However, these descriptions must be prepared in advance and, in many cases, simply reading an audio description is insufficient. Tactile printouts, which allow users to feel graphics and obtain information through touch, are more effective [2]. A tactile image is a printed graphic in which shapes are represented by raised dots and lines. This requires careful preparation according to strict rules and guidelines so that a blind person can accurately interpret the shapes [19]. Several factors influence the perception of tactile graphics:

- Arrangement of objects: Overlapping or widely spaced objects can introduce errors and prolong the understanding process.
- Limiting the number of objects: Too many objects can disorient the user during analysis.
- Complexity of the structure of objects: More complicated objects take longer to understand.
- Differentiation of objects: Features of objects should be distinctly different to avoid confusion.

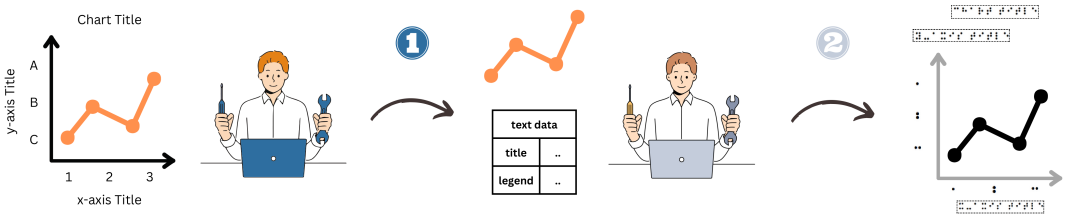


Figure 24: Conventional process of converting raster images into tactile material

Given the challenges of making visual information accessible to PVI individuals, the conventional approach as shown in Figure 24, often involves having expert sighted users manually redraw components using graphic software. This process typically requires extensive time and effort, as it involves simplifying complex visuals, ensuring clarity, and adhering to accessibility guidelines.

A further step involves using intelligent approaches to streamline this process through the automatic or supported creation of accessible visual content. By utilizing current advancements in vision and language models, it is possible to significantly reduce the time and effort required to adapt charts and diagrams for PVI individuals. As we demonstrated earlier, one could use these models to help extract metadata such as data points from line charts and textual content with few user interactions [134]. Many researchers have followed similar approaches for different types of visualizations [1], yet they all require human involvement as the models are trained to perform a single task: extracting the data. Hence, we ask: *how can we develop models that not only perform the primary task but also comply with accessibility standards, ensuring end-to-end usable content for PVI individuals?*. This question drives our exploration into creating the ChartFormer model and dataset.

4.3.1 *Dataset & Standards*

These tactile charts are often created using vector graphics software such as Inkscape and LibreOffice Draw, and saved in the Scalable Vector Graphics (SVG) format. SVGs are XML-based files that store geometrical shapes using mathematical formulas in a hierarchical structure. This format offers several advantages for creating accessible graphics [46]. First, each element in an SVG can be assigned different styles, which translates into distinctive textures in the tactile version. Second, SVG files can hold supplementary textual descriptions, enhancing interactivity when used with screen readers or tactile displays. Third, SVGs can be resized without blurring or distortion, making them ideal for varying paper sizes or zooming on tactile displays [131]. Creating an SVG chart from a raster image requires careful simplification of both textual and visual content to support tactile formats while preserving the integrity of the chart information. This process includes reducing textual content, such as limiting the number of axis labels, and focusing on crucial visual elements, like emphasizing significant scatter points in a scatter plot. Due to these complexities, crafting vector graphics is not a trivial task.

Therefore, we first addressed the need for a benchmark to better evaluate and train deep learning models for this task. By establishing a comprehensive bench-

mark, we can assess the performance of these models and ensure they comply with accessibility standards.

Dataset. ChartAssistants [124], a recent work, stands out for its comprehensive collection of chart images, each paired with detailed metadata. Although it lays a solid groundwork for converting charts to code, it cannot generate accessible visualizations. In contrast, datasets oriented towards accessibility, like VisText [173], have focused on making visualizations accessible through chart summarization tasks, but none have considered the tactile modality. **To our knowledge, we propose the first dataset for the task of *Chart2Tactile* conversion.**

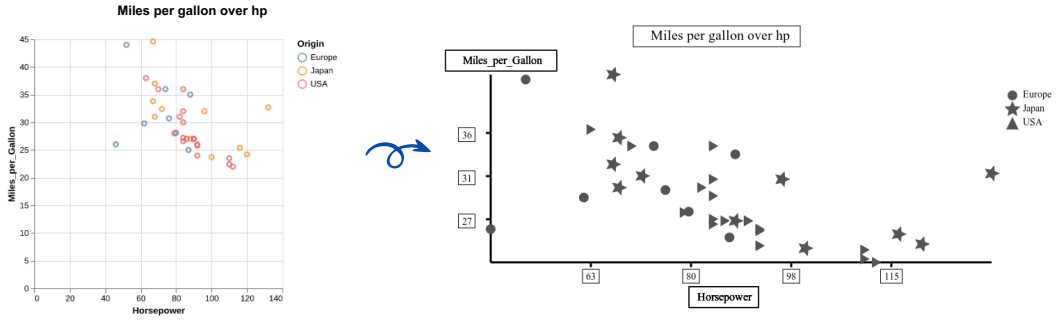


Figure 25: A scatter plot sample: (a) the original synthesized raster image; (b) the tactile version following accessibility guidelines.

Our dataset comprises 10,000 tactile chart images, spanning 4 categories (line, bar, scatter and error-bar charts) each accompanied by time series data and a raster version. We identified the VisText [173] and ChartX [190] datasets as the most suitable choices. VisText offers 8,822 images, complete with their data tables and accessible summarizations, featuring univariate time series. ChartX contains 48K chart data covering 22 topics, 18 chart types, with each chart including four modalities: image, CSV, Python code, and text description. A sample raster image is presented in Figure 25-(a).

Standards. Rendering the metadata as tactile charts necessitates adherence to established guidelines to ensure that the charts are accessible by individuals with visual impairments. This process involves not only the translation of visual information into a tactile format but also the thoughtful consideration of how various elements can be differentiated by touch. We followed various tactile printing guidelines [49, 141] to create accessible SVGs. The key requirements we adhered to are summarized as follows:

1. Elements should be distinguishable by touch, using varying thicknesses or symbol types such as dotted or dashed patterns.
2. Thin elements should be avoided.
3. Text in tactile illustrations should be in Braille, oriented horizontally.

Additionally, we collaborated with an expert from the Center for Digital Accessibility and Assistive Technology[5] at Karlsruhe Institue of Technology, specializing

5 https://www.access.kit.edu/english/index.php

in converting educational materials for people with blindness and visual impairments. Their feedback included the following recommendations:

1. Enclose text content with a bounding box for better exploration and distinguishing separate texts more effectively.
2. For dense charts such as scatter plots, only significant, non-overlapping points should be drawn to avoid clutter.
3. Embed description tags for both text and visual elements to enable accessibility via screen readers.

For transforming metadata into SVGs, we used the *svgwrite* Python package[6]. For each time series, we synthesized an SVG template and rendered a raster image using *Vega-Lite*[7]. To ensure accuracy, we manually selected samples from each category of the data and conducted a thorough verification process. A tactile sample is illustrated in Figure 25-(b).

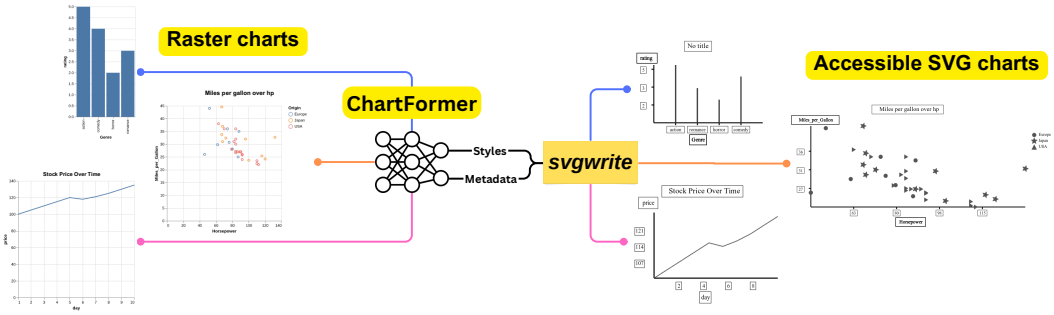


Figure 26: The ChartFormer takes a raster x-y plot as an input. The essential metadata and styles are extracted, which are then used to populate the svgwrite templates. For better viewing resolution, please visit our project page.

4.3.2 *ChartFormer Model*

We selected the state-of-the-art publicly available model LLaVA-1.5 [111] as our baseline. This choice was motivated by the desire to provide accessibility centers with a cost-effective solution that does not require additional expenditures. At the time of selection, LLaVA-1.5 was top-performing in vision tasks such as document question answering, which we anticipated would expedite the training process for chart data.

We utilized the baseline weights from ChartLLama [70] and adopted the same hyperparameters for training. The model was then fine-tuned for 10 epochs using our dataset, aiming to analyze x-y raster chart images and extract simplified metadata with appropriate styling for *svgwrite* code, see Figure 26). The specific information extracted includes:

- Chart Type and Titles: Identification of the x-y chart type and extraction of titles for the plot, axes, and legend. This step is crucial for providing context and clarity.

6 https://svgwrite.readthedocs.io/en/latest/
7 https://vega.github.io/vega-lite/

- Axes Range and Labels: Determination of the axes range with 3 or 4 labels that cover the entire period. Labels must conform to specific encodings (e.g., int, float, fraction, date/time, text) to ensure accuracy.
- Time-Series Data Extraction: Extraction of time-series data necessary for drawing the chart, forming the core of the visual representation.

The extracted data is rendered using predefined svgwrite code templates for each chart category. For scatter plots, we ensure clarity by drawing 10 points per label unit and separating overlapping points, following the SVG Guidelines described above.

4.3.3 *Experiments*

We conducted a pilot user study with four people with PVI (three males and one female, aged 21-29) to evaluate the effectiveness of the generated SVGs on a HyperBraille 2D tactile display [8].

Procedure. The user study images were randomly sampled from the LG dataset [134], which includes real charts. At the beginning of the session, a test graph was provided to introduce the participants to the available interactions. Afterwards, the 3 line charts shown in Figure 27 were displayed on the HyperBraille and the participants were asked to explain and identify the key elements, titles, labels and legends and count lines, in each graph, as well as name few points intersection and line trend. For the full user study materials refer to Appendix A.2

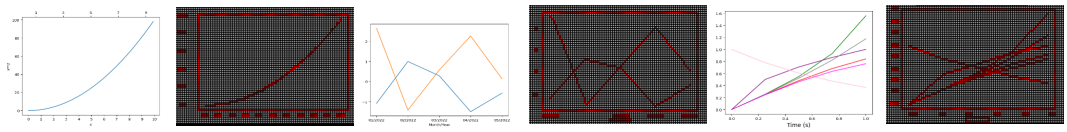


Figure 27: SVG-formatted line charts used in the user study, showcasing varying complexities: (A) a single line; (B) two lines; (C) six lines. For better viewing resolution, please visit our project page.

Results & Discussion. All participants successfully completed tasks related to charts (A) and (B), which involved identifying intersections and counting lines. They could also accurately describe the line trend as increasing, decreasing, or constant. However, in chart (C), participants encountered difficulties in counting all intersections, likely due to the chart's high density. Two participants used zoom features on the tactile display to discern closely positioned elements and intersections more clearly. Audio descriptions were also suggested by one participant as a way to facilitate access to the chart's textual elements. A common suggestion from all participants was regarding SVG rendering, specifically to address the staircasing effect in the tactile output. The need for smoother line rendering to avoid jagged or stair-like appearances was emphasized.

8 https://metec-ag.de/en/produkte-graphik-display.php

Limitations Although our system has been positively received by the participating people with blindness and visual impairments and collaborators, we believe there is still significant room for improvement: (1) Our system mainly targets x-y plots with two axes and charts of a single type. Future implementations could encompass other chart types. (2) Adding an interface to our system could allow sighted individuals to modify the chart before exporting it, ensuring that textual and visual details are accurately represented. (3) Conducting a larger, formal user study is necessary to assess the performance and furthermore, to experiment with different types of charts beyond just line charts.

4.4 CHAPTER CONCLUSION

In this chapter, we introduced two new tasks for the chart analysis field: image based chart retrieval and chart-to-tactile tasks. Additionally, we developed a novel vision-language model application for assistive technologies that not only streamlines the creation of accessible materials but also promotes good practices and compliance with standards.

First, we presented a new alt-text authoring interface that assists users by retrieving similar charts to use as a reference for authoring high-quality alt-text. This chart retrieval task is achieved through our novel model, which retrieves charts based on both contextual and visual elements such as label formats, colors, markers, etc. Our pilot user study demonstrated that presenting similar charts to authors is more effective than AI-assisted approaches, which cannot always ensure the generated alt-text conforms to standards.

We also investigated the tactile modality and proposed the chart-to-tactile task. We conducted an extensive analysis of our dataset and demonstrated the potential of using vision-language models for tactile creation.

In summary, this chapter represents the first attempts to investigate the compliance of vision-language models with accessibility standards and guidelines. It also lays the groundwork for the community to address further limitations and challenges identified in the proposed tasks. Below is a more detailed summary of the contributions from each section in this chapter:

Contribution 1: We present the image-based chart retrieval task, where charts are retrieved based on contextual and visual semantics. The Alt4Blind model and benchmark were proposed, along with an application interface developed for alt-text authoring to assist sighted individuals in creating high-quality alt-text.

Contribution 2: For the first time, we introduce the image-based chart-to-tactile task. This challenge is accompanied by a comprehensive benchmark, and we propose an approach to adapt vision-language models for vector graphics tactile material creation.

5

CAPTURING THE DIVERSE NATURE OF VISUALIZATIONS

Authors often present complex data using appealing designs to help users envision the topic, increase engagement, and enhance memorability. However, we strive to make visualizations accessible and address these issues while improving generalization across different chart styles. Current approaches tend to train on specific classes of visualizations with limited styles, primarily because their data are mostly synthetically generated. Consequently, there is a essential need for systems that can adapt to the diverse nature of visualizations. This raises the question: How can document assistive technologies and models cope with the diverse designs and input formats of visualizations? In this chapter, we discuss how to benchmark model robustness to the diverse nature of charts and how to learn better representations to handle real-life scenarios effectively.

Part of this chapter is based on the publication [135] at (ICDAR 2024).

5.1 INTRODUCTION

Authored Visual Diversity. Where would we be without charts? Not only are charts more appealing to our eye than raw data, but they also make it much simpler for us to spot trends. The old saying "a picture is worth a thousand words" also applies to charts. Managers don't always want to look at every record in a spreadsheet; they would rather make their decisions based on trends, whether they be good or bad. That is why there has been recent advancement in graphing software giving us new tools to compose artistic charts rival those of professional artists, see Figure 28. Most of these software provide visualization models, where data serves as the starting point for the design or analysis process, after which designers, developers, and analysts select can alter the visual mappings, labels, notions, to make that data more legible and actionable. Over the past 50 years, a large body of research has successfully focused on developing and optimizing visual mappings and interactions, creating a diversity of different visualization genres tailored to unique data, tasks, audiences, and contexts. For example pictorial visualizations that uses icon-based language to visualize dataset points.

Source-Based Variations. In fact, author variations are not the only challenge for adaptability. Visualizations captured in the wild using different devices, such as cameras or scanners, also introduce another type of variation, as illustrated in Figure 29. Orientation, distortion, resolution differences, lighting conditions, etc., are all potential factors that can affect the captured images. According to our obser-

vations, perturbed chart images cause significant performance drops among all state-of-the-art chart analysis models. A simple experiment involving the input of two identical charts but collected from scanned and phone captures sources into GPT-4.0 demonstrated the instability and hallucination in the authored summarization.

Most available datasets include only digital charts with up to 3-4 types that are synthetically generated using uniform templates, severely lacking in variability. So far, state-of-the-art chart analysis models have primarily focused on extracting data from charts and performing knowledge-based tasks such as question-answering, summarization, and chart-to-table conversion. This lack of style and input diversity hinders the development of a multi-domain, general chart analysis field. Consequently, current research is not adequately addressing these requirements or considering robustness aspects, leading to more hallucinations and poor context capturing. This issue is particularly critical for assistive technology applications. Therefore, in this section, we raise the question of how vision-language models can adapt to these scenarios. We first comprehensively analyze the robustness of available models for both chart summary and chart-QA tasks. Then, we present a new training approach to equip large models with adaptability skills, specifically focusing on enhancing performance for the chart summary task.

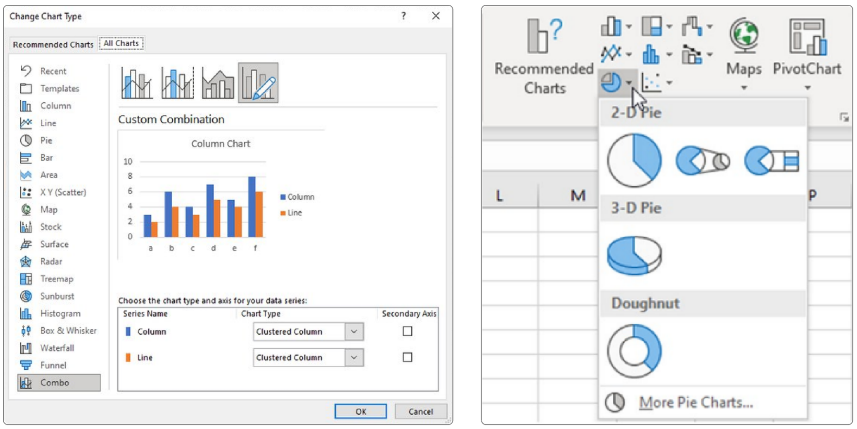
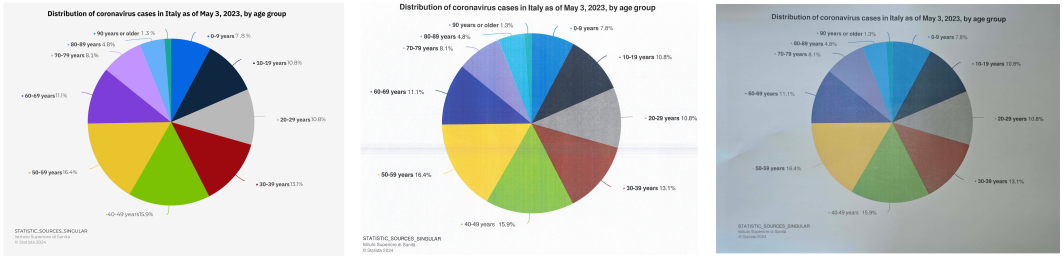


Figure 28: Variations of the chart in Microsoft Excel. (a) Various chart types (b) Sample pie chart variations.



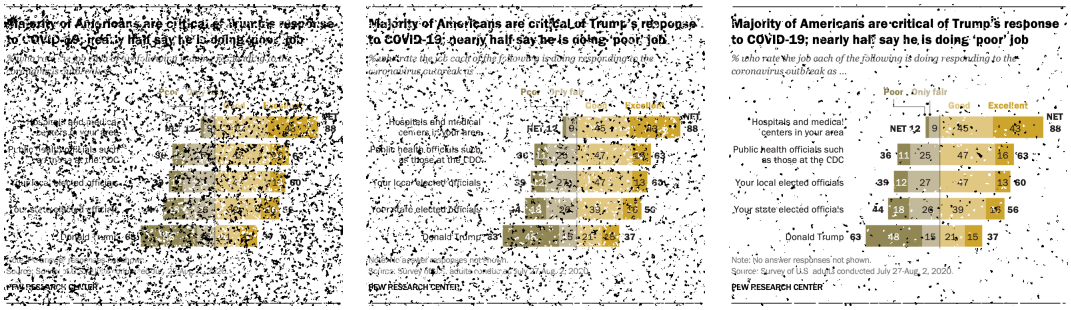
The chart shows the distribution of coronavirus cases in ... The highest percentage of cases is among the 50-59 age group at 16.4%, followed by the 40-49 age group at 15.9%. The lowest percentage is in the 90 years or older age group at 1.3% ...

The image is a pie chart... The segments are colored differently and labeled with the following age groups and percentages: 0-9 years: 18.2%, 10-19 years: 5.8%, ...

The chart illustrates the distribution of coronavirus cases in Italy by age group. The highest percentage of cases is in the 50-59 years group, which is highlighted in purple, ... The 10-19 years group, shows a higher percentage than expected at 20.8%.

(a) (b) (c)

Figure 29: Three images of the same chart collected from different sources: (a) digital, (b) scanned, and (c) printed version. The summaries contain inaccuracies, with the false statements underlined. For full responses refer to Appendix A.3.



(a) (b) (c)

Figure 30: Sample chart from the "blotches" perturbation at three levels: (a) level 10, (b) level 5, and (c) level 1.

5.2 ROBUSTNESS BENCHMARK

The democratization of artificial intelligence and deep learning systems has raised public concern regarding the reliability of these new technologies, particularly in terms of accessibility and robustness. These concerns, among others, have motivated the creation of new European regulations known as the Accessibility Act and the AI Act, which aim to regulate the use of these systems by the public.

In the context of a substantial increase in incoming documents, both at the individual level (e.g., mail, educational materials) and within large companies, document visual analysis based on computer vision techniques has proven to be an effective method for automatically extracting data from documents such as ID cards, invoices, and tax notices.

Unfortunately, these techniques have proven to be inaccessible to people with visual impairments and vulnerable to different input sources, such as phone-captured or scanned documents, as we have discussed previously. The robustness of chat comprehension models, particularly from an accessibility perspective, has not yet been thoroughly investigated. In fact, there have been few attempts and benchmarks for document robustness analysis, and the conclusions of these studies might not apply to visualizations and accessibility. Charts, for example, have different semantic information than most images; they contain interrelated text, diverse

layouts, and often a stylish design. This is why the robustness of visual models for chart analysis (e.g., alt-text authoring) should be evaluated in an appropriately controlled setting. In this section, we evaluate the robustness of several state-of-the-art visual models and define the different robustness levels based on human user studies for the chart-QA task.

5.2.1 *Robustness Benchmark Dataset*

For our experiments, we used the ChartQA dataset [121], a large-scale collection of real-world charts and question-answer pairs designed for benchmarking chart-related question answering. Masry et al. compiled this dataset by crawling charts from four sources: (1) Statista (statista.com), an online platform with charts on topics like economy, politics, and industry; (2) Pew Research (pewresearch.org), which publishes reports with various charts on social and economic issues, demographic trends, and public opinion; (3) Our World In Data (OWID) (ourworldindata.org), a platform with thousands of charts on global issues such as economy, finance, and society; and (4) OECD (oecd.org), a global organization providing reports and data analysis for policymaking. The dataset includes 9.6K human-authored question-answer pairs and 23.1K machine-generated ones. Sample data are presented in Appendix A.4.

5.2.2 *Human Perception for Perturbed Visualizations*

To understand model perception of noisy chart image inputs, we first need to evaluate human perception and abilities for this task. To address the varying difficulty levels of noisy images, we conducted a structured user study with a large group of participants. This study aimed to determine the different levels of difficulty, which can be compared to the performance of vision-language models. By involving human chart question-answering, we identified the levels of understanding, creating a more structured evaluation approach for these models. In the following section, we discuss the details of the user study.

Participants To evaluate human perception against AI, we designed a user study on JotForm[1] and recruited 43 participants who are over 18 years old, university graduates, experienced in reading charts and answering questions, and able to complete the survey in English. We adhered to best practices for ethical human subjects survey research, and all participants provided consent for their responses to be used for academic research purposes. We also complied with GDPR data protection regulations. The study could be conducted on laptops or phones, and three randomly selected participants were compensated with a prize of 10 euros.

Procedure. We selected 10 perturbations suitable for accessibility settings (see Table 8), i.e., challenges PVI users may experience when capturing or scanning visuals. The perturbations as well as their severity levels were adapted from [33]. Participants were presented with a different chart image for each perturbation, shown at severity level 10, and asked whether they could understand the content of the

1 www.jotform.com

Table 8: List of perturbations tested in the user study with examples from PVI cases.

Perturbation	Real-life Experience for PVI Users
Blotches	Spots or smudges on the camera lens or printed chart.
Color	Faded print colors.
Dilation	Overexposure causing text to appear overly thick and blurred.
Erosion	Worn out or faded text on old or damaged documents.
Elastic Transform	Curved text from taking a picture of a bent document.
Fibrous Noise	Scratches or lines from a dirty scanner glass.
Gaussian Blur	Out-of-focus images due to shaky hands while capturing.
Motion Blur	Blurred images from moving the phone while taking a picture.
Shifting	Misaligned text from poorly scanned documents.
Uneven Brightness	Uneven lighting causing dark or overly bright areas.

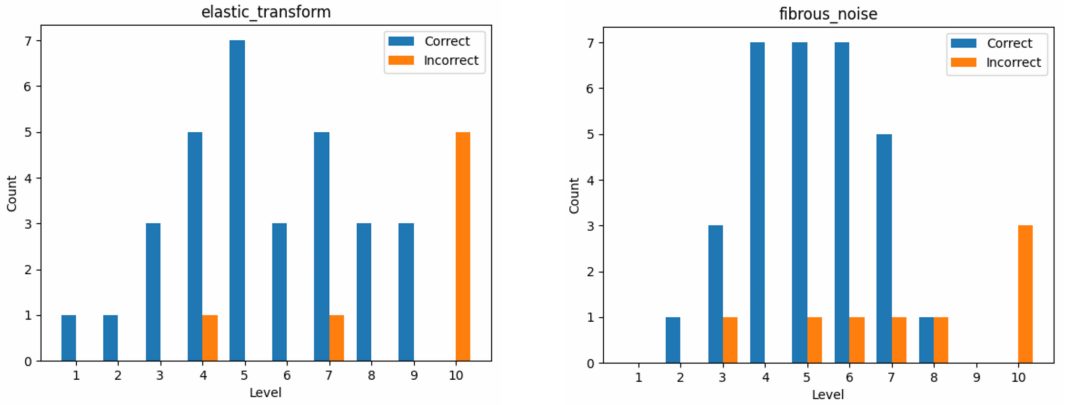


Figure 31: Statistical results from the robustness user study, showing the frequency count of correct and incorrect responses versus the perspective level chosen by the participants.

chart. If they answered "yes," they were then asked a question about the chart and required to type their response. If "no" then lower severity level is shown. Upon completion, they moved to the next perturbation type. See Figure 30 for an example of different levels of the perturbation images. For more information about the survey, refer to Appendix A.4.

Findings Most of the participants (93%) were able to answer the questions correctly. We only considered the correct answers in the statistical evaluations. We analyzed the participants' results to determine the three suitable levels to compare with the vision-language models. For each perturbation, we counted the number of correct answers per level, as shown in the bar chart in Figure 31. The "difficult" level is determined by "the highest level where at least one person answered correctly." The "medium" level is identified as the mode of correct answers, while the "easy" level is defined as the level where 90% of participants answered correctly. Table 9 below shows each level number number of each perturbation. One exception was the "color" perturbation, where most participants failed. They were confident that the chart was visually clear, but the missing color information misled them

when trying to trace the two intersecting lines to determine the right answer (see Figure 40 in Appendix A.4). For other perturbations, it was visually obvious that something couldn't be fully understood, so all participants' responses fell into similar levels.

Table 9: The chosen perturbation levels for each category based on the analytical results of the user study.

Perturbation	Easy Level	Medium Level	Difficult Level
Tutorial	2	7	10
Blotches	2	6	10
Color	5	7	10
Dilation	4	9	10
Elastic Transform	3	5	9
Erosion	4	5	10
Fibrous Noise	3	4	8
Gaussian Blur	4	5	10
Motion Blur	3	5	10
Shifting	1	7	10
Uneven Brightness	3	4	10

5.2.3 *Vis-Lang Models Robustness Evaluation*

Following the human perception evaluations, we conducted a thorough and comprehensive evaluation of state-of-the-art chart understanding models on the task of chart QA. In line with Chart-QA [121], we used a *relaxed accuracy* measure for numeric answers, allowing for minor inaccuracies that may result from the automatic data extraction process. Specifically, we considered an answer correct if it was within 5% of the gold answer. For example, if the ground-truth answer is 100, any answer between 95 and 105 would be considered correct. For non-numeric answers, an exact match was required to consider an answer correct. We call our benchmark as "CHAOS" Chart Analysis with Outlier Samples. The average results for each method are presented in Table 10. For the sake of comprehensiveness, we also experimented with document-related VLMs and general-oriented VLMs. As shown in the table below, even the introduction of fine noise at the easy level significantly degrades the model's perception abilities. At the hard level, the performance degradation is even more pronounced. Among the models, TinyChart [196] consistently demonstrated decent performance across all levels of perturbation, likely due reduce the burden of learning numerical computations through a Program-of-Thoughts (PoT) learning strategy, which trains the model to generate Python programs for numerical calculations. This suggests that specialized chart-related models, particularly those with robust analysis tools, are better suited for handling noisy and perturbed data. The overall trend shows that as the perturbation level increases, the model's ability to accurately interpret chart data diminishes,

emphasizing the need for further enhancements in model robustness to handle real-world visual challenges effectively.

Table 10: Results on CHAOS benchmark of ChartQA.

Model	#Param	Resolution	Inference Throughput	ChartQA Clean	**CHAOS** Easy	Mid	Hard
General							
Llava1.5 [114]	13B	336×336	1.94 it/s	55.32	33.10	6.31	0.61
Qwen-VL [9]	9.6B	448×448	1.65 it/s	61.60	39.00	9.22	0.40
Document-related							
UReader [195]	7B	224×224(×20)	1.67 it/s	59.30	40.29	13.80	0.55
DocOwl1.5 [79]	8B	448×448(×9)	1.56 it/s	70.50	53.90	16.00	1.10
Chart-related							
ChartInstruct [122]	7B	-	-	66.64	52.90	24.13	2.74
ChartLlama [70]	13B	336×336	1.94 it/s	69.66	60.49	13.72	0.92
ChartAst [124]	13B	448×448	1.47 it/s	79.90	62.97	18.32	1.80
TinyChart@768 [196]	3B	768×768	3.14 it/s	**83.60**	77.50	49.35	5.30

5.3 ADAPT TO DIVERSITY

Generated descriptions through Vision-Language (VL) models, unlike deterministic systems, are faster to obtain and require no expertise, but are highly susceptible to hallucinations [110]. Recent advancements in large VL models have significantly improved chart analysis field, enabling tasks like chart2text [93], chart2table [34], and chart2code [124], among others. However, these models are often trained on synthetically generated datasets or existing real chart corpus, which are either limited in size or do not meet accessibility guidelines. This limitation can result in semantically weak summaries that are brief and potentially inaccurate. To address this, we introduce introduces AltChart, a dataset particularly suitable for VL models with 10,000 real chart images with human-authored chart summarize, adhering to accessibility guidelines and semantically rich.

VL models have shown improvement in overall performance when scaling the pretraining corpus [70, 186]. The increased sample size enables these models to learn fine-grained representations. However, the use of synthetically generated data raises concerns about model robustness and biases towards certain visualization styles [11, 23]. To address these limitations, we investigate whether vision encoders could develop better representations through different training means.

Pretext tasks are among the methods that have shown promising performance in preparing models for complex tasks [156]. They challenge models to resolve smaller image-level tasks as a preliminary step before the mainstream one. Consequently, we have conducted experiments with multi-pretext tasks and found that it can achieve state-of-the-art performance on widely recognised chart summarization benchmarks. We also find that other available annotation types in datasets

(e.g., bounding boxes, segmentation masks, key points, etc..) could be made useful for vision language models with pretext tasks as they can be defined accordingly.

5.3.1 *Background*

High-quality alternative text is characterized by its semantic richness. Semantics could be words, phrases, or sentences grouped to define the theme of the overall text. Each semantic element helps further streamline the interpretation of the described content. For example, see Figure 32. Defining and extracting these textual semantics guides the model to form a representation of the patterns that an accessible description could embody. In this work, we use the term 'semantic' to refer to the textual keywords as shown in the aforementioned figure.

We build on the work of Lundgard, A., et al. [93], who conducted a thorough examination of visualizations, focusing specifically on the semantic depth of effective chart descriptions. They expanded the summarization guidelines framework with a more general conceptual model covering four levels of semantic content:

- L1: Chart construction properties (e.g., axes, encodings, title).
- L2: Statistical concepts and relations (e.g., outliers, correlations, statistics).
- L3: Perceptual and cognitive phenomena (e.g., trends, patterns).
- L4: Domain-specific insights (e.g., context relevant to the data).

The study suggests two key points: Firstly, it indicates that captions should communicate key trends and statistics, while also considering the preferences of the reader. Secondly, it highlights the importance of using existing accessibility guidelines as a foundation to enrich chart summarization research.

5.3.2 *AltChart Dataset*

While models for high-level understanding and especially for question answering have made extraordinary strides in recent years, there is still a wide gap to human performance. They still can't cope with the diverse nature of data a human can observe in daily life. We analyzed five current benchmarks and developed our dataset, designed to bridge some of the identified gaps in current research. Our dataset specifically targets L1 and combines L2 and L3 to simplify the annotation process. We decided to exclude L4 from our current dataset due to the domain knowledge required beyond input chart images, such as document-level topics. With the interest to explore this level in future research.

Existing Datasets. Table 11 lists the top five related datasets for chart summarization task. The Chart-to-Text dataset [93] compiles descriptions for charts from Pew[2] and Statista[3], covering line, bar, pie, and area charts. ChartSumm [155] expands on this by nearly doubling the dataset size and including longer summaries. However, our analysis shows that both datasets focus mainly on Statistical and Perceptual

2 https://www.pewresearch.org/
3 https://www.statista.com/

(L2L3) sentences, with 91% and 94% of their content, respectively, missing foundational visual sentences (L1). AutoChart [202], while offering a balanced mix of sentence levels through synthesized charts and template-based captions, suffers from limited variation. These datasets fall short on accessibility standards. In contrast, our dataset uses real charts and summaries collected from accessible venues, doubling the chart categories to include new challenging types like Compose and Panel charts.

More recently, datasets like HCI Alt Text [35] and VisText [173] have been developed to specifically address chart summarization for BVI individuals. Both datasets are rich with L1 and L2L3 semantics. VisText creates synthetic chart images using the Vega-Lite visualization tool, then used crowdsourcing for L2L3 summarize, while machine learning models are employed for L1 captions. In contrast, HCI Alt Text compiles figures from accessibility venues, filtering for those with alternative text. However, this dataset, intended primarily for analysis, comprises only 511 chart images. This limited size makes it challenging to train effective data-driven methods. To overcome these constraints, we adopted a similar methodology to HCI Alt Text, but expanded our collection to 10,000 chart images and manually annotated them with 10 text semantics.

Table 11: Overview of the five most related datasets. Our AltChart dataset includes real-charts and real-summarize, with a broader range of categories and semantics.

Name	**Data Type**		**Categories**	**Semantics**	**Image Count**
	Images	**Descriptions**			
Chart-to-Text [93]	real	real	4	✗	44,085
HCI Alt Text [35]	real	real	2	2	511
ChartSumm [155]	real	mixed	3	✗	84,363
AutoChart [202]	synthetic	synthetic	3	✗	23,543
Vistext [173]	synthetic	mixed	3	2	8,822
AltChart (Ours)	**real**	**real**	**8**	**10**	**10,000**

5.3.3 *Dataset Construction*

Considering the existing limitations in available data, such as the lack of semantically rich descriptions, short descriptions, or adherence to accessibility guidelines, we dedicated efforts to creating the AltChart dataset. We began by crawling HCI publications from five ACM[4] conferences (CHI, ASSETS, DIS, UIST, W4A) spanning 2015 to 2023. Our focus was on papers containing alt-text tags. This process yielded 8,000 PDFs and 43,510 images.

To capture high-quality images with alt-text in our corpus, we undertook three steps: (1) We fine-tuned a BERT-based classifier [42] on the HCI Alt Text dataset to determine whether the alt text corresponded to a chart and for sentence-level classification (L1/L2L3). The model achieved an F1 score of 93% on the test set. (2) We

4 https://dl.acm.org/conferences/

Table 12: Comparison of three leading datasets in terms of comprehensive summarization. *AltChart* stands out with significantly higher average sentence and word counts—nearly double those of the others—and showcases the most balanced L1 to L2/L3 sentence ratio.

Dataset	**Avg. Summ.**		**L1:L2L3 Ratio**
	Sentence Count	**Word Count**	
ChartSumm	2.0	45.44	1.17 : 98.83
Vistext	2.26	42.6	56.2 : 43.8
HCI Alt Text	3.66	77.0	74.2 : 25.8
AltChart	**5.67**	**136.35**	**44.9 : 55.1**

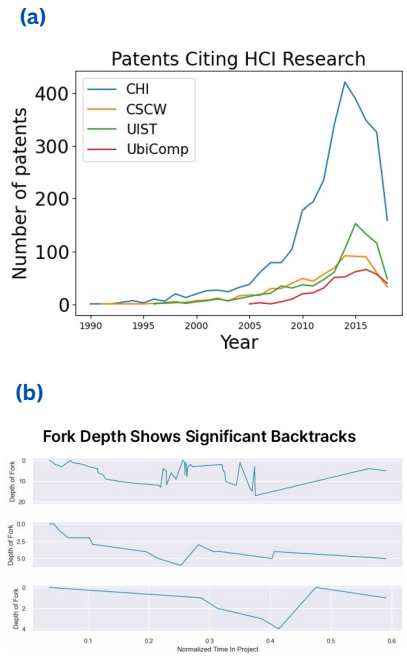
reviewed the predictions and filtered out the false positives. (3) Throughout the annotating phase, we further eliminated images lacking L1/L2L3 descriptions, those shorter than three sentences, or not adhering to alt-text guidelines (e.g., presenting incorrect information), ultimately retaining 10,000 images. These steps ensured that our corpus was semantically rich and included longer author-written descriptions. A comparison analysis was conducted to verify this, as illustrated in Table 12. We randomly split our dataset into training, validation, and test sets using chart IDs to prevent data leakage across sets, resulting in an approximate 80:10:10 ratio. Next, we discuss our dataset annotation process.

Data Annotations and Properties For each image, we recorded the paper's DOI, the figure number, and both the image caption and its alt-text. To annotate descriptions, we followed the protocol outlined by Lundgard, A., et al [117]. In a given batch of 300 images, each description was semantically tagged using 10 attributes keys as seen in Figure 32. GPT-4.0[5] was then employed to tag the remaining descriptions. Each tagged result underwent verification by our annotators. While our primary aim in using these semantic tags was to facilitate our pretext tasks, the annotations can also be useful for analyzing the data structure and enhancing accessibility by identifying missing attributes. The AltChart dataset encompasses a range of eight chart types: line, bar, area, scatter, multivariate, panel, pie, and box charts. For clarification, multivariate and panel charts represent two new categories not previously addressed in earlier benchmarks. Multivariate charts refer to those displaying more than one data type (e.g., combining lines and bars), while panel charts (Figure 32-b) are a collection of multiple charts within a single figure sharing common elements, such as a unified legend or axis.

5.3.4 *Capturing Context*

Building a pretrained multimodal foundation model typically involves two steps. First, textual inputs are encoded using a language model such as T5 [154], BERT [42], or recent architectures like Llama 2.0 [175]. Second, a vision encoder processes the input image, which may focus on parts of images [32], by using Fas-

5 https://openai.com/gpt-4



<L1> This <chart_type> line plot </chart_type>, titled <chart_title> "Patents Citing HCI Research," </chart_title> represents <y_title> the number of patents citing HCI research </y_title> from the <x_encoding> years </x_encoding> <x_range> 1990 to 2015 </x_range> across four key venues: <legend> CHI, CSCW, UIST, and UbiComp </legend>**</L1>**. **<L2L3>**<statistics> Within the last five years, 2010-2015, the citations have significantly increased across all venues </statistics>. <extrema> CHI is highlighted as the most cited venue, with 400 citations in 2015, followed by UIST with 150 in the same year </extrema>. This suggests an implicit correlation: papers from these venues that are cited by patents often achieve a higher number of academic citations within the HCI community, indicating their significant impact and relevance in both academic research and practical applications.**</L2L3>**

<L1> A plot labeled <chart_title> "Fork Depth Shows Significant Backtracks" </chart_title> shows <chart_type> three horizontally stacked line graphs </chart_type> with a <encoding> shared X axis </encoding> labeled <x_title> "normalized time in project" </x_title> which ranges <x_range> from 0 to 1 </x_range>. <y_title> The Y axis is labeled "Depth of Fork" </y_title>, which is the distance from the root node of the version tree, and ranges <y_range> from 0 to 20, 0 to 5, and 0 to 4 on each of the graphs </y_range>**</L1>**. **<L2L3>** <statistics> All three graphs show a steadily decreasing line with sharp increases </statistics>. <comparison> The top image is noticeably more scattered than the other two </comparison>.**</L2L3>**

Text Semantics

- **L1**
  - **Chart type**
  - **Title**
  - **Range**
  - **Encoding**
  - **Legend**
- **L2L3**
  - **Statistics**
  - **Extrema**
  - **Comparison**

Figure 32: Two chart samples from *AltChart* with their annotated summaries. Semantics are indicated by a color code, where <semantic-name> marks the beginning and </semantic-name> marks the end of the semantic segment.

tRCNN [158], or use recent, larger transformers (e.g, VIT [44]) that encode the entire image [103]. These models are initially trained on extensive web content for comprehension tasks such as text-to-text [57] and image-to-text [153] are mainly trained with natural images. To adapt these models for specific domain tasks, a second fine-tuning iteration is often necessary to ensure that the model develops a meaningful latent space for the targeted task. For instance, Donut [94] introduced an OCR-free Transformer, trained end-to-end for document understanding. Subsequently, Nougat [17] fine-tuned Donut, making it effective for converting academic documents into markdown language. Charts, however, present a unique challenge compared to natural images or textual documents. The complexity of user questions often involves sophisticated mathematical calculations. As a result, multimodal foundation models often struggle when addressing tasks related to charts [53].

Recent works have addressed chart-related tasks using various techniques. Some approaches involve modifying the architecture by developing adapters to interpret charts, while others introduce more comprehensive benchmarks for fine-tuning. Matcha [109] builds upon Pix2Struct [99], incorporating numerical reasoning knowledge into the image-to-text model by learning from textual math datasets. UniChart [120] employs a visual instruction tuning approach [115] and fine-tunes the Donut base model with real charts for multiple low-level tasks (e.g., extracting table data) and high-level tasks (e.g., generating summaries). Unlike UniChart, a recent model named ChartLlama [70], based on LLaVA-1.5 [112], proposes an extensive chart-related benchmark leveraging GPT-4. This benchmark is synthetically created with multiple steps to ensure high quality.

Although these models generate generally appealing outputs for sighted individuals, they raise significant concerns regarding accessibility. It is important to remember that while a summary accessible to blind individuals is also accessible to sighted individuals, the reverse is not necessarily true. J. Tang and Bogust et al. [173] were the first to experiment with the abilities of VLMs to generate accessible summaries, but only with synthetically generated charts. In contrast, our proposal, AltChart, ensures that our pretraining corpus comprises real charts from accessible resources, which are semantically rich for everyone.

The aforementioned state-of-the-art approaches mainly follow a similar strategy, extending the size of the pretraining corpus, which leads to higher performance on specific benchmarks but also tends to suffer from catastrophic forgetting [30, 192] and lacks consistent summary structure among similar visual inputs. We instead question whether "less can be more." We demonstrate in our work how pretraining vision encoders with multiple pretext tasks such as, classification and colorization, can achieve state-of-the-art performance. Pretext tasks have already shown promising performance with vision models in previous studies [86, 163]. Furthermore, Pretext tasks could enable the use of other annotations format that were previously not possible to train with VLMs, such as segmentation masks. We also believe that pretext tasks could help us address challenging samples (e.g., those with high loss values) with simpler substream tasks.

Pretext Training. Although larger datasets may improve performance, as demonstrated by several SOTA works, the critical factor is how effectively the model learns from this synthetic data. The capability of these models to handle the variations and complexities of real-world charts remains a challenging issue. With this in mind, we pose a question: **Can we improve the vision encoder hidden representations to a degree that minimizes our reliance on synthetic data?** To address this, our approach leverages pretext tasks to guide the vision encoder in capturing essential covariant and invariant chart features, thereby reducing hallucinations in descriptions. Next, we discuss the details of our pretext task implementations, as outlined in Figure 33.

Chart Pretext Tasks. An effective feature extraction process should include both covariant and invariant features [127]. Covariant features, which adapt to transformations such as scaling or rotation (e.g., vertical axis labels, font sizes), enhance vision encoders' ability to recognize objects despite spatial changes. Meanwhile, invariant features maintain consistency by capturing key characteristics that remain constant across various scenarios (e.g., multi-line charts often include legends). This duality is crucial to ensure a comprehensive and reliable interpretation of chart images. These features, covariant and invariant, are developed through pretext tasks using both self-supervised and supervised training methods, respectively.

Self-Supervised Tasks. is to learn image representations directly from pixels, without relying on predefined semantic annotations. This process typically involves applying transformations to input images and training sub-models to predict the properties of the transformation. In this work, we have chosen three traditional tasks, namely:

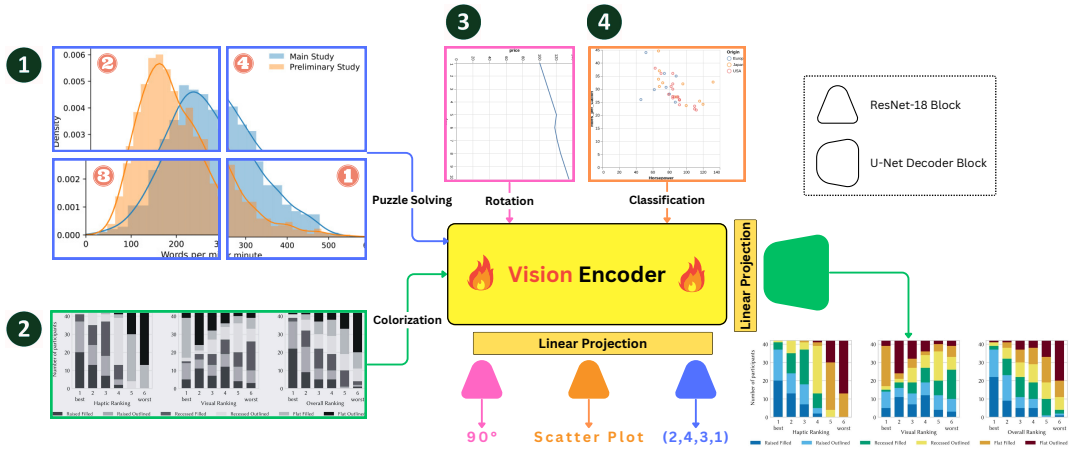


Figure 33: Overview of our vision encoder's training approach, starting from the top-left with tasks including puzzle solving, colorization, rotation, and classification. Sample outputs for each corresponding task are shown on the bottom-right of the figure.

1. **Rotation Prediction [62]**: rotating chart images by various degrees and having a sub-model to predict the angle of rotation. This helps in learning the orientation and geometry of chart components.

2. **Jigsaw Puzzle Solving [140]**: scrambling charts to multiple segments and training a sub-model to reorder them correctly. This teaches the model about the spatial relationships within chart elements.

3. **Colorization [zhang2016colorful]**: feeding grayscale chart images into a sub-model for colorization. This helps in developing features that distinguish between chart components (e.g., different pies in a pie chart).

Supervised Tasks. defined as the utilization of a small amount of labelled data to capture consistent representations. Given the chart analysis topic, different datasets provide varying types of annotations. For example, AltChart and Vistext lack pixel-level semantic annotations, while the Chart-to-Text benchmark offers bounding box annotations for text. To ensure applicability across all benchmarks, we leverage the classification of chart categories task. However, one may explore additional approaches, such as masked element identification, when segmentations or bounding boxes are available.

5.3.5 *Implementation Details*

All pretext tasks need to be evaluated against quantitative metrics to ensure their effectiveness. To facilitate this evaluation, we employ a shallow backbone network complemented by a task-specific head. In the following sections, we will detail the transformation functions, sub-models, and our formation of the loss function.

Transformation Functions. Given an image I, we apply the transformation functions $g(.)$, $f(.)$ and $h(.)$ to generate a transformed image and corresponding ground-truth labels for rotation, puzzle solving, and colorization tasks, respectively.

For rotation, the image undergoes one of four rotational levels: 0°, 90°, 180°, and 270°. This process follows the methodology described in [62], where the network is tasked with identifying the correct rotation angle. In the puzzle-solving task, the image is divided into a 3x3 grid, resulting in nine 64x64 pixel patches. To avoid overfitting, each patch's location is randomly jittered by up to seven pixels, in line with the approach of N. Mehdi et al. [140], thus creating nine distinct 64x64 pixel tiles. We define 100 possible permutations (puzzle configurations), each associated with a unique index. The function $f(.)$ outputs the nine image tiles along with the index of the corresponding permutation, serving as the ground-truth label. The model's objective is to accurately classify the correct permutation index. For the colorization operation, the image is transformed into a grayscale image using the formula $\frac{R+G+B}{3}$ as in [40], and the sub-model is trained to predict the a and b color channels in the CIE Lab color space. The development of the loss function for these operations is discussed in the following section.

Loss Function. three pretexts and one supervised task are present in our formulations. three of which; rotation, puzzle solving, and the supervised tasks end with a softmax activation layer hence we utilized the traditional Cross-Entropy Loss $\mathcal{L}_{rotation}$, $\mathcal{L}_{puzzle}$ and $\mathcal{L}_{categ}$ respectively to output probability values between 0 and 1. For the colorization we utilized the conditional GAN loss $\mathcal{L}_{cGAN}$ with regression mean absolute error loss $\mathcal{L}_{L1}$ as proposed in Pix2Pix [84]. Given N images in a batch, The colorization loss $\mathcal{L}_{color}$ is then computed as follows:

$$\mathcal{L}_{cGAN}(G, D) = \frac{1}{N}\sum_{i=1}^{N} \log D(I_g^i, I_{ab}^i) + \log\left(1 - D(I_g^i, G(I_g^i, z))\right) \tag{1}$$

$$\mathcal{L}_{L1}(G) = \frac{1}{N}\sum_{i=1}^{N} |I_{ab}^i - G(I_g^i, z)| \tag{2}$$

$$\mathcal{L}_{color} = \mathcal{L}_{cGAN}(G, D) + \alpha\mathcal{L}_{L1}(G) \tag{3}$$

In Equation 1, the generator, G, takes a grayscale image I_g^i and produces a 2-channel image I_{ab}. The discriminator, D, concatenates both images to decide whether they are fake or real. It's important to note that both models are conditioned on the grayscale image, meaning the noise vector is omitted [84]. The mean absolute error $\mathcal{L}_{L1}(G)$ in Equation 2 aims for pixel-wise comparison between the generated image and the ground truth to introduce a form of self-supervision. Since L1 has been shown to produce contrastive results , the parameter α is introduced to balance its overall impact. Our final loss $\mathcal{L}_{total}$ is formed as follows:

$$\mathcal{L}_{total} = \gamma_1\mathcal{L}_{color} + \gamma_2\mathcal{L}_{rotation} + \gamma_3\mathcal{L}_{puzzle} + \gamma_4\mathcal{L}_{categ} \tag{4}$$

In calculating the total loss, we sum each one with its respective gamma parameter γ_{1-4}, allowing us to fine-tune their contributions.

Convolutional network. In our experiments, we utilize a ResNet-18 network [74]. Two modifications are applied: (1) The standard 3-channel convolutional layer input is adjusted to align with the vision encoder output's shape (2) After the final FC layer, following the final FC (Fully Connected) layer, we implement average pooling and linearly project the output to match the number of classes for each specific pretext task. For the colorization task, a U-Net decoder [161] is concatenated to the vision encoder to function as the generator. The discriminator is again a ResNet-18.

Hyperparameters. For our pretext tasks, we use a default image resolution of 224×224, unless specified otherwise. We set the parameter $\alpha = 100$ in Equation 3 and an equal impact for all losses, $\gamma_{1-4} = 0.25$ in Equation 4. Both the ResNet and the U-Net decoder are initialized with their pre-trained weights, obtained from MMpretrain[6] and MMsegmentation[7] respectively.

5.3.6 *Experiments*

In this section, we conduct both quantitative and qualitative comparisons of four SOTA methods against AltChart.

Experimental Setups. In this study, we conducted all training processes on a cluster equipped with four NVIDIA-40 GPUs. We utilized publicly available source code from GitHub for each model. Initially, all models were trained on the Chart-2-Text dataset. We then fine-tuned these pre-trained models on the other datasets listed in Table 13 (Vistext and AltChart), with the exception of our baseline model, which is described subsequently. For fine-tuning, the training epochs were set to 5, and the LoRa adapter was employed. We did not alter the input resolution for any of the approaches from their initial configurations. Each experimental run took approximately 8-10 hours to complete.

Baselines & Evaluation Metrics. We compare our model against four baselines: (1) Vistext [173], a VL-T5-based model that achieves SOTA results in generating accessible chart summaries. (2) Matcha [109], an adaptation of Pix2Struct for charts, pre-trained on mathematical reasoning and chart data extraction tasks. (3) UniChart [120], a model based on Donut [94], further pre-trained on multiple chart analysis tasks, achieving SOTA on Chart-2-Text [93] and ChartQA [121]. (4) ChartLLama [70], a fine-tuned LLaVA 1.5 [112] model trained on a large chart corpus synthetically generated with GPT-4.

To conduct the comparison on our benchmarks, we first reproduced and ran inference with each of the baseline models, evaluating their summarization performance using the BLEU4 score [146]. Furthermore, given that the BLEU4 score [146] primarily focuses on n-gram matching between the generated and reference texts, it may overlook essential aspects such as semantic similarity, informativeness, and factual correctness [162]. Hence, we also performed a qualitative evaluation and error analysis of the outputs.

6 https://github.com/open-mmlab/mmpretrain
7 https://github.com/open-mmlab/mmsegmentation

Training Details. As a baseline, we utilized the Donut model, primarily chosen for its relatively low number of parameters, scaling to millions, in contrast to the LLaVA model, which scales to billions. Initially, we employed the base Donut weights, pretrained for text reading tasks as an alternative to OCR engines. These base weights cannot comprehend chart images. We initially train the transformer encoder for 3 epochs on pretext tasks as previously described, followed by training the entire model (vision+language components) for an additional 2 epochs on summarization tasks.

Table 13: Results of state-of-the-art methods on three datasets of chart summarization are presented, with BLEU4 as the evaluation metric. The number of training parameters is also reported.

Model	#Params	VisText			Chart-2-Text			AltChart		
		L1	L2/L3	avg.	Pew	Statista	avg.	L1	L2/L3	avg.
Vistext - image guided [173]	224M	9.0	2.0	5.5	14.2	44.2	29.2	-	-	-
Matcha [108]	282M	6.0	4.0	5.0	12.2	39.4	25.8	16.5	8.0	12.2
Unichart [120]	201M	6.3	5.2	5.75	12.4	38.2	25.3	22.7	13.9	18.3
ChartLLama - 13B [70]	500M	35.0	6.0	20.5	14.2	40.7	27.45	35.0	14.2	24.6
Ours (AltChart)	**180M**	**37.6**	5.6	**21.6**	**15.1**	**46.0**	**30.55**	**44.1**	**14.6**	**29.3**

Comparison of State-of-the-Art Models. Our novel chart summarization model achieves state-of-the-art performance on diverse datasets, pushing the boundaries of both efficiency and quality. As shown in Table 13, we evaluate our method against different chart models, such as MatCha, UniChart, and ChartLlama on three datasets: VisText, Chart-2-text, and our proposed AltChart. Each dataset offers unique challenges, with VisText and AltChart testing different summarization levels and Chart-2-text utilizing data from two distinct sources: Pew and Statista. Despite having fewer trainable parameters (180M) than other text-based models, our approach boasts superior efficiency and adaptability. It readily applies to different benchmarks and swiftly trains for downstream tasks.

Apart from the significant efficiency, our method has outstanding performance, yielding state-of-the-art scores on all three datasets. For instance, on VisText, it delivers a remarkable 37.6% score on L1 test, outperforming ChartLlama by 2.6%, and achieves exceptional results on L2/L3 test. Similar success manifests in the Chart-2-text dataset, where it scores 15.1% and 46.0% across the two data sources. It shows the genearlizability of the proposed method. Our proposed AltChart dataset unveils the potential of our method. In the structured summarization test, it surpasses the previous best score by 9.06% (reach 44.1% on L1) and obtains a best 14.6% on the more challenging L2/L3 test. These remarkable improvements across diverse datasets showcase the effectiveness and generalizability of our pre-training method. Our model sets a new standard for both efficiency and quality.

Qualitative Analysis. To present the qualitative results of the proposed AltChart benchmark, we conduct a visualization comparison in Fig. 34. The visualization showcases a chart summarization example from the proposed AltChart dataset, comparing our method with other models like UniChart and ChartLlama. Each case displays the input chart image followed by summaries generated by all three

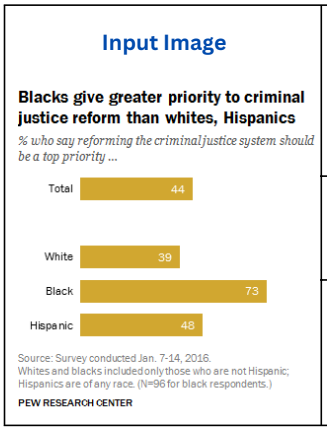
Input Image		
	UniChart	According to a survey conducted by the Pew Research Center, blacks give greater priority to criminal justice reform than whites. Specifically, 73% of blacks prioritize criminal justice reform as a top priority, while only 44% of whites prioritize it. Hispanics prioritize criminal justice reform as a top priority, while only 48% of Hispanics prioritize it as a priority. The chart shows that the majority of blacks. The chart reveals a top priority, with only 39% of them as a top priority, where Hispanics, where Hispanics priorities, where Hispanics priorities, where Hispanics
	ChartLLama	The data indicates a significant divergence in the perception of prioritizing the reform of the criminal justice system among different racial groups. A striking 73% of blacks express the view that this should be a top priority, contrasting with 48% of Hispanics and merely 39% of whites who share the same sentiment.
	AltChart (Ours)	The image displays a horizontal bar chart from a Pew Research Center survey conducted from January 7-14, 2016. The chart, titled "Blacks give greater priority to criminal justice reform than whites, Hispanics." It shows that 73% of Black respondents view criminal justice reform as a top priority, significantly higher compared to 39% of White and 48% of Hispanic respondents. Overall, 44% of the total survey participants consider it a top priority.

Figure 34: Qualitative analysis of chart summarization.

models. On the left side of the visualization is the same input chart image that is a horizontal bar chart, and on the right side from top to bottom are the prediction from UniChart, ChartLLama, and ours. Among them, UniChart made some wrong description about this chart, such as "44% of whites" is incorrect. UniChart struggles in this case, producing repetitive and unclear summaries. ChartLlama also falters, offering basic and incomplete descriptions that miss key details like the total value. In contrast, our method delivers accurate and detailed summaries for the input bar chart. For instance, it can correctly identify the chart type as "horizontal bar chart" and timeframe "January 7-14, 2016". Apart from that, our method can provide insightful descriptions such as "Overall, 44% of the total survey participants consider it a top priority". This comparative analysis demonstrates our model's ability to generate comprehensive summaries, surpassing existing models.

Table 14: Ablation study of the prefix tasks. The metric used is the average BLEU4 score of L1 and L2L3.

Self-supervised	Supervised	Result
✓		27.90
	✓	25.30
✓	✓	**29.35**

Ablation Study. To isolate the impact of our proposed pre-training tasks on chart summarization performance, we conduct an ablation study on the proposed AltChart dataset. As shown in Table 14, we divide the experiments into two groups: self-supervised and supervised training, allowing for clear comparisons between different pre-training paradigms. The reported score is the average of the BLEU4 scores from the L1 and L2/L3 summarizations. Self-supervised training can achieve a score of 27.9%, while supervised training alone can reach 25.3%. This 2.6% gain highlights the effectiveness of self-supervised learning in enriching the model's understanding of charts. Pushing the boundaries further, our combined approach that leverages both self-supervised and supervised training delivered the best score of 29.35% in structured chart summarization.

5.4 CHAPTER CONCLUSION

In this chapter, we address the emerging need for robust assistive technologies. These technologies not only need to meet accessibility guidelines but also ensure robustness against the input variations that PVI may query. For the first time, we present a robustness benchmark for chart analysis models. We conducted a human user study to understand the needs and level of perception of humans, categorizing the results into three levels: easy, mid, and hard.

We then evaluated eight state-of-the-art chart understanding models and demonstrated the importance of considering robustness and consistency in model performance. Our findings highlight that while synthetic sample training is a common trend in chart analysis, it may harm overall robustness, even though it performs well with digital samples due to their high-quality and high-resolution nature.

To address these challenges, we introduced AltChart, a state-of-the-art chart summarization model that generates rich, accessible summaries. Our model employs pretext tasks as a pretraining technique, reducing reliance on synthetic data. This approach allows the use of various annotation formats during training to acquire robust feature representations, which were missing in earlier models. Additionally, we present the AltChart dataset, the largest accessibility-compliant real visualization summarization dataset with rich semantics.

In summary, this chapter represents a significant step towards enhancing the robustness of VLMs in the context of accessibility and assistive technologies. Developing models is the first step in solving emerging problems, but we also recommend conducting robustness benchmarks and improving datasets to address the needs of PVI. Here is a more detailed summary of the contributions of each section in this chapter:

Contribution 1: Introduced CHAOS, a robustness benchmark for chart analysis models. Based on human perception evaluations through a user study, we categorized chart images into three levels: easy, mid, and hard. The emerging need for robustness improvements was highlighted, and suggestions for future research were discussed.

Contribution 2: Proposed the AltChart model, achieving state-of-the-art performance on chart summarization tasks. For the first time, the model utilized pretext tasks before training to equip vision encoders in VLMs with better chart embedding capabilities.

Contribution 3: Released the richest textual alt-text dataset publicly for the vision, NLP, and accessibility communities. The dataset consists of 10k visualization images with rich semantic annotations.

6 CONCLUDING REMARKS

This thesis has advanced the research field of chart analysis to enhance document visual accessibility. Previous approaches have either trained on digital origin data, avoiding real-life cases where PVI might capture charts with noisy cameras, or have not addressed accessibility concerns at all. Our methodological contributions enable networks to perform new tasks for chart comprehension. We demonstrated how supervised learning can guide networks to comply with accessibility standards and guidelines. Additionally, our contributions extend to novel user interfaces and interaction patterns that save labor time and enhance output quality. Here, we summarize the main contributions and open pathways for exploring the accessibility of visualizations and documents.

6.1 IMPACT ON THE FIELD

While AI has made significant strides in chart image analysis, including summarization and visual reasoning, high-level understanding of charts has been limited. In this thesis, we address the problem of enhancing chart comprehension tasks via deep learning models, specifically from the accessibility perspective. We target diverse types of charts frequently used in documents and educational materials, analyzing various structures rich and chart type variations. Our focus is on accessibility for PVI, maintaining the high-quality creation of accessible modalities, namely, alternative texts and tactile materials. These modalities have challenging guidelines that inexperienced users might find difficult to adhere to. To that end, we enhance accessibility on three different yet related levels:

- We first introduce new methods to extract metadata from charts to streamline the conventional process of creating accessible modalities. Among these methods, we propose a line tracing approach for line charts, an interactive display for sighted assistants, and a tactile interface using a bounding box paradigm for PVI. Our approaches were thoroughly evaluated through structured user studies with the target group.
- We propose two new tasks for the chart understanding field: chart retrieval and chart-to-tactile conversion. Both tasks advance the capabilities of VLMs for better adherence to accessibility guidelines.
- We explore the robustness capabilities of chart understanding models. Through a thorough comparison with human perception, we provide insights and recommendations to the community. Notably, our AltChart model achieves state-of-the-art performance in chart summarization tasks.

6.1.1 *New research directions*

Next, we summarize the key contributions made in this thesis:

Better Representation of Deep Learning Outputs. With the increasing requirement for accessibility of documents and visuals, they need to be presented to a wide range of people, including those with visual impairments, who require specific representations such as tactile materials and alt-text. In Section 3.2, we provide insights on how deep learning output paradigms, such as detection bounding boxes and segmentation masks, are valuable not only for extracting metadata but also for creating more comprehensive interactions and insights for PVI. These representations aim to facilitate computer vision research with a greater consideration of different output usability for various use cases.

Conventional Processes to Intelligent Ones. One major challenge for sighted assistants who help convert educational materials is the required expertise and time-consuming conversion process. In this thesis, particularly in Section 3.4, we discuss various approaches to assist this process with deep learning models that offer intelligent interactions, making the process more engaging and less labor-intensive.

High-Quality Alt-Text Authoring. Beyond AI-generated alt-text, we experimented with the image retrieval approach to guide authors in writing high-quality alt-text (see Section 4.2). We collected and annotated valuable visuals available with alt-text from HCI conferences and developed the first image-based chart retrieval model that also facilitates text queries.

VLMs Generating Tactile Materials. For the first time, we considered the task of converting images to tactile formats for document visuals. An end-to-end approach was proposed in Section 4.3 to equip VLMs with the capability to generate accessible SVGs that can be printed or accessed digitally. We formalized this task and provided a publicly available open-source dataset, ChartFormer.

Out-of-Domain Samples Adaptation In real life, we often encounter changes in environmental conditions or sensor types. Research on domain-invariant representations is vital for assistive technologies, which are highly susceptible to distributional shifts. In Section 5, we discuss the problem of low robustness of VLMs in understanding scanned, captured, and new variations of visualizations. These improvements and advancements represent significant steps towards making assistive technologies more robust and accessible, ensuring that they can better serve the needs of diverse user groups, including those with visual impairments.

6.1.2 *Collected Datasets and Benchmarks*

Throughout this thesis, we introduce multiple tasks for chart understanding that were not previously addressed and collect several accompanying datasets. Most prominently, the AltChart dataset encompasses 10K chart images with semantically annotated and rich alt-text collected from HCI conferences. We ensured that the images include a high variety of graphical and structural types.

Additionally, we synthesized the first chart-to-tactile dataset, ChartFormer, using real RGB chart images and synthesized SVG files. This dataset demonstrates how the capabilities of VLMs can be extended beyond image generation to vector graphic material creation.

For the task of chart metadata extraction, we propose the LG dataset, which includes over 500 visualizations with fine-grained semantic segmentation labels for 10 classes. Unlike previous methods that include a single label per pixel, our dataset supports multi-label hierarchical annotations where regions can belong to several semantic categories (e.g., label and x-axis). We demonstrated the possibilities of this dataset with a line tracing system.

Finally, we propose a novel robustness benchmark for the chart analysis field, CHAOS. This evaluation benchmark scheme assesses VLM models on summarization and chart QA tasks and is based on human perception evaluation through a thorough user study. To foster further research on accessibility enhancements, we have made all our datasets publicly available.

6.1.3 *New tools and insights*

In this thesis, we developed three intelligent user interfaces to streamline the creation and transformation of accessibility materials. Two of these interfaces assist sighted individuals in the process, while one is a tactile interface designed for PVI.

We introduced Chart4Blind (see Demo[1]), an interface that utilizes our deep learning backend system to trace chart data and present it in a tactile view before export. This system requires minimal interaction from the sighted user to create educational tactile materials. We also later present the Alt4Blind[2] tool for the alt-text modality. This tool offers a new approach to authoring alt-text by presenting similar charts to the user-uploaded image as references. This method helps in developing good practices and raises awareness about the importance of alt-text.

Finally, we developed Layout4Blind (see Demo[3]), a tactile representation of documents on 2D refreshable tactile displays for PVI. The interface is equipped with intelligent control features integrated with deep learning models in the backend. These features include audio feedback on request to classify and locate document blocks. The interface not only enables users to understand document structures and locate visuals but also assists in quickly skimming and scanning the document, which is crucial when time is limited or doing literature review.

1 https://moured.github.io/chart4blind/
2 https://moured.github.io/alt4blind/
3 https://github.com/moured/accessible-document-layout/tree/main

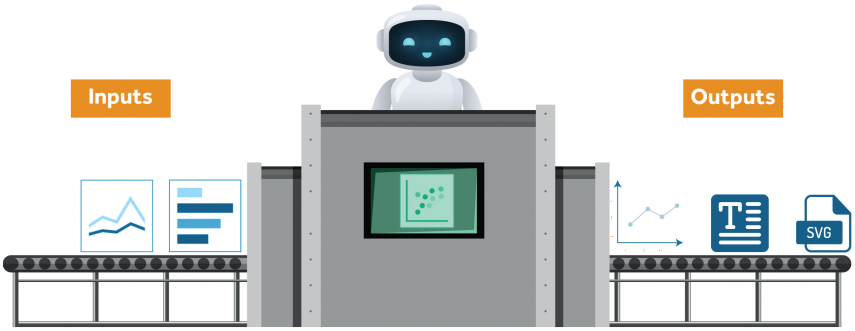


Figure 35: Towards unified accessibility-oriented model.

6.2 OPEN QUESTIONS FOR FUTURE WORK

While models for high-level understanding and especially for question answering have made extraordinary strides in recent years, there is still a wide gap to human performance. Due to the nature of our proposed approaches, we can directly infer the reasoning of our models and, thus, we can easily deduct weak spots in the data and overall architecture. Our thesis opens doors to new methodologies for improving model robustness and accessibility, providing a foundation for future advancements in these areas. Below, we highlight some of the directions we are working on and considering for our future work.

Unified Accessibility Model. In our current work, we focus on investigating tactile and alt-text modalities separately. However, future efforts should aim at developing unified and general understanding models that are more robust and capable of better visual perception. There is a rising need for models that can interact with PVI users and handle different modalities seamlessly. Such a model would be able to manage various user interfaces and actions through APIs and other methods.

Addressing Other Impairments. While blindness and visual impairments are significant sensory deficiencies, other types such as color blindness and hearing impairments should also be considered. Charts are rich with color information, so future work could explore how color information can be effectively represented by VLMs. This includes recolorization, summarization, and chart QA based on required sensory considerations.

Chart Referral. For sighted users, visual question answering is a common way to gain insights from a query image (e.g., "What is the highest value in this line trend?"). Sighted users can easily spot hallucinations and redirect the model with follow-up questions. However, for PVI users, confirming the correctness of results is challenging. We are working on developing a grounding dataset and model for the chart QA task, aiming to provide localization and a way for PVI users to interact with VLMs more effectively.

Part I

APPENDIX

APPENDIX

A.1 *chart4blind* QUESTIONNAIRE

The online questionnaire implemented in SoSciSurvey consisted of five pages: an initial page to confirm that the user has the necessary materials before starting, followed by three pages, one for each chart. All three pages had the same questions, with the only difference being the name of the chart (chartx). The survey concludes with a final goodbye page.

A.1.1 *Start Page*

Thank you for participating in our survey! Please ensure you have downloaded the three files sent to you by email, named chart1.svg, chart2.svg, and chart3.svg. Also, confirm you have received the postal mail containing three embossed papers. Each embossed paper corresponds to one of the digital chart files. You can identify each chart by checking the top left corner, where the chart ID – chart1, chart2, or chart3 – is located.

(Selection Question Type)

- Have you successfully downloaded the three chart files (chart1.svg, chart2.svg, and chart3.svg) sent to your email? (Yes/No)
- Have you received the postal mail with the three embossed papers? (Yes/No)
- Were you able to identify each chart by the chart ID in the top left corner? (Yes/No)

If you encounter any issues with the steps mentioned above, please contact us by sending an email to omar.moured@kit.edu Next, there will be three pages, one for each chart, with two questions on each page. Please proceed by clicking the "Next" button.

A.1.2 *Pages 2-4*

First, use the screen reader to go through 'chartx.svg', and then answer the following question.

(Text Input)

- How would you describe your experience with using a screen reader to access and interpret the information in chart1.svg? Please include details about the ease of navigation, clarity of the alt-text, and any challenges you faced.

Next, examine the printed tactile version of chartx and answer the following question.

- Please describe your experience interacting with the tactile printed chartx. Were you able to effectively perceive and understand both the visual and textual elements? Share any challenges or observations you encountered.

A.1.3 *Goodbye Page*

Thank you for completing this questionnaire! We would like to thank you very much for helping us. Your answers were transmitted, you may close the browser window or tab now.

A.1.4 *Embossed Tactile Images*

The following printed tactile charts were sent to our blind and visually impaired participants for the user study.

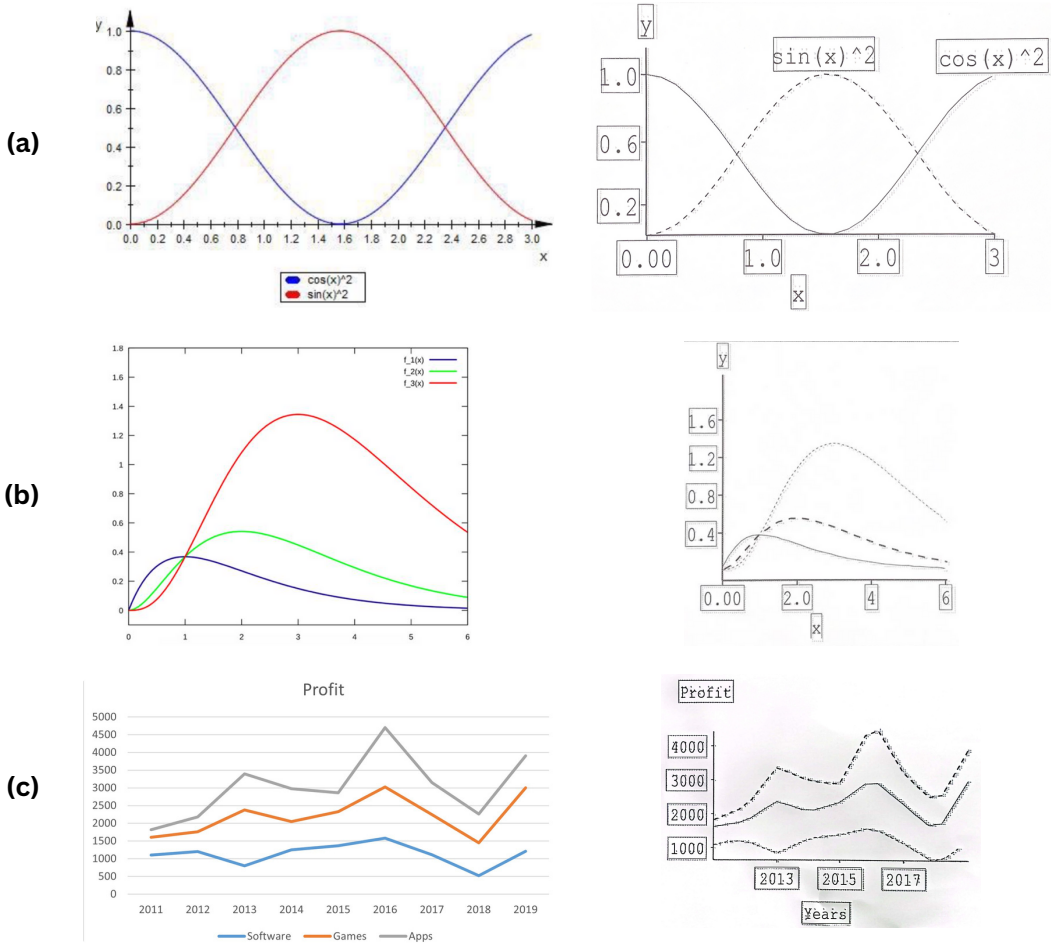


Figure 36: Three printed tactile charts sent to our BVI individuals. The left row displays the original chart images, while the right row presents the tactile versions.

A.2 CHARTFORMER USER STUDY SAMPLES

Figure 37 illustrates the HyperBraille view of the samples generated with the ChartFormer model.

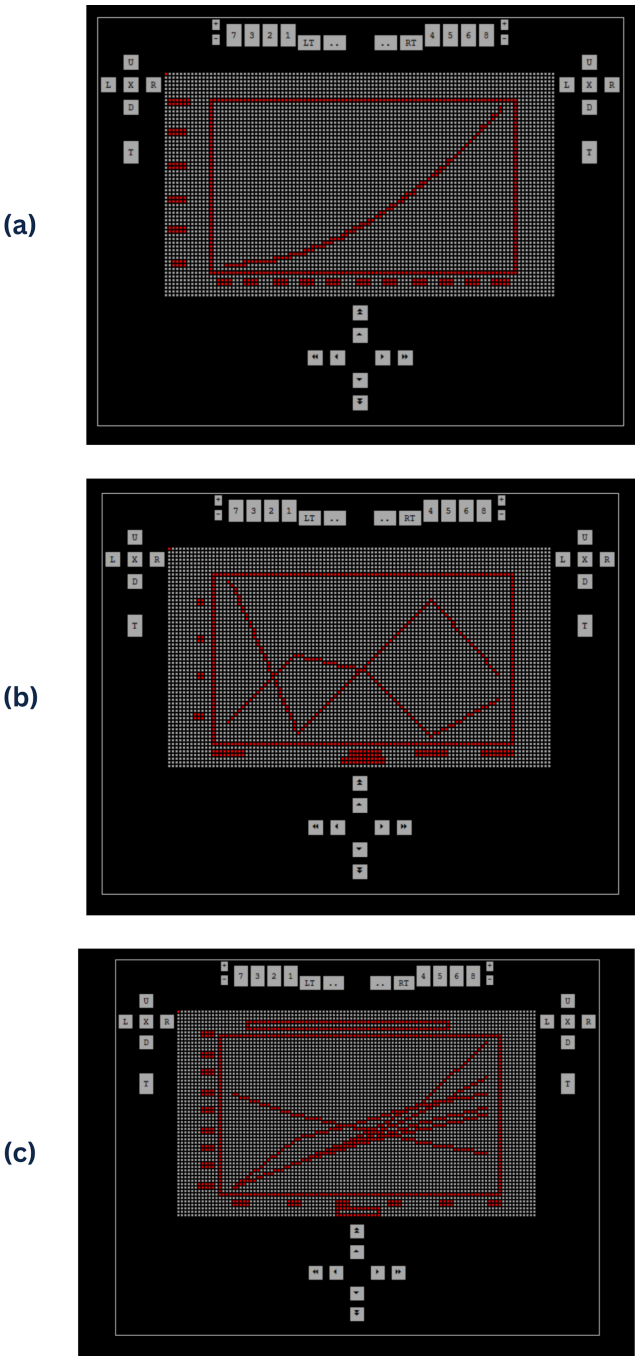
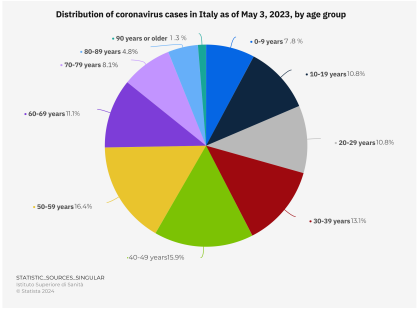


Figure 37: HyperBraille view of user study samples.

A.3 SCANNED & CAPTURED CHART COMPREHENSION WITH CHATGPT4

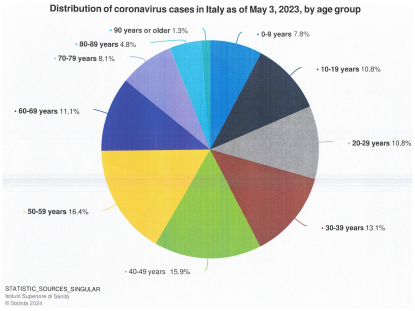
ChatGPT-4 Vision module was used to author alt-text for each of the samples in Figure 38 below. Used prompts: "Author alt-text including the key points of the following chart. Make sure that the alt-text adhere to accessibility guidelines."

(a)



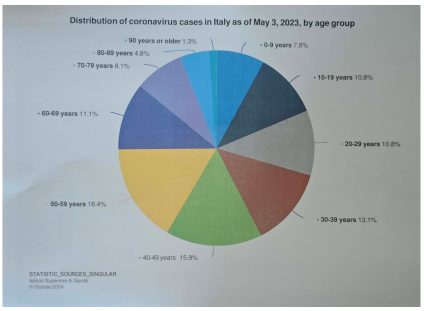
The chart shows the distribution of coronavirus cases in Italy as of May 3, 2023, by age group. The highest percentage of cases is among the 50-59 age group at 16.4%, followed by the 40-49 age group at 15.9%. The lowest percentage is in the 90 years or older age group at 1.3%. Data sources are Istituto Superiore di Sanità and Statista 2024.

(b)



The image is a pie chart titled "Distribution of coronavirus cases in Italy as of May 3, 2023, by age group." The segments are colored differently and labeled with the following age groups and percentages: 0-9 years: 18.2%, 10-19 years: 5.8%, 20-29 years: 10.8%, 30-39 years: 23.1%, 40-49 years: 5.9%, 50-59 years: 16.4%, 60-69 years: 11.1%, 70-79 years: 2.1%, 80-89 years: 24.8%, 90 years or older: 0.3%.

(c)



The chart illustrates the distribution of coronavirus cases in Italy by age group. The highest percentage of cases is in the 50-59 years group, which is highlighted in purple, at 16.4%. The 40-49 years group, shown in orange, follows at 15.9%. The 10-19 years group, shows a higher percentage than expected at 20.8%.

Figure 38: Alt-text created by GPT4 vision module for three same charts collected from different sources, digital, scanned and captured with phone.

A.4 CHART-QA ROBUSTNESS USER STUDY

The online questionnaire implemented in JotForm [1] consisted of 10 image pertubations applied to 10 disntinused charts. The users are [something] with an initial page to confirm that the user has understood the necessary steps before starting, followed tutorial pertubation. Each chart has its own question and users are requested to answer based on the pertubation level they find convinent.

A.4.1 *Start Page*

- **Survey Structure**: You will be shown images affected by 10 different types of perturbations (noise categories). Each type includes 10 levels of noise intensity, starting from level 10 (most perturbed) down to level 1 (least perturbed).
- **Your Task**: Your initial task is to determine if you can comfortably read and clearly understand the image content in a way that allows you to answer any question about it.—respond with "Yes" or "No." If you answer "Yes," a follow-up question about the chart's content will be posed. If you answer "No," the next image displayed will have reduced level of noise, continuing until the original image (without noise). You should stop at the level where you can understand the content; ideally, this should be before reaching level 1.
- **Tutorial**: We'll guide you through a quick tutorial before beginning the actual survey.
- **Privacy and Data Use**: No personal details are required to participate. Please note that if you exit the survey before reaching the final page, your data will not be saved.
- **Need Help?** If you encounter any unclear steps or issues, feel free to reach out to the responsible researchers, Omar Moured, Yufan Chen or Jiaming Zhang at their email addresses.

A.4.2 *Perturbation Pages*

Ten perturbations were experimented with. For each perturbation, we have 10 levels. Users start from the most severe level, level 10, and answer the question "Can you see and understand the chart?" If they click "yes," a follow-up question appears, and they enter their answer. If they click "no," the next lower level (e.g., level 9) is presented. This process continues until they enter their answer, after which the next perturbation is shown. The study takes approximately 15 minutes on average to complete. Below, we show three examples from the total page: "blotches" noise (Figure 39), "color" perturbation (Figure 40) at both level 10, and medium noise, level 5 for elastic transform (Figure 41).

1 https://jotform.com/

Figure 39: A sample bar chart image from the tutorial page with "blotches" noise at level 10.

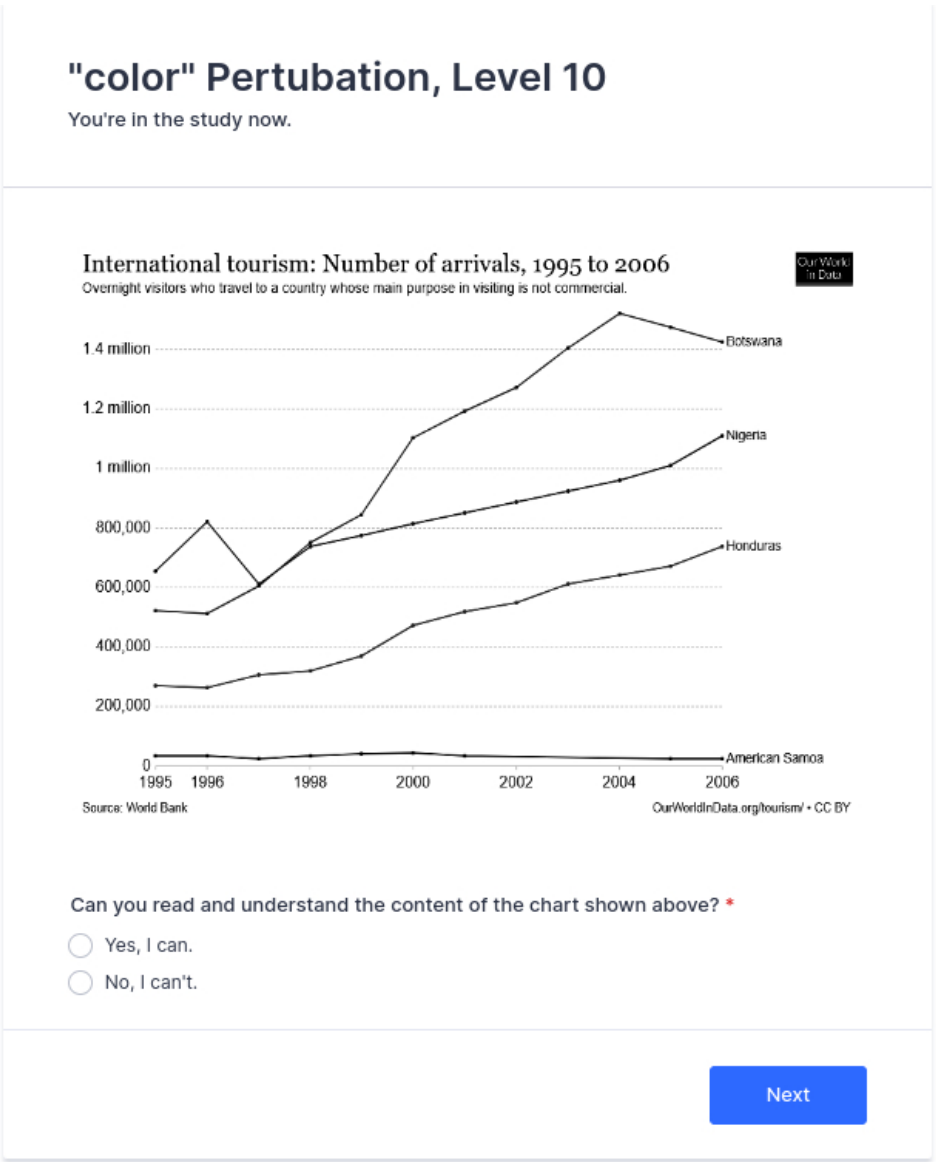


Figure 40: A sample multi-line chart image from the second page with missing "color" information at level 10.

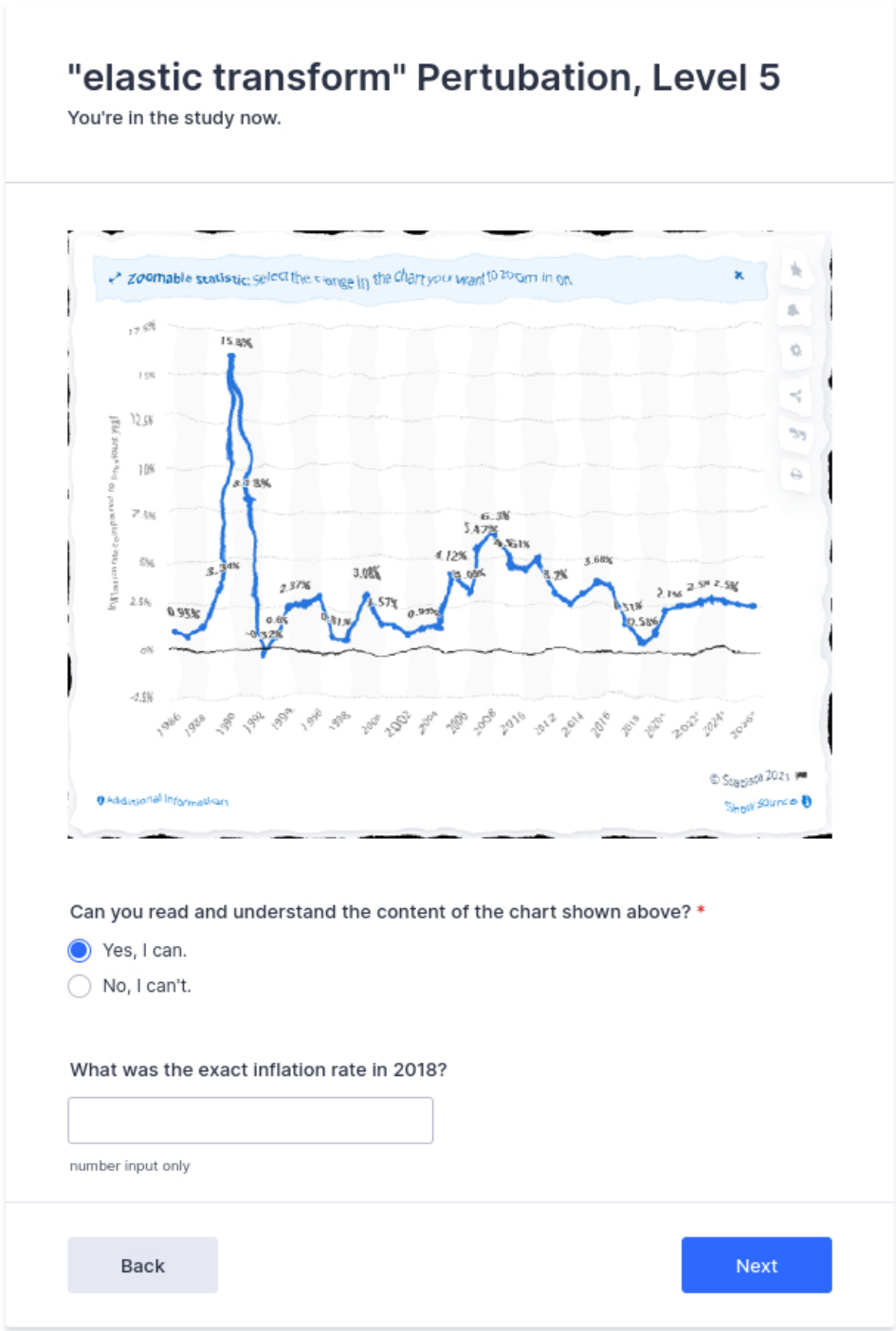


Figure 41: A sample line chart image from the "elastic transform" perturbation at medium level 5.

B

SHORT CV

Omar Moured

LinkedIn: https://www.linkedin.com/in/omar-moured/

Education and Career Highlights

Aug 2021 - Today	**PhD Student in Computer Vision**, *Karlsruhe Institute of Technology*, Topic: *"Towards Accessible Visualizations with Vision-Language Models"*
Sep 2020 - Aug 2021	**Research Assistant**, *Royal Academy of Engineering, UK*, Topic: *"Smart Embedded Camera with Multi-exposure & Multi-view capability for autonomous vehicles."*
Jun 2018 - Aug 2021	**R&D Engineer (part-time)**, *ISSD Elektronik*
Oct 20214 - May 2018	**B.Sc. and M.Sc. in Computer Science**, *Middle East Technical University, Turkey*

Awards and other Achievements

Aug 2021	**Marie-Curie scholarship for early stage researchers**, *INTUITIVE project* https://www.intuitive-itn.eu/,
Jan 2022 - today	**Supervision of student theses**, Successfully supervised computer science students (two Master and two Bachelor) during their final theses projects.
Jun 2014 and 2018	**Ranked among the top 10 bachelor's and master's students with high honors** at *METU*.

C

OWN PUBLICATIONS

This doctoral research resulted in the following peer-reviewed publications, which are incorporated in whole or in part in this thesis.

[1] Omar Moured, Sara Alzalabny, Anas Osman, Thorsten Schwarz, Karin Müller, and Rainer Stiefelhagen. "ChartFormer: A Large Vision Language Model for Converting Chart Images into Tactile Accessible SVGs." In: *Computers Helping People with Special Needs*. Cham: Springer Nature Switzerland, 2024, pp. 299–305.

[2] Omar Moured, Sara Alzalabny, Thorsten Schwarz, Bastian Rapp, and Rainer Stiefelhagen. "Accessible Document Layout: An Interface for 2D Tactile Displays." In: *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*. 2023, pp. 265–271.

[3] Omar Moured, Morris Baumgarten-Egemole, Karin Müller, Alina Roitberg, Thorsten Schwarz, and Rainer Stiefelhagen. "Chart4blind: An intelligent interface for chart accessibility conversion." In: *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 2024, pp. 504–514.

[4] Omar Moured, Shahid Ali Farooqui, Karin Müller, Sharifeh Fadaeijouybari, Thorsten Schwarz, Mohammed Javed, and Rainer Stiefelhagen. "Alt4Blind: A User Interface to Simplify Charts Alt-Text Creation." In: *Computers Helping People with Special Needs*. Cham: Springer Nature Switzerland, 2024, pp. 291–298.

[5] Omar Moured, Jiaming Zhang, Alina Roitberg, Thorsten Schwarz, and Rainer Stiefelhagen. "Line Graphics Digitization: A Step Towards Full Automation." In: *International Conference on Document Analysis and Recognition*. Springer. 2023, pp. 438–453.

[6] Omar Moured, Jiaming Zhang, M Saquib Sarfraz, and Rainer Stiefelhagen. "AltChart: Enhancing VLM-based Chart Summarization Through Multi-Pretext Tasks." In: *arXiv preprint arXiv:2405.13580* (2024). **Accepted at ICDAR 2024, to be presented on September 2nd, 2024.**

[7] Gaspar Ramôa, Omar Moured, Thorsten Schwarz, Karin Müller, and Rainer Stiefelhagen. "Enabling People with Blindness to Distinguish Lines of Mathematical Charts with Audio-Tactile Graphic Readers." In: *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*. 2023, pp. 384–391.

[8] Sara Zalabny, Omar Moured, Karin Müller, Thorsten Schwarz, Bastian Rapp, and Rainer Stiefelhagen. "Touch for Accessibility: Haptic SVG Diagrams for Visually Impaired and Blind Individuals." In: *2024 IEEE Haptics Symposium (HAPTICS)*. IEEE. 2024, pp. 79–84.

BIBLIOGRAPHY

[1] Md Zubair Ibne Alam, Shehnaz Islam, and Enamul Hoque. "SeeChart: Enabling Accessible Visualizations Through Interactive Natural Language Interface For People with Visual Impairments." In: *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 2023, pp. 46–64.

[2] Frances K Aldrich and Linda Sheppard. "Tactile graphics in school education: perspectives from pupils." In: *British Journal of Visual Impairment* 19.2 (2001), pp. 69–73.

[3] Helmut Alt and Michael Godau. "Computing the Fréchet distance between two polygonal curves." In: *International Journal of Computational Geometry & Applications* 5.01n02 (1995), pp. 75–91.

[4] Kerstin Altmanninger and Wolfram Wöß. "Dynamically generated scalable vector graphics (SVG) for barrier-free web-applications." In: *Computers Helping People with Special Needs: 10th International Conference, ICCHP 2006, Linz, Austria, July 11-13, 2006. Proceedings 10*. Springer. 2006, pp. 128–135.

[5] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. "Docformer: End-to-end transformer for document understanding." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 993–1003.

[6] Kosuke Araki, Tetsuya Watanabe, and Kazunori Minatani. "Development of Tactile Graph Generation Software Using the R Statistics Software Environment." In: *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility*. ASSETS '14. Rochester, New York, USA: Association for Computing Machinery, 2014, 251–252. ISBN: 9781450327206.

[7] Christoph Auer, Ahmed Nassar, Maksym Lysak, Michele Dolfi, Nikolaos Livathinos, and Peter Staar. "ICDAR 2023 competition on robust layout segmentation in corporate documents." In: *International Conference on Document Analysis and Recognition*. Springer. 2023, pp. 471–482.

[8] "Automatic document processing: A survey." In: *Pattern Recognition* 29.12 (1996), pp. 1931–1952. ISSN: 0031-3203. DOI: https://doi.org/10.1016/S0031-3203(96)00044-1.

[9] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. "Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities." In: *arXiv preprint arXiv:2308.12966* (2023). URL: https://doi.org/10.48550/arXiv.2308.12966.

[10] Abhijit Balaji, Thuvaarakkesh Ramanathan, and Venkateshwarlu Sonathi. "Chart-text: A fully automated chart image descriptor." In: *arXiv preprint arXiv:1812.10636* (2018).

[11] Hritik Bansal and Aditya Grover. "Leaving Reality to Imagination: Robust Classification via Generated Datasets." In: *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*. 2023.

[12] Safwane Benbba. "Comparison of D3. js and Chart. js as visualisation tools." B.S. thesis. 2021.

[13] Jeffrey P. Bigham, Ryan S. Kaminsky, Richard E. Ladner, Oscar M. Danielsson, and Gordon L. Hempton. "WebInSight: Making Web Images Accessible." In: *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*. Assets '06. Portland, Oregon, USA: Association for Computing Machinery, 2006, 181–188. ISBN: 1595932909.

[14] Galal M Binmakhashen and Sabri A Mahmoud. "Document layout analysis: a comprehensive survey." In: *ACM Computing Surveys (CSUR)* 52.6 (2019), pp. 1–36.

[15] Sanket Biswas, Pau Riba, Josep Lladós, and Umapada Pal. "Beyond document object detection: instance-level segmentation of complex layouts." In: *International Journal on Document Analysis and Recognition (IJDAR)* 24.3 (2021), pp. 269–281.

[16] Matthew Blanco, Jonathan Zong, and Arvind Satyanarayan. "Olli: An extensible visualization library for screen reader accessibility." In: *IEEE VIS Posters* 6 (2022).

[17] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. "Nougat: Neural Optical Understanding for Academic Documents." In: *The Twelfth International Conference on Learning Representations*. 2023.

[18] Yevgen Borodin, Jeffrey P. Bigham, Glenn Dausch, and I. V. Ramakrishnan. "More than Meets the Eye: A Survey of Screen-Reader Browsing Strategies." In: *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*. W4A '10. Raleigh, North Carolina: Association for Computing Machinery, 2010. ISBN: 9781450300452. DOI: 10.1145/1805986.1806005. URL: https://doi.org/10.1145/1805986.1806005.

[19] Braille Authority of North America. *Guidelines and Standards for Tactile Graphics*. http://www.brailleauthority.org/tg/ (accessed on 6 November 2022). 2022.

[20] Egil Bru, Thomas Trautner, and Stefan Bruckner. "Line Harp: Importance-Driven Sonification for Dense Line Charts." In: *arXiv preprint arXiv:2307.16589* (2023).

[21] Matthew Butler, Leona M Holloway, Samuel Reinders, Cagatay Goncu, and Kim Marriott. "Technology Developments in Touch-Based Accessible Graphics: A Systematic Review of Research 2010-2020." In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. Yokohama, Japan: Association for Computing Machinery, 2021.

[22] Xu Canhui, Li Yuteng, Shi Cao, Zhang Honghong, Bi Hengyue, and Chen Yinong. "HiM: hierarchical multimodal network for document layout analysis." In: *Applied Intelligence* 53.20 (2023), pp. 24314–24326.

[23] Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gul Varol, Aude Oliva, Vicente Ordonez, Rogerio Feris, et al. "Going beyond nouns with vision & language models using synthetic data." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 20155–20165.

[24] Diagram Center. *General Guidelines*. 2019. URL: http://diagramcenter.org/general-guidelines-final-draft.html#4.

[25] Diagram Center. *Specific Guidelines: Art, Photos & Cartoons*. 2019.

[26] Arindam Chaudhuri, Krupa Mandaviya, Pratixa Badelia, Soumya K Ghosh, Arindam Chaudhuri, Krupa Mandaviya, Pratixa Badelia, and Soumya K Ghosh. *Optical character recognition systems*. Springer, 2017.

[27] Kai Chen, Mathias Seuret, Marcus Liwicki, Jean Hennebert, and Rolf Ingold. "Page segmentation of historical document images with convolutional autoencoders." In: *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. 2015, pp. 1011–1015.

[28] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. "MMDetection: Open mmlab detection toolbox and benchmark." In: *arXiv preprint arXiv:1906.07155* (2019).

[29] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. "Encoder-decoder with atrous separable convolution for semantic image segmentation." In: *ECCV*. 2018.

[30] Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. "Recall and Learn: Fine-tuning Deep Pretrained Language Models with Less Forgetting." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 7870–7881.

[31] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. "Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks." In: *Transactions on Machine Learning Research* (2023).

[32] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. "Uniter: Universal image-text representation learning." In: *European conference on computer vision*. Springer. 2020, pp. 104–120.

[33] Yufan Chen, Jiaming Zhang, Kunyu Peng, Junwei Zheng, Ruiping Liu, Philip Torr, and Rainer Stiefelhagen. "RoDLA: Benchmarking the Robustness of Document Layout Analysis Models." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 15556–15566.

[34] Zhi-Qi Cheng, Qi Dai, and Alexander G Hauptmann. "Chartreader: A unified framework for chart derendering and comprehension without heuristic rules." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 22202–22213.

[35] Sanjana Shivani Chintalapati, Jonathan Bragg, and Lucy Lu Wang. "A dataset of alt texts from HCI publications: Analyses and uses towards producing more descriptive alt texts of data visualizations in scientific papers." In: *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. 2022, pp. 1–12.

[36] Jinho Choi, Sanghun Jung, Deok Gun Park, Jaegul Choo, and Niklas Elmqvist. "Visualizing for the non-visual: Enabling the visually impaired to use visualization." In: *Computer Graphics Forum*. Vol. 38. 3. Wiley Online Library. 2019, pp. 249–260.

[37] Christopher Clark and Santosh Divvala. "Pdffigures 2.0: Mining figures from research papers." In: *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. 2016, pp. 143–152.

[38] World Wide Web Consortium et al. "Web content accessibility guidelines 1.0." In: (1999).

[39] Kenny Davila, Bhargava Urala Kota, Srirangaraj Setlur, Venu Govindaraju, Christopher Tensmeyer, Sumit Shekhar, and Ritwick Chaudhry. "ICDAR 2019 competition on harvesting raw tables from infographics (chart-infographics)." In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. 2019, pp. 1594–1599.

[40] Aditya Deshpande, Jason Rock, and David Forsyth. "Learning large-scale automatic image colorization." In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 567–575.

[41] Heather Devine, Andres Gonzalez, and Matthew Hardy. "Making accessible PDF documents." In: *Proceedings of the 11th ACM symposium on Document engineering*. 2011, pp. 275–276.

[42] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[43] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. "Benchmarking adversarial robustness on image classification." In: *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 321–331.

[44] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." In: *International Conference on Learning Representations*. 2020.

[45] Xuefeng Du, Xin Wang, Gabriel Gozum, and Yixuan Li. "Unknown-aware object detection: Learning what you don't know from videos in the wild." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 13678–13688.

[46] Bernhard Dürnegger, Christina Feilmayr, and Wolfram Wöß. "Guided Generation and Evaluation of Accessible Scalable Vector Graphics." In: *Computers Helping People with Special Needs*. Ed. by Klaus Miesenberger, Joachim Klaus, Wolfgang Zagler, and Arthur Karshmer. Springer Berlin Heidelberg, 2010, pp. 27–34.

[47] Polly Edman. *Tactile graphics*. American Foundation for the Blind, 1992.

[48] Christin Engel, Emma Franziska Müller, and Gerhard Weber. "SVGPlott: an accessible tool to generate highly adaptable, accessible audio-tactile charts for and from blind and visually impaired people." In: *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. 2019, pp. 186–195.

[49] Christin Engel, Emma Franziska Müller, and Gerhard Weber. "SVGPlott: an accessible tool to generate highly adaptable, accessible audio-tactile charts for and from blind and visually impaired people." In: PETRA '19. 2019.

[50] Christin Engel and Gerhard Weber. "Improve the accessibility of tactile charts." In: *Human-Computer Interaction-INTERACT 2017: 16th IFIP TC 13 International Conference, Mumbai, India, September 25–29, 2017, Proceedings, Part I 16*. Springer. 2017, pp. 187–195.

[51] Christin Engel and Gerhard Weber. "A user study to evaluate tactile charts with blind and visually impaired people." In: *Computers Helping People with Special Needs: 16th International Conference, ICCHP 2018, Linz, Austria, July 11-13, 2018, Proceedings, Part II 16*. Springer. 2018, pp. 177–184.

[52] European Blind Union. *Facts and Figures*. Accessed: 2024-06-04. 2024. URL: https://www.euroblind.org/about-blindness-and-partial-sight/facts-and-figures.

[53] Ali Mazraeh Farahani, Peyman Adibi, Alireza Darvishy, Mohammad Saeed Ehsani, and Hans-Peter Hutter. "Automatic chart understanding: a review." In: *IEEE Access* (2023).

[54] Ali Mazraeh Farahani, Peyman Adibi, Mohammad Saeed Ehsani, Hans-Peter Hutter, and Alireza Darvishy. "Automatic Chart Understanding: A Review." In: *IEEE Access* 11 (2023), pp. 76202–76221.

[55] Hao Feng, Shaokai Liu, Jiajun Deng, Wengang Zhou, and Houqiang Li. "Deep unrestricted document image rectification." In: *IEEE Transactions on Multimedia* (2023).

[56] Jon Ferraiolo, Fujisawa Jun, and Dean Jackson. *Scalable vector graphics (SVG) 1.0 specification*. iuniverse Bloomington, 2000.

[57] Luciano Floridi and Massimo Chiriatti. "GPT-3: Its nature, scope, limits, and consequences." In: *Minds and Machines* 30 (2020), pp. 681–694.

[58] The Document Foundation. *LibreOffice Draw*. 2020. URL: https://www.libreoffice.org/discover/draw/.

[59] Jiayun Fu, Bin B Zhu, Haidong Zhang, Yayi Zou, Song Ge, Weiwei Cui, Yun Wang, Dongmei Zhang, Xiaojing Ma, and Hai Jin. "Chartstamp: Robust chart embedding for real-world applications." In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 2786–2795.

[60] Gillian Gale. "Guidelines on Conveying Visual Information." In: *Round Table on Information Access for People with Print Disabilities Inc. Available from http://printdisability. org/guidelines/guidelines-on-conveying-visual-information-2005* (2005).

[61] Andrea Gemelli, Simone Marinai, Lorenzo Pisaneschi, and Francesco Santoni. "Datasets and annotations for layout analysis of scientific articles." In: *International Journal on Document Analysis and Recognition (IJDAR)* (2024), pp. 1–23.

[62] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. *Unsupervised Representation Learning by Predicting Image Rotations*. 2018. arXiv: 1803.07728 [cs.CV].

[63] Max Göbel, Tamir Hassan, Ermelinda Oro, and Giorgio Orsi. "ICDAR 2013 table competition." In: *2013 12th international conference on document analysis and recognition*. IEEE. 2013, pp. 1449–1453.

[64] A Jonathan R Godfrey and M Theodor Loots. "Statistical software (R, SAS, SPSS, and Minitab) for blind students and practitioners." In: *Journal of Statistical Software* 58 (2014), pp. 1–25.

[65] Cagatay Goncu and Kim Marriott. "Tactile Chart Generation Tool." In: *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility*. Assets '08. Halifax, Nova Scotia, Canada: Association for Computing Machinery, 2008, 255–256. ISBN: 9781595939760.

[66] Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, and Liqing Zhang. "Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 4583–4592.

[67] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. "SegNeXt: Rethinking Convolutional Attention Design for Semantic Segmentation." In: *arXiv preprint arXiv:2209.08575* (2022).

[68] Yong Guo, David Stutz, and Bernt Schiele. "Robustifying token attention for vision transformers." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 17557–17568.

[69] Martin Hahner, Christos Sakaridis, Mario Bijelic, Felix Heide, Fisher Yu, Dengxin Dai, and Luc Van Gool. "Lidar snowfall simulation for robust 3d object detection." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16364–16374.

[70] Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. "Chartllama: A multimodal llm for chart understanding and generation." In: *arXiv preprint arXiv:2311.16483* (2023).

[71] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. "Evaluation of deep convolutional nets for document image classification and retrieval." In: *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. 2015, pp. 991–995.

[72] Monica Haurilet, Ziad Al-Halah, and Rainer Stiefelhagen. "Spase-multi-label page segmentation for presentation slides." In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2019, pp. 726–734.

[73] Monica Haurilet, Alina Roitberg, Manuel Martinez, and Rainer Stiefelhagen. "Wise—slide segmentation in the wild." In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. 2019, pp. 343–348.

[74] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[75] Thomas Hermann, Andy Hunt, John G Neuhoff, et al. *The sonification handbook*. Vol. 1. Logos Verlag Berlin, 2011.

[76] Leona M Holloway, Cagatay Goncu, Alon Ilsar, Matthew Butler, and Kim Marriott. "Infosonics: Accessible infographics for people who are blind using sonification and voice." In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–13.

[77] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. "Searching for mobilenetv3." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1314–1324.

[78] Anwen Hu, Shizhe Chen, and Qin Jin. "Question-controlled text-aware image captioning." In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 3097–3105.

[79] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. "mplug-docowl 1.5: Unified structure learning for ocr-free document understanding." In: *arXiv preprint arXiv:2403.12895* (2024).

[80] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. "Layoutlmv3: Pre-training for document ai with unified text and image masking." In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 4083–4091.

[81] Joseph A Huwaldt and S Steinhorst. *Plot digitizer*. 2023. URL: http://plotdigitizer.sourceforge.net.

[82] Sanasam Inunganbi. "A systematic review on handwritten document analysis and recognition." In: *Multimedia Tools and Applications* 83.2 (2024), pp. 5387–5413.

[83] Tatsuya Ishihara, Hironobu Takagi, Takashi Itoh, and Chieko Asakawa. "Analyzing Visual Layout for a Non-Visual Presentation-Document Interface." In: *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*. Assets '06. Portland, Oregon, USA: Association for Computing Machinery, 2006, 165–172. ISBN: 1595932909. DOI: 10.1145/1168987.1169016. URL: https://doi.org/10.1145/1168987.1169016.

[84] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. *Image-to-Image Translation with Conditional Adversarial Networks*. 2018. arXiv: 1611.07004 [cs.CV].

[85] JaidedAI. *EasyOCR*. https://github.com/JaidedAI/EasyOCR. 2021.

[86] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. "A survey on contrastive self-supervised learning." In: *Technologies* 9.1 (2020), p. 2.

[87] Chandrika Jayant, Matt Renzelmann, Dana Wen, Satria Krisnandi, Richard Ladner, and Dan Comden. "Automated Tactile Graphics Translation: In the Field." In: *Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility*. Assets '07. Tempe, Arizona, USA: Association for Computing Machinery, 2007, 75–82. ISBN: 9781595935731.

[88] Patrick W Jordan, Bruce Thomas, Ian Lyall McClelland, and Bernard Weerdmeester. *Usability evaluation in industry*. CRC Press, 1996.

[89] Daekyoung Jung, Wonjae Kim, Hyunjoo Song, Jeong-in Hwang, Bongshin Lee, Bohyoung Kim, and Jinwook Seo. "ChartSense: Interactive Data Extraction from Chart Images." In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA: Association for Computing Machinery, 2017, 6706–6717. ISBN: 9781450346559.

[90] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. "Dvqa: Understanding data visualizations via question answering." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 5648–5656.

[91] Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Ko Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. "Opencqa: Open-ended question answering with charts." In: *arXiv preprint arXiv:2210.06628* (2022).

[92] Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. *Chart-to-Text: A Large-Scale Benchmark for Chart Summarization*. 2022. arXiv: 2203.06486 [cs.CL].

[93] Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. "Chart-to-Text: A Large-Scale Benchmark for Chart Summarization." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022, pp. 4005–4023.

[94] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. "OCR-Free Document Understanding Transformer." In: *European Conference on Computer Vision*. 2022, pp. 498–517.

[95] Nam Wook Kim, Shakila Cherise Joyner, Amalia Riegelhuth, and Y Kim. "Accessible visualization: Design space, opportunities, and challenges." In: *Computer Graphics Forum*. Vol. 40. 3. Wiley Online Library. 2021, pp. 173–188.

[96] Koichi Kise, Akinori Sato, and Motoi Iwata. "Segmentation of page images using the area Voronoi diagram." In: *Computer Vision and Image Understanding* 70.3 (1998), pp. 370–382.

[97] Stephen E. Krufka and Kenneth E. Barner. "Automatic Production of Tactile Graphics from Scalable Vector Graphics." In: *Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility*. Assets '05. Baltimore, MD, USA: Association for Computing Machinery, 2005, 166–172. ISBN: 1595931597.

[98] Jay Lal, Aditya Mitkari, Mahesh Bhosale, and David Doermann. *LineFormer: Rethinking Line Chart Data Extraction as Instance Segmentation*. 2023. arXiv: 2305.01837 [cs.CV].

[99] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. "Pix2struct: Screenshot parsing as pretraining for visual language understanding." In: *International Conference on Machine Learning*. PMLR. 2023, pp. 18893–18912.

[100] D.J. Leiner. *SoSci Survey – the Solution for Professional Online Questionnaires*. 2019. URL: https://www.soscisurvey.de/en/index.

[101] Jingyi Li, Son Kim, Joshua A Miele, Maneesh Agrawala, and Sean Follmer. "Editing spatial layouts through tactile templates for people with visual impairments." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–11.

[102] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. "Dit: Self-supervised pre-training for document image transformer." In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 3530–3539.

[103] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." In: *International Conference on Machine Learning*. PMLR. 2022, pp. 12888–12900.

[104] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. "DocBank: A benchmark dataset for document layout analysis." In: *arXiv preprint arXiv:2006.01038* (2020).

[105] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. "Microsoft coco: Common objects in context." In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer. 2014, pp. 740–755.

[106] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. "Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models." In: *arXiv preprint arXiv:2311.07575* (2023).

[107] Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. "Deplot: One-shot visual language reasoning by plot-to-table translation." In: *arXiv preprint arXiv:2212.10505* (2022).

[108] Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Martin Eisenschlos. "Matcha: Enhancing visual language pretraining with math reasoning and chart derendering." In: *arXiv preprint arXiv:2212.09662* (2022).

[109] Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Martin Eisenschlos. *MatCha: Enhancing Visual Language Pretraining with Math Reasoning and Chart Derendering*. 2023. arXiv: 2212.09662 [cs.CL].

[110] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. "A Survey on Hallucination in Large Vision-Language Models." In: *arXiv preprint arXiv:2402.00253* (2024).

[111] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. *Improved Baselines with Visual Instruction Tuning*. 2023. arXiv: 2310.03744 [cs.CV].

[112] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. *Improved Baselines with Visual Instruction Tuning*. 2023.

[113] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. "Improved baselines with visual instruction tuning." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 26296–26306.

[114] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. *Visual Instruction Tuning*. 2023.

[115] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. "Visual instruction tuning." In: *Advances in neural information processing systems* 36 (2024).

[116] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin transformer: Hierarchical vision transformer using shifted windows." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[117] Alan Lundgard and Arvind Satyanarayan. "Accessible visualization via natural language descriptions: A four-level model of semantic content." In: *IEEE transactions on visualization and computer graphics* 28 (2021), pp. 1073–1083.

[118] Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. "Chartocr: Data extraction from charts images via a deep hybrid framework." In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2021, pp. 1917–1925.

[119] Kim Marriott, Bongshin Lee, Matthew Butler, Ed Cutrell, Kirsten Ellis, Cagatay Goncu, Marti Hearst, Kathleen McCoy, and Danielle Albers Szafir. "Inclusive data visualization for people with disabilities: a call to action." In: *Interactions* 28.3 (2021), pp. 47–51.

[120] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. "Unichart: A universal vision-language pretrained model for chart comprehension and reasoning." In: *arXiv preprint arXiv:2305.14761* (2023).

[121] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. "Chartqa: A benchmark for question answering about charts with visual and logical reasoning." In: *arXiv preprint arXiv:2203.10244* (2022).

[122] Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. "ChartInstruct: Instruction Tuning for Chart Comprehension and Reasoning." In: *arXiv preprint arXiv:2403.09028* (2024).

[123] Damien Masson, Sylvain Malacria, Daniel Vogel, Edward Lank, and Géry Casiez. "ChartDetective: Easy and Accurate Interactive Data Extraction from Complex Vector Charts." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215.

[124] Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. *ChartAssisstant: A Universal Chart Multimodal Language Model via Chart-to-Table Pre-training and Multitask Instruction Tuning*. 2024. arXiv: 2401.02384 [cs.CV].

[125] Metec AG. *laptop for the blind*. Accessed: 11 September 2023. 2023. URL: https://metec-ag.de/en/produkte-graphik-display.php.

[126] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. "Plotqa: Reasoning over scientific plots." In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020, pp. 1527–1536.

[127] Ishan Misra and Laurens van der Maaten. "Self-supervised learning of pretext-invariant representations." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 6707–6717.

[128] Apostolos Modas, Rahul Rade, Guillermo Ortiz-Jiménez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. "Prime: A few primitives can boost robustness to common corruptions." In: *European Conference on Computer Vision*. Springer. 2022, pp. 623–640.

[129] Ajoy Mondal, Madhav Agarwal, and CV Jawahar. "Dataset agnostic document object detection." In: *Pattern Recognition* 142 (2023), p. 109698.

[130] Omar Moured, Sara Alzalabny, Anas Osman, Thorsten Schwarz, Karin Müller, and Rainer Stiefelhagen. "ChartFormer: A Large Vision Language Model for Converting Chart Images into Tactile Accessible SVGs." In: *Computers Helping People with Special Needs*. Cham: Springer Nature Switzerland, 2024, pp. 299–305.

[131] Omar Moured, Sara Alzalabny, Thorsten Schwarz, Bastian Rapp, and Rainer Stiefelhagen. "Accessible Document Layout: An Interface for 2D Tactile Displays." In: *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*. 2023, pp. 265–271.

[132] Omar Moured, Morris Baumgarten-Egemole, Karin Müller, Alina Roitberg, Thorsten Schwarz, and Rainer Stiefelhagen. "Chart4blind: An intelligent interface for chart accessibility conversion." In: *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 2024, pp. 504–514.

[133] Omar Moured, Shahid Ali Farooqui, Karin Müller, Sharifeh Fadaeijouybari, Thorsten Schwarz, Mohammed Javed, and Rainer Stiefelhagen. "Alt4Blind: A User Interface to Simplify Charts Alt-Text Creation." In: *Computers Helping People with Special Needs*. Cham: Springer Nature Switzerland, 2024, pp. 291–298.

[134] Omar Moured, Jiaming Zhang, Alina Roitberg, Thorsten Schwarz, and Rainer Stiefelhagen. "Line Graphics Digitization: A Step Towards Full Automation." In: *International Conference on Document Analysis and Recognition*. Springer. 2023, pp. 438–453.

[135] Omar Moured, Jiaming Zhang, M Saquib Sarfraz, and Rainer Stiefelhagen. "AltChart: Enhancing VLM-based Chart Summarization Through Multi-Pretext Tasks." In: *arXiv preprint arXiv:2405.13580* (2024). **Accepted at ICDAR 2024, to be presented on September 2nd, 2024.**

[136] NV Access. *NVDA Screen Reader*. Accessed on 09/20/2018. 2020. URL: https://www.nvaccess.org/.

[137] H Nagao and S Hatanaka. "Braille Books and Translation of Illustration by Using PCs." In: *Dokusho Kobo, Tokyo* (2005).

[138] George Nagy, Thomas A. Nartker, and Stephen V. Rice. "Optical character recognition: an illustrated guide to the frontier." In: *Document Recognition and Retrieval VII*. Ed. by Daniel P. Lopresti and Jiangying Zhou. Vol. 3967. International Society for Optics and Photonics. SPIE, 1999, pp. 58 –69.

[139] George Nagy and Sharad C Seth. "Hierarchical representation of optically scanned documents." In: (1984).

[140] Mehdi Noroozi and Paolo Favaro. *Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles*. 2017. arXiv: 1603.09246 [cs.CV].

[141] Braille Authority of North America. *Guidelines and Standards for Tactile Graphics*. 2010. URL: https://www.brailleauthority.org/tg/web-manual/index.html.

[142] Lawrence O'Gorman. "The document spectrum for page layout analysis." In: *IEEE Transactions on pattern analysis and machine intelligence* 15.11 (1993), pp. 1162–1173.

[143] Jason Obeid and Enamul Hoque. "Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model." In: *arXiv preprint arXiv:2010.09142* (2020).

[144] World Health Organization et al. "World report on vision." In: (2019).

[145] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. "Training language models to follow instructions with human feedback." In: *Advances in neural information processing systems* 35 (2022), pp. 27730–27744.

[146] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a method for automatic evaluation of machine translation." In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[147] Yanting Pei, Yaping Huang, Qi Zou, Xingyuan Zhang, and Song Wang. "Effects of image degradation and degradation removal to CNN-based image classification." In: *IEEE transactions on pattern analysis and machine intelligence* 43.4 (2019), pp. 1239–1253.

[148] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter Staar. "Doclaynet: A large human-annotated dataset for document-layout segmentation." In: *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*. 2022, pp. 3743–3751.

[149] E. Pontelli, D. Gillan, W. Xiong, E. Saad, G. Gupta, and A. I. Karshmer. "Navigation of HTML Tables, Frames, and XML Fragments." In: Assets '02. Edinburgh, Scotland: Association for Computing Machinery, 2002, 25–32. ISBN: 1581134649. DOI: 10.1145/638249.638256.

[150] Denise Prescher, Jens Bornschein, and Gerhard Weber. "Production of accessible tactile graphics." In: *Computers Helping People with Special Needs: 14th International Conference, ICCHP 2014, Paris, France, July 9-11, 2014, Proceedings, Part II 14*. Springer. 2014, pp. 26–33.

[151] Round Table on Information Access for People with Print Disabilities. "Guidelines for producing accessible graphics / Round Table on Information Access for People with Print Disabilities Inc." In: (2022). URL: https://printdisability.org/guidelines/graphics-2022/.

[152] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. "Learning transferable visual models from natural language supervision." In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[153] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. "Learning transferable visual models from natural language supervision." In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[154] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer." In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 5485–5551.

[155] Raian Rahman, Rizvi Hasan, and Abdullah Al Farhad. "ChartSumm: A large scale benchmark for Chart to Text Summarization." PhD thesis. Department of Computer Science and Engineering (CSE), Islamic University of . . ., 2022.

[156] Veenu Rani, Syed Tufael Nabi, Munish Kumar, Ajay Mittal, and Krishan Kumar. "Self-supervised learning: A succinct review." In: *Archives of Computational Methods in Engineering* 30.4 (2023), pp. 2761–2775.

[157] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." In: *Advances in neural information processing systems* 28 (2015).

[158] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." In: *Advances in neural information processing systems* 28 (2015).

[159] John Roach. "What's that? Microsoft's latest breakthrough, now in Azure AI, describes images as well as people do." In: *The AI Blog, Microsoft* (2020).

[160] Ankit Rohatgi. "WebPlotDigitizer user manual version 3.4." In: (2014), pp. 1–18. URL: https://automeris.io/WebPlotDigitizer/userManual.pdf.

[161] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].

[162] Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. "A survey of evaluation metrics used for NLG systems." In: *ACM Computing Surveys (CSUR)* 55.2 (2022), pp. 1–39.

[163] Madeline C Schiappa, Yogesh S Rawat, and Mubarak Shah. "Self-supervised learning for videos: A survey." In: *ACM Computing Surveys* 55.13s (2023), pp. 1–37.

[164] J Serrano, GM Bruscas, JV Abellán, and R Lázaro. "Study of additive manufacturing techniques to obtain tactile graphics." In: *IOP Conference Series: Materials Science and Engineering*. 1. IOP Publishing. 2021, p. 012117.

[165] Ather Sharif, Sanjana Shivani Chintalapati, Jacob O Wobbrock, and Katharina Reinecke. "Understanding screen-reader users' experiences with online data visualizations." In: *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 2021, pp. 1–16.

[166] VP Shivasankaran, Muhammad Yusuf Hassan, and Mayank Singh. "LineEX: Data Extraction from Scientific Line Charts." In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2023, pp. 6202–6210.

[167] B. Shneiderman. "The eyes have it: a task by data type taxonomy for information visualizations." In: *Proceedings 1996 IEEE Symposium on Visual Languages*. 1996, pp. 336–343. DOI: 10.1109/VL.1996.545307.

[168] Ben Shneiderman. "The eyes have it: A task by data type taxonomy for information visualizations." In: *Proceedings 1996 IEEE symposium on visual languages*. IEEE. 1996, pp. 336–343.

[169] Mandhatya Singh, Muhammad Suhaib Kanroo, Hadia Showkat Kawoosa, and Puneet Goyal. "Towards accessible chart visualizations for the non-visuals: Research, applications and gaps." In: *Computer Science Review* 48 (2023), p. 100555. ISSN: 1574-0137.

[170] Ray Smith. "An overview of the Tesseract OCR engine." In: *Ninth international conference on document analysis and recognition (ICDAR 2007)*. Vol. 2. IEEE. 2007, pp. 629–633.

[171] Charalambos Strouthopoulos and Nikos Papamarkos. "Text identification for document image analysis using a neural network." In: *Image and Vision Computing* 16.12-13 (1998), pp. 879–896.

[172] Saiganesh Swaminathan, Thijs Roumen, Robert Kovacs, David Stangl, Stefanie Mueller, and Patrick Baudisch. "Linespace: A Sensemaking Platform for the Blind." In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. San Jose, California, USA: Association for Computing Machinery, 2016, 2175–2185. ISBN: 9781450333627.

[173] Benny Tang, Angie Boggust, and Arvind Satyanarayan. "VisText: A Benchmark for Semantically Rich Chart Captioning." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 7268–7298.

[174] Juan Terven, Diana-Margarita Córdova-Esparza, and Julio-Alejandro Romero-González. "A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas." In: *Machine Learning and Knowledge Extraction* 5.4 (2023), pp. 1680–1716.

[175] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. "Llama 2: Open foundation and fine-tuned chat models." In: *arXiv preprint arXiv:2307.09288* (2023).

[176] Tuan Anh Tran, Kanghan Oh, In-Seop Na, Guee-Sang Lee, Hyung-Jeong Yang, and Soo-Hyung Kim. "A robust system for document layout analysis using multilevel homogeneity structure." In: *Expert Systems With Applications* 85 (2017), pp. 99–113.

[177] B. Tummers. *DataThief III*. 2006. URL: https://datathief.org/.

[178] Russ Unger and Carolyn Chandler. *A Project Guide to UX Design: For user experience designers in the field or in the making*. New Riders, 2012.

[179] Shai Vaingast. *im2graph*. 2014. URL: https://www.im2graph.co.il/.

[180] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In: *Advances in neural information processing systems* 30 (2017).

[181] World Wide Web Consortium (W3C). *Web Content Accessibility Guidelines (WCAG) 2.1*. Accessed on 06/20/2024. 2024. URL: https://www.w3.org/TR/WCAG21/.

[182] Friedrich M Wahl, Kwan Y Wong, and Richard G Casey. "Block segmentation and text extraction in mixed text/image documents." In: *Computer graphics and image processing* 20.4 (1982), pp. 375–390.

[183] Jingdong Wang et al. "Deep high-resolution representation learning for visual recognition." In: *TPAMI* (2021).

[184] Lucai Wang, Hongda Qin, Xuanyu Zhou, Xiao Lu, and Fengting Zhang. "R-YOLO: A robust object detector in adverse weather." In: *IEEE Transactions on Instrumentation and Measurement* 72 (2022), pp. 1–11.

[185] Lucy Lu Wang, Isabel Cachola, Jonathan Bragg, Evie Yu-Yen Cheng, Chelsea Haupt, Matt Latzke, Bailey Kuehl, Madeleine N van Zuylen, Linda Wagner, and Daniel Weld. "SciA11y: Converting Scientific Papers to Accessible HTML." In: *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS '21. Virtual Event, USA: Association for Computing Machinery, 2021. ISBN: 9781450383066. DOI: 10.1145/3441852.3476545. URL: https://doi.org/10.1145/3441852.3476545.

[186] Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Hao Tan, and Yu Qiao. "Scaling data generation in vision-and-language navigation." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 12009–12020.

[187] Tetsuya Watanabe, Toshimitsu Yamaguchi, and Masaki Nakagawa. "Development of software for automatic creation of embossed graphs: Comparison of non-visual data presentation methods and development up-to-date." In: *Computers Helping People with Special Needs: 13th International Conference, ICCHP 2012, Linz, Austria, July 11-13, 2012, Proceedings, Part I 13*. Springer. 2012, pp. 174–181.

[188] Chung-Chih Wu, Chien-Hsing Chou, and Fu Chang. "A machine-learning approach for analyzing document layout structures with two reading orders." In: *Pattern recognition* 41.10 (2008), pp. 3200–3213.

[189] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. "Automatic alt-text: Computer-generated image descriptions for blind users on a social network service." In: *proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 2017, pp. 1180–1192.

[190] Renqiu Xia et al. *ChartX & ChartVLM: A Versatile Benchmark and Foundation Model for Complicated Chart Reasoning*. 2024. arXiv: 2402.12185 [cs.CV].

[191] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. "SegFormer: Simple and efficient design for semantic segmentation with transformers." In: *NeurIPS*. 2021.

[192] Ying Xu, Xu Zhong, Antonio Jose Jimeno Yepes, and Jey Han Lau. "Forget me not: Reducing catastrophic forgetting for domain adaptation in reading comprehension." In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2020, pp. 1–8.

[193] Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C Lee Giles. "Learning to extract semantic structure from documents using multimodal fully convolutional neural networks." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 5315–5324.

[194] Cong Yao, Jianan Wu, Xinyu Zhou, Chi Zhang, Shuchang Zhou, Zhimin Cao, and Qi Yin. "Incidental scene text understanding: Recent progresses on icdar 2015 robust reading competition challenge 4." In: *arXiv preprint arXiv:1511.09207* (2015).

[195] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. "UReader: Universal OCR-free Visually-situated Language Understanding with Multimodal Large Language Model." In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023, pp. 2841–2858.

[196] Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, Yichen Xu, Qin Jin, Ji Zhang, and Fei Huang. "Tinychart: Efficient chart understanding with visual token merging and program-of-thoughts learning." In: *arXiv preprint arXiv:2404.16635* (2024).

[197] Mingliang Zhang, Zhen Cao, Juntao Liu, Liqiang Niu, Fandong Meng, and Jie Zhou. *WeLayout: WeChat Layout Analysis System for the ICDAR 2023 Competition on Robust Layout Segmentation in Corporate Documents*. 2023. arXiv: 2305.06553 [cs.CV]. URL: https://arxiv.org/abs/2305.06553.

[198] Peng Zhang, Can Li, Liang Qiao, Zhanzhan Cheng, Shiliang Pu, Yi Niu, and Fei Wu. "VSR: a unified framework for document layout analysis combining vision, semantics and relations." In: *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16*. Springer. 2021, pp. 115–130.

[199] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. "Pyramid scene parsing network." In: *CVPR*. 2017.

[200] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. "Publaynet: largest dataset ever for document layout analysis." In: *2019 International conference on document analysis and recognition (ICDAR)*. IEEE. 2019, pp. 1015–1022.

[201] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. "Understanding the robustness in vision transformers." In: *International Conference on Machine Learning*. PMLR. 2022, pp. 27378–27394.

[202] Jiawen Zhu, Jinye Ran, Roy Ka-Wei Lee, Zhi Li, and Kenny Choo. "AutoChart: A Dataset for Chart-to-Text Generation Task." In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. 2021, pp. 1636–1644.

[203] Jiawen Zhu, Jinye Ran, Roy Ka-Wei Lee, Zhi Li, and Kenny Choo. "AutoChart: A Dataset for Chart-to-Text Generation Task." In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Ed. by Ruslan Mitkov and Galia Angelova. Held Online: INCOMA Ltd., Sept. 2021, pp. 1636–1644. URL: https://aclanthology.org/2021.ranlp-1.183.

[204] Hong Zou and Jutta Treviranus. "ChartMaster: A tool for interacting with stock market charts using a screen reader." In: *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. 2015, pp. 107–116.