SlideCraft: Synthetic Slides Generation for Robust Slide Analysis

Travis Seng, Axel Carlier, Thomas Forgione, Vincent Charvillat, Wei Tsang Ooi

SlideCraft: Synthetic Slides Generation for Robust Slide Analysis

Travis Seng[1,2⋆][0000−0002−1548−8841], Axel Carlier[1,2][0000−0002−6838−3445], Thomas Forgione[3][0009−0005−2085−9503], Vincent Charvillat[1][0009−0006−6755−5774], and Wei Tsang Ooi[4][0000−0001−8994−1736]

[1] IRIT Toulouse, France
[2] IPAL, IRL2955
[3] Polymny Studio
[4] National University of Singapore

Abstract. The increasing amount of slide presentations in various sectors has amplified the need for effective slide layout and semantic analysis. However, we found that current slide datasets contain inconsistencies, mislabels, and incomplete annotations. Using them as a basis for developing deep learning-based slide analysis models could lead to models that are not robust and suboptimal. Addressing these challenges, we introduce SlideCraft, a tool for creating synthetic slide datasets that imitate real-world presentations. This tool overcomes the drawbacks of existing datasets by allowing users to create balanced, diverse, and accurately annotated slide data. We demonstrate SlideCraft's efficacy in enhancing slide layout analysis algorithms, focusing on its capability to improve dataset quality and object detection performance. Our code and a demo can be found at this address.

Keywords: Slide datasets · Slide analysis · Open source tool · Synthetic dataset.

1 Introduction

Presentations in sectors such as education and research are increasingly being recorded and shared online, particularly since the pandemic of the early 2020s. This trend fueled a renewed interest in innovative applications for processing and interacting with recorded digital presentations, such as summarization [30,31], indexing [32,23,2,16], and enhanced browsing [32,23,16,2]. Slides are important visual aids for a presentation that often drive the structure and key points of the content of a presentation. Extracting the semantics from the slides, therefore, is an important fundamental task that enables these downstream applications. Much of the slides available online, however, are in image or video format, making computational analysis and understanding of the slides a non-trivial problem.

Recent efforts in slide analysis have pivoted to data-driven, learning-based, methods [15,28]. For such methods to be effective, there needs to be a large

⋆ Correspondence to travis.seng@irit.fr

corpus of data with high-quality annotations. However, existing slide datasets [9,8,15,28] often either suffer from inconsistencies and inaccuracies, or lack the essential annotations [3] necessary for the development of effective analysis algorithms. Beyond these issues, there are additional concerns that slow down progress in this domain. Firstly, the process of accurately annotating slide datasets is labor-intensive and expensive, requiring significant expertise and time to ensure high-quality annotations. Secondly, this challenge is compounded by the scarcity of comprehensive slide datasets available in the public domain, limiting the diversity and scope of training data for analysis algorithms. Finally, there is often an imbalance in the type of elements present on the slides within these datasets. For example, some classes of visual or textual elements may be overrepresented while others are underrepresented, leading to skewed learning outcomes and limiting the algorithms' ability to generalize across different slide layouts and content types.

To address these challenges, we introduce SlideCraft, a tool developed for generating synthetic slide datasets. SlideCraft first tackles the issue of labor-intensive annotations. Generating content with annotation eliminates the costly and time-consuming nature of manual annotation, ensuring high-quality, consistent annotations across the dataset. Second, it mitigates the scarcity of comprehensive slide datasets by generating a large and diverse array of realistic slides. Finally, it addresses the imbalance in slide elements by allowing for customizable element distribution in generated slides. This ensures a balanced representation of both visual and textual elements. In summary, SlideCraft's design not only imitates real-world presentations but also methodically overcomes the existing limitations of slide datasets, thereby facilitating the development of slide analysis algorithms.

In this paper, we present the development, capabilities, and potential impact of SlideCraft. Our focus is on its contribution to overcoming the current limitations in slide dataset availability and quality, thereby paving the way for future advancements in digital presentation analysis. **We show that by augmenting the existing dataset with slides generated from our tool, we can improve the mAP50 up to 13% on object detection models.** The source code of SlideCraft will be made publicly available for research purposes, fostering further innovation and collaboration in the field.

2 Related Work

Studies in slide dataset creation have primarily focused on collating slides from existing sources, with an emphasis on real-world diversity. However, these datasets often face challenges in terms of size, annotation accuracy, consistency, and imbalance. WiSe [9] and SPaSe [8] datasets, while offering segmentation masks across 25 diverse classes, are limited by their small size, containing only 2000 images – a quantity insufficient for comprehensive analysis. FitVid [15] introduces a more subject-diverse object detection dataset over 12 classes, yet it comprises only 5527 images. In contrast, Slideshare-1M [3], despite being the

largest slide collection, lacks annotations, because it was made for content retrieval. Meanwhile, SlideVQA [28] stands out as the most extensive annotated slide dataset available, with 52,480 slides across 9 classes designed for object detection, showcasing a step forward in annotated slide data availability. Despite this, it faces two notable issues: a disparity between text and visual content, and the presence of inconsistencies as shown in Figure 1.

Document layout analysis has recently garnered increased attention in the machine learning community, a development largely fueled by the introduction of more popular benchmarks and high-quality datasets such as DocBank [18], DocLayNet [24], and PubLayNet [33]. These datasets have been instrumental in advancing the field, offering comprehensive and diverse collections of document layouts that facilitate the training and testing of sophisticated layout analysis algorithms. However, despite these advancements, a significant gap remains in the specific domain of slide layout analysis. Unlike standard documents, slides have distinct characteristics: they are typically more visual, presented in landscape format, and contain less structured text. Furthermore, the classification needs for slide elements differ markedly from those in traditional document layouts. For instance, slides often require specific classes for elements like bullet points, titles, and visual aids, which are not commonly accounted for in general document layout datasets. The existing datasets, while robust for general document analysis, fall short of addressing the unique demands of slide presentations and potential applications.

Specialized datasets focused on document elements like tables [17], charts [21], diagrams [14], plots ([22]), equations [6] and handwriting [20] have been developed to encourage better analysis and recognition. Given that slide presentations often incorporate such elements, leveraging these specific datasets within SlideCraft presents an excellent opportunity to refine and broaden the analysis capabilities of models trained using SlideCraft-generated slides.

Historically, synthetic datasets have been approached with caution in machine learning applications due to their tendency to diverge from real-world scenarios. This divergence often results in a performance gap when algorithms trained on synthetic data are applied to actual tasks. In contrast to this trend, some researchers have employed advanced techniques like Generative Adversarial Networks (GANs) to create more realistic synthetic datasets [7,5,4]. These methods involve training neural networks to generate data that closely mimic real-world scenarios, a technique that has shown promise in fields such as image and document generation but which cannot be easily used to create reliable annotations. Others have utilized photorealistic open-world computer games to generate large datasets, offering reliable pixel-level semantic annotations for enhanced accuracy in various applications [27].

Furthermore, another approach in the generation of synthetic datasets for document layout analysis has involved creating layouts first by using a generative model and then populating them with content [25]. This method targets article layouts, where the focus is on arranging text and graphical elements in a manner consistent with traditional document formats, thereby creating syn-

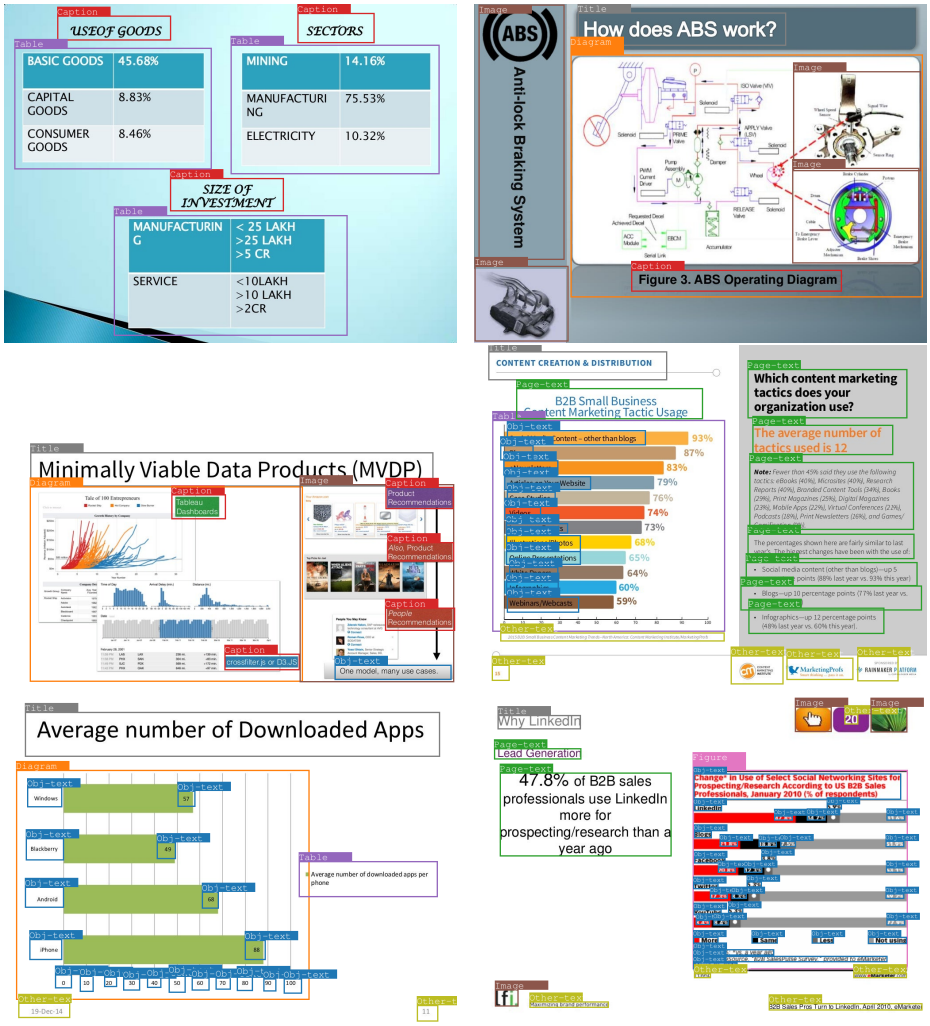

Fig. 1. SlideVQA contains some inconsistencies and errors. The Obj-Text class which represents all the text seen on any graphical elements (Image, Diagram, Table, Figure) is not always annotated. Moreover, there is some confusion between the class Figure and Diagram shown in the last two slides. Even though they are both horizontal charts, one is labeled as Diagram and in the other, as Figure. This affects training performance and evaluation reliability.

thetic datasets that resemble the structure of real-world documents. However, these techniques are predominantly tailored to the needs of standard document layouts, such as articles and reports, and do not directly address the unique requirements of slide presentations, such as having more visual content such as charts, graphs, and images, as well as the necessity to accommodate diverse layouts. The semantic organization within slides, how information is structured and prioritized, differs significantly from traditional documents, necessitating tools, and datasets specifically designed to understand and generate these visually oriented formats.

We provide a detailed analysis of SlideCraft in the next section.

3 SlideCraft

SlideCraft is a tool designed for generating extensive, labeled datasets for training learning-based slide layout analysis methods. Its primary function is to facilitate the production of a wide array of annotated slide data. This tool operates through a set of flexible rules that ensure the generation of coherent, well-structured, and readable presentations. These rules can be dynamically adjusted by users, allowing for the creation of customized layouts and styles to suit specific requirements.

The tool can be broken into four components, each responsible for generating a different element: content, layout, style, and annotation. Figure 2 shows the whole pipeline of SlideCraft.

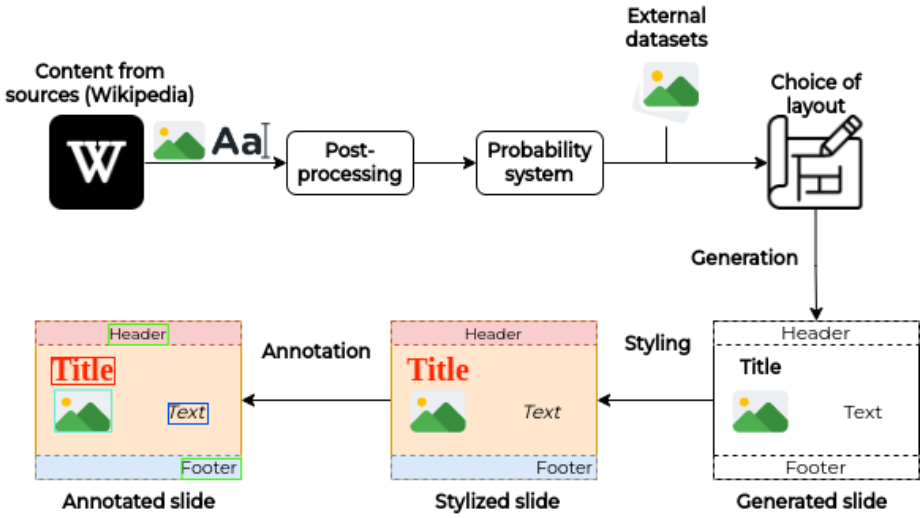

Fig. 2. Pipeline of SlideCraft

3.1 Content

The content component of SlideCraft is integral to generating the textual and visual elements of the slides, with primary reliance on Wikipedia articles for source material in our current implementation (the source material could come from anywhere). Each section and subsection of an article is adeptly converted into an individual slide, aligning with the original content's structure. To adapt the typically lengthy and detailed text from Wikipedia articles to a slide-friendly format, we employed Mistral-7B [11], a Large Language Model, for summarization. This LLM was tasked with distilling the textual elements into bullet points and short sentences, which are more suitable for the concise and direct nature of slide presentations.

However, given that Wikipedia articles are predominantly text-based and slides necessitate a strong visual component, we introduce the possibility of supplementing them with various graphical elements from other public datasets. This addition includes charts (ChartQA [21]), plots (PlotQA [22]), diagrams (AI2D [14]), tables (TableBank [17]), equations (im2latex-100k [6]), handwriting (IAM-OnDB [20]), logos (Logodet-3k [29]) and more. While incorporating these diverse visual elements could introduce a degree of incoherence with the original text, it is a deliberate choice aimed at enhancing the overall visual diversity of the slides. This step is beneficial for the primary objective of the tool - to improve the performance of slide layout analysis algorithms. We show in our experiments that the use of this additional material helps to improve object detection performance in several classes.

By expanding the range of visual content in the slides, we provide a richer and more challenging dataset for training these algorithms, ultimately contributing to their enhanced accuracy and effectiveness in real-world applications.

3.2 Layout

The layout component is crucial in determining the arrangement and presentation of content within each slide. This component is designed to emulate traditional layouts commonly found in popular presentation tools like Google Slides and PowerPoint. It assesses the quantity and type of content elements - be it text, images, or graphical data - and selects the most suitable layout for each slide. The decision-making process is guided by principles of design and readability, ensuring that the final output is not only visually appealing but also easy to comprehend. To actualize these layouts, we utilize Marp [1], a versatile tool that enables the creation of slides using Markdown, HTML, and CSS. Marp's flexibility and simplicity allow for efficient translation of the chosen layouts into polished slides. Based on the content, we choose a template tailored to the content, and we generate a markdown file that contains instructions for Marp to translate into a slide in HTML format. Templates are written in HTML and CSS and contain code to create layouts with columns, headers, footers, and different positions for elements. This approach ensures that SlideCraft can produce slides that are aesthetically consistent with standard presentation formats,

thereby making the generated datasets ideal for enhancing slide layout analysis algorithms. Figure 3 shows some examples of layouts we can generate.

A probability system also plays a part in determining the composition of each slide. This system allows for control over the occurrence probabilities of various elements within the slides. Users can adjust the generation process by specifying the weights for different classes and determining their inclusion in a slide in a greedy way. The system plays a role both in the content phase, where it determines the likelihood of incorporating external content, and in the layout phase, deciding whether to include specific classes like titles, headers, and footers. This feature is particularly beneficial when aiming to balance an existing dataset or when focusing on enhancing specific classes that may be underrepresented or inadequately portrayed in the dataset. By adjusting these probabilities, we can tailor the layout of the slides to address specific needs or deficiencies in the dataset. For example, if the initial analysis indicates that slides with diagrams are less frequent, the system can be configured to generate more slides with diagrams. This level of customization in the layout component not only adds versatility to SlideCraft, but also ensures that the resulting dataset is well-rounded and accurately reflective of diverse presentation scenarios. Such targeted adjustments are instrumental in fine-tuning the dataset for more effective training and evaluation of slide layout analysis models.

3.3 Style

The style component is responsible for the aesthetic diversity of the slides. This component operates by randomly selecting various stylistic parameters to ensure that each slide is distinct from the others, but can also be modified to target certain styles of slides. These parameters include a wide range of elements such as colors for different text sections (like paragraphs, titles, headers, and footers), font types, text sizes, the degree of text boldness, and more. Additionally, it also encompasses choices for background styles and colors. This is done through the modification of the CSS in the markdown files read by Marp.

Randomization in selecting these stylistic features is key to producing a dataset with high variability. This diversity is not merely cosmetic; it is needed for the robustness of the slide layout analysis algorithms. By exposing these algorithms to diverse styles, we challenge and enhance their ability to accurately analyze and interpret slides across a broad spectrum of designs. In doing so, the style component contributes to the generation of a comprehensive and varied dataset. Figure 4 shows some outputs of SlideCraft with different layouts and styles, and Figure 5 shows the same but for the same source material.

While SlideCraft represents the first attempt at creating a synthetic slide dataset for slide layout analysis, it is important to acknowledge its limitations in replicating the full spectrum of slide styles. The tool is not designed to mimic every possible type of slide, as the variety and complexity of presentation styles are vast and often context-specific. However, our goal with SlideCraft is to cover the most common and widely used slide styles. By focusing on these prevalent

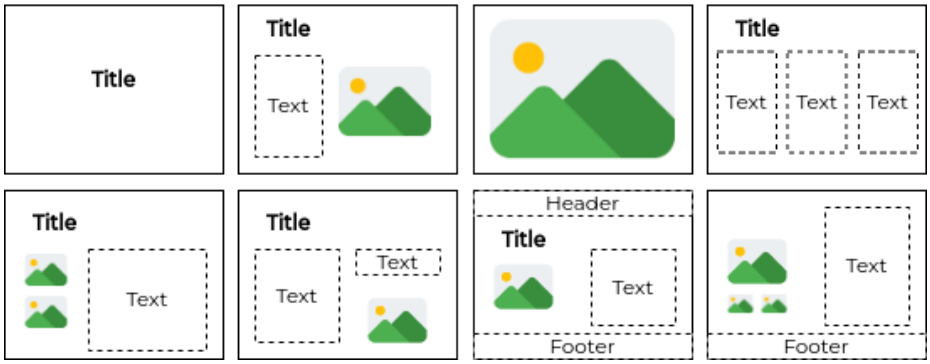

Fig. 3. Examples of different layouts: different number of columns, display of header and footer, display of the title. The number of elements taken from the content can be adjusted to choose between more visual or more textual output.

formats, we aim to provide a comprehensive and representative dataset that reflects the majority of real-world scenarios. This strategic approach allows us to optimize the tool's effectiveness in training and improving slide layout analysis algorithms, while realistically addressing the practical constraints of such a generative system.

3.4 Annotations

SlideCraft generates annotations through the processing of the HTML generated by Marp for each slide. We employ JavaScript to access each text element within the Document Object Model (DOM) and obtain precise bounding boxes for these elements. To enhance accuracy, we modify the HTML by adding span tags around paragraphs and words, enabling us to capture more detailed bounding boxes around the text at a paragraph and word level. This approach ensures precise demarcation of text elements, critical for effective layout analysis.

In parallel, segmentation masks are created by isolating individual elements on the slides with CSS and computing the image difference. This method is particularly effective for identifying and delineating graphical elements within the slides.

For the classification of text elements, we use two information. Primarily, we rely on HTML tags to determine their classes (e.g., 'p' for general text, 'li' for bullet points, 'h1' for titles, etc.). Additionally, we incorporate class information derived during the layout computation phase, further refining the accuracy of our text annotations.

Graphical elements, on the other hand, are annotated based on the classifications available from the labeled images we use from existing datasets. By incorporating these pre-labeled images into our slides, we ensure that each graphical element is accurately categorized, significantly enhancing the reliability of our annotations. We also apply OCR to graphical elements to extract bounding

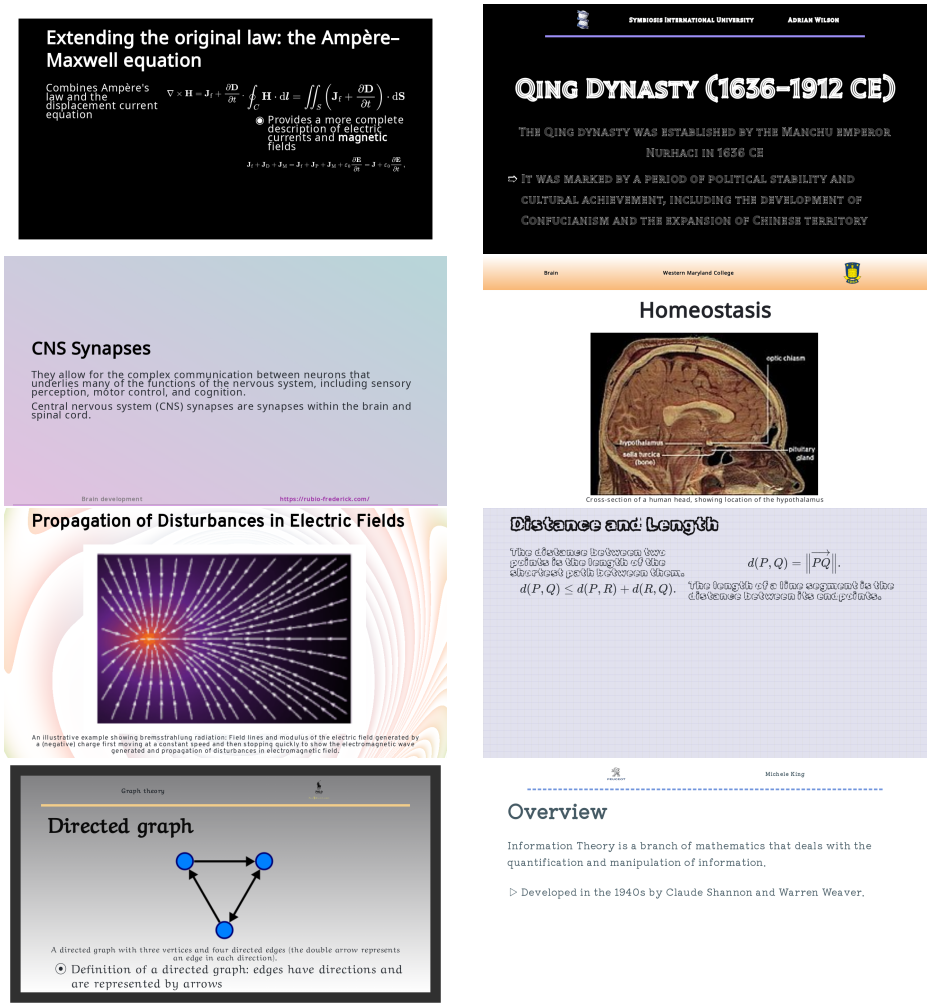

Fig. 4. Examples of SlideCraft's generated slides. Different styles and layouts are shown. Colors, backgrounds, font size, and font style are chosen randomly to increase the diversity of the generation.

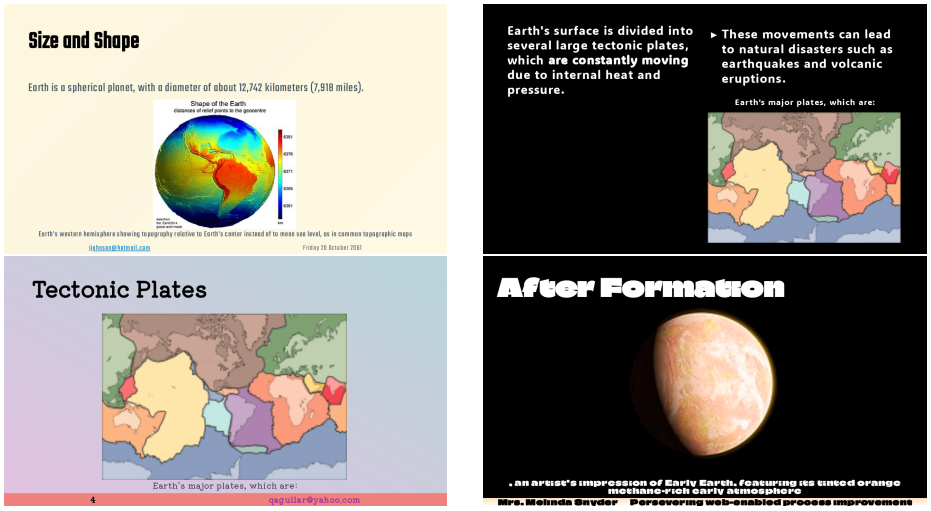

Fig. 5. Different styles of slides from the same source (Wikipedia: Earth). We can see diverse fonts and backgrounds, various font sizes and colors, the presence of a footer or not, and different numbers of elements.

boxes for any text they contain, enhancing text recognition across diverse slide components.

This comprehensive approach to annotation, combining DOM manipulation, image processing, and existing dataset labels, ensures that SlideCraft not only creates visually varied slides but also provides richly annotated data.

The tool can be customized to generate custom classes as needed as shown in Figure 6 where we generate different classes depending on the dataset we want to extend.

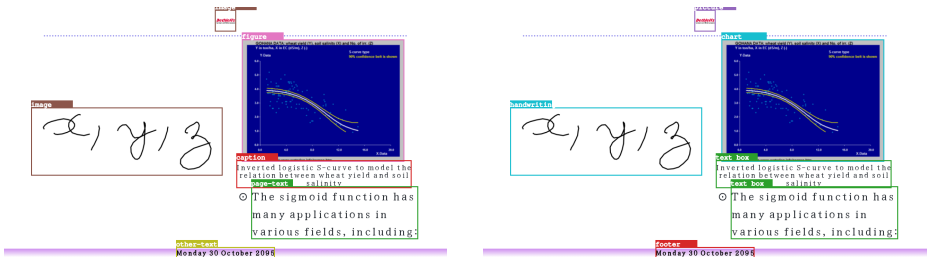

Fig. 6. Example of generated slides for SlideVQA annotations (Left) and Fitvid annotations (Right). Handwriting, header, footer, and equation, called Figure in FitVid, do not exist in SlideVQA.

4 Experiments

To demonstrate SlideCraft's effectiveness, we evaluate object detection performance using two models and use our generated synthetic dataset to extend two existing datasets.

4.1 Datasets

We selected the two largest annotated datasets available. The FitVid dataset, featuring 5,527 images, covers a wide array of 12 classes, including Title, Text Box, Picture, Chart, Figure, Diagram, Table, Schematic Diagram, Header, Footer, Handwriting, and Instructor. Given that there is no test dataset, we create a split by carefully separating slides from the same presentation. We obtain a train set with 4,421 slides and a test set with 1,106 slides. Similarly, the SlideVQA dataset brings together a vast collection of 52,480 slides, categorized into 9 distinct classes: Title, Page-Text, Obj-Text (text on graphical elements), Caption, Other-Text, Diagram, Table, Image, and Figure. The authors provide a train of 37,023 images, a validation set of 5,839 images, and a test set of 7,727 images.

4.2 Models

We selected FasterRCNN [26] with a ResNet50FPN [10,19] backbone and YoloV8 [12] as our models, noting that FasterRCNN tends to be more sensitive to class imbalance, while YoloV8 appears to show less sensitivity, probably due to the use of focal loss. We trained FasterRCNN with a batch size of 16 using the SGD optimizer, starting with a learning rate of 0.005 and a momentum of 0.95, coupled with a Cosine Annealing Learning Rate scheduler. Data augmentation techniques included color jittering, random gamma adjustments, and brightness and contrast modulation, alongside blurring. For YoloV8, we used a batch size of 64 and adhered to standard settings but changed the input resolution to 1280 and omitted mosaic and flip augmentations. We trained both models for 50 epochs with an early stopping (patience of 5 epochs).

4.3 Experiments

Dataset completion. Utilizing SlideCraft, we generate 25,000 synthetic slides for each dataset, enriching them with classes underrepresented in existing data, such as Diagram, Table, Plot, Handwriting, and Equation, by incorporating additional datasets for balance. We specifically target the underrepresented classes with the SlideCraft probability system while trying to fit the original dataset by adjusting the layout and style components. For example, in SlideVQA, to imitate the Obj-Text class, we run publicly available OCR Tesseract [13] on top of all the graphical elements. To imitate captions, we include a randomly generated smaller text close to the top or the bottom of a graphical element. We now compare the performance of each model trained on each dataset with and without added synthetic slides, evaluating their respective test datasets.

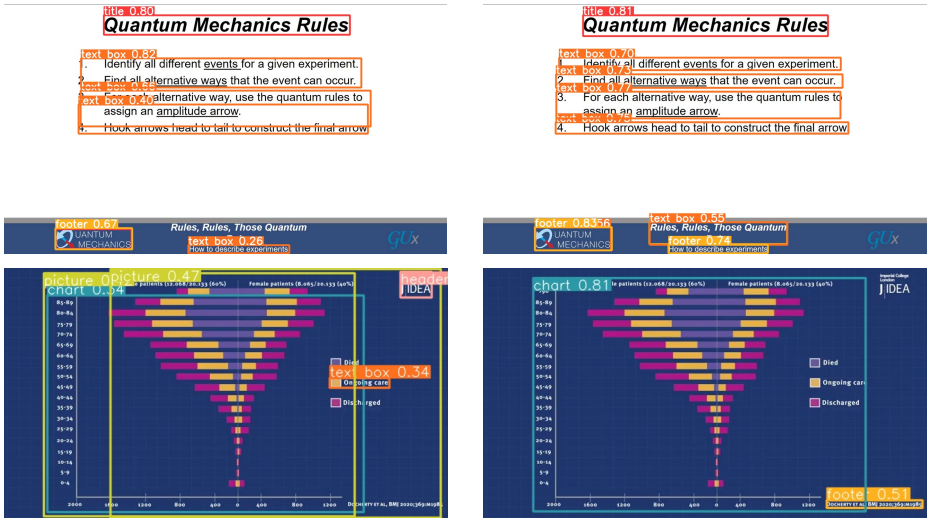

Fig. 7. Predictions on FitVid test dataset with our models. **Left**: Inference with YOLOv8 trained without added synthetic slides. **Right**: Inference with YOLOv8 trained with added synthetic slides. The model trained with synthetic data from SlideCraft demonstrates higher precision in box predictions, correlating with the marked improvement in mAP scores.

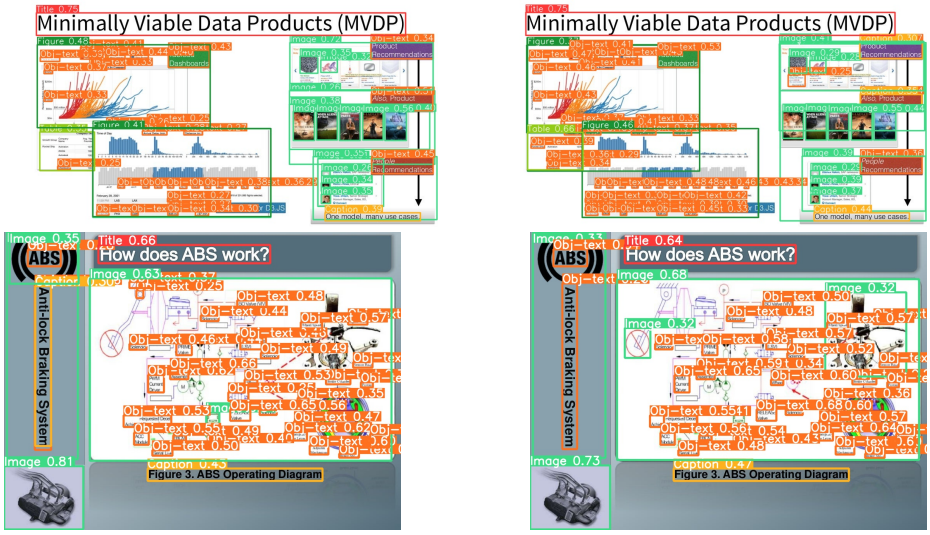

Fig. 8. Predictions on SlideVQA test dataset with our models. **Left**: Inference with YOLOv8 trained without added synthetic slides. **Right**: Inference with YOLOv8 trained with added synthetic slides. We can see that the model trained with synthetic data contradicts the ground-truth, despite looking accurate. For example, here, we predict more Obj-Text with added synthetic slides, but the ground truth in Figure 1 does not contain these elements.

Table 1. mAP50 of FasterRCNN and YoloV8 on SlideVQA test set and FitVid test set, training with and without the slides generated by SlideCraft.

Dataset	FasterRCNN			Yolov8		
	Original mAP50	Mix mAP50	Improvement	Original mAP50	Mix mAP50	Improvement
SlideVQA	0.641	**0.645**	0.59%	0.682	**0.685**	0.44%
FitVid	0.485	**0.530**	9.22%	0.513	**0.581**	13.26%

Our first results, displayed in Table 1, reveal that mixing SlideCraft-generated data into training can lead to superior performance. With synthetic slides, SlideVQA observes a slight improvement of the mAP50 by 0.59% and 0.44% with FasterRCNN and YoloV8 respectively. Conversely, FitVid benefits significantly from the synthetic slides, showing an improvement of the mAP50 by 9.22% and 13.26% for FasterRCNN and YoloV8, respectively. This gain is illustrated in Figure 7. These outcomes underscore the importance of large datasets for achieving optimal performance, as evidenced by the more pronounced improvements in FitVid compared to SlideVQA when training with an extended and larger dataset.

The marginal gains on SlideVQA suggest its extensive size may inherently limit performance improvements. However, we think that the dataset inconsistencies likely skew results. For example, we get a lower mAP50 (-0.3%) for the Obj-Text class on our model trained with SlideCraft's slides than the one trained solely with the original dataset. Despite this, our predictions, as highlighted in the images from Figure 8, show that we predict Obj-Text elements that should be correct based on how this class was defined. We also predict two Figures and a Table, which seem correct. However, as showcased in Figure 1, these Obj-Text elements do not exist in the ground truth, and the whole left part of the slide is labeled as Diagram. Such cases happen in other images, as highlighted by Figure 1.

Our first experiment shows that extending a dataset with SlideCraft-generated data can improve performance, especially on smaller datasets.

Ground Truth Data ratio. Furthermore, we conduct an additional experiment where we vary the ratio of real to synthetic slides from SlideVQA to see the impact of SlideCraft on different dataset sizes.

The findings from our second experiment, illustrated in Figure 9, reveal that smaller datasets experience more significant enhancements when trained with synthetic slides. Our performance only decreases by 2% when we train with 50% of the SlideVQA dataset augmented with synthetic slides, compared to training with the entire SlideVQA dataset.

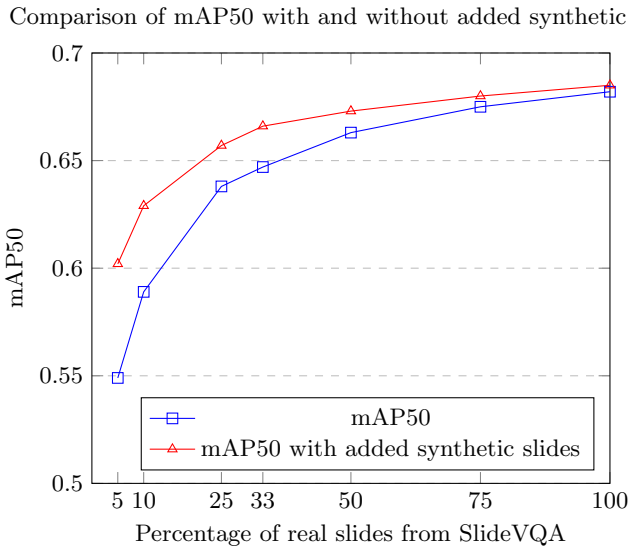

Fig. 9. Comparison of mAP50 obtained after training YoloV8 with and without added 25,000 SlideCraft's generated slides. Testing on SlideVQA test dataset. The results show a direct correlation between dataset size and the impact from adding synthetic slides: smaller datasets see greater benefits from the inclusion of synthetic slides.

This highlights SlideCraft's potential to significantly reduce the necessity for manually labeled slides in training object detection models, while potentially improving performance. This efficiency suggests a strategic approach in dataset annotation, emphasizing quality and consistency over sheer volume.

Balancing dataset. We also study whether balancing the datasets affects the performance of FasterRCNN. To achieve this, we created additional synthetic datasets comprising 25,000 slides each, deliberately designed to match the imbalance of the original datasets. The new sets were then combined with the original SlideVQA and FitVid datasets to observe the effects of skewed class distribution. We then train a FasterRCNN on the imbalanced datasets.

Table 2. Comparison of mAP50 obtained after training FasterRCNN on original dataset, balanced dataset and imbalanced dataset. The balanced and imbalanced dataset were created after adding SlideCraft's generated slides with different distribution. The distribution was chosen to balance or not the original dataset.

Dataset	Original	Imbalanced	Balanced
SlideVQA	0.641	0.637	**0,645**
FitVid	0.511	0.522	**0.530**

The results shown in Table 2 reveal that training on balanced datasets give a slightly better performance on FasterRCNN. For SlideVQA, the model trained on the imbalanced dataset has a mAP50 of 0.637 which is slightly lower than training only on the original SlideVQA dataset. The FasterRCNN trained on the more balanced dataset, however, displays a mAP50 of 0.645 which is slightly better. For FitVid, FasterRCNN gets better performance with added synthetic slides but the balanced dataset gets the best performance. This suggests that balancing the dataset has a positive impact on performance, while an imbalanced dataset can lead to worse performance.

This result underscores the value of SlideCraft's probability system in enabling dataset balancing.

5 Discussion and Future Work

We demonstrated that integrating SlideCraft's synthetic data generation enhances existing datasets while minimizing the need for extensive manual labeling. Prioritizing high-quality annotations and supplementing them with SlideCraft-generated data emerges as a strategic approach to boost model performance. However, challenges persist, notably the inconsistencies within and between datasets, particularly in how ambiguous classes such as diagrams are annotated, and the lack of standardized classes for dataset creation. SlideCraft offers a solution by accommodating these discrepancies, enabling researchers to tailor datasets to specific needs and applications, marking a step toward more standardized and effective dataset creation for object detection models. This adaptability paves the way for more focused and customized research, enhancing the overall quality and applicability of object detection within diverse contexts. In order to further enhance SlideCraft's capabilities, we see potential in refining the generation of specific classes, such as captions, headers, and footers and expanding the diversity of layouts and styles to better mirror real-world presentations. Although not extensively studied yet, SlideCraft also possesses the ability to generate other types of annotations such as pixel-level segmentation masks and slide to Markdown code, broadening its utility for visual analysis. Another potential use of SlideCraft is to enhance presentations' clarity and adaptability, aiming to improve accessibility for diverse audiences.

References

1. marp-team/marp (Feb 2024), https://github.com/marp-team/marp, original-date: 2018-03-25T12:47:38Z
2. Adcock, J., Cooper, M., Denoue, L., Pirsiavash, H., Rowe, L.A.: TalkMiner: a search engine for online lecture video. In: Proceedings of the international conference on Multimedia - MM '10. p. 1507. ACM Press, Firenze, Italy (2010). https://doi.org/10.1145/1873951.1874263, http://dl.acm.org/citation.cfm?doid=1873951.1874263
3. Araujo, A., Chaves, J., Lakshman, H., Angst, R., Girod, B.: Large-Scale Query-by-Image Video Retrieval Using Bloom Filters. arXiv **1604.07939** (2015)

4. Blanc-Beyne, T., Carlier, A., Mouysset, S., Charvillat, V.: Unsupervised human pose estimation on depth images. In: Machine Learning and Knowledge Discovery in Databases: Applied Data Science Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part IV. pp. 358–373. Springer (2021)
5. Capobianco, S., Marinai, S.: DocEmul: A Toolkit to Generate Structured Historical Documents. pp. 1186–1191 (Nov 2017). https://doi.org/10.1109/ICDAR.2017.196
6. Deng, Y., Kanervisto, A., Ling, J., Rush, A.M.: Image-to-markup generation with coarse-to-fine attention (2017)
7. Ferreira, A., Nowroozi, E., Barni, M.: VIPPrint: Validating Synthetic Image Detection and Source Linking Methods on a Large Scale Dataset of Printed Documents. Journal of Imaging **7**(3), 50 (Mar 2021). https://doi.org/10.3390/jimaging7030050, https://www.mdpi.com/2313-433X/7/3/50, number: 3 Publisher: Multidisciplinary Digital Publishing Institute
8. Haurilet, M., Al-Halah, Z., Stiefelhagen, R.: SPaSe - Multi-Label Page Segmentation for Presentation Slides. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 726–734. IEEE, Waikoloa Village, HI, USA (Jan 2019). https://doi.org/10.1109/WACV.2019.00082, https://ieeexplore.ieee.org/document/8659181/
9. Haurilet, M., Roitberg, A., Martinez, M., Stiefelhagen, R.: WiSe — Slide Segmentation in the Wild. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 343–348. IEEE, Sydney, Australia (Sep 2019). https://doi.org/10.1109/ICDAR.2019.00062, https://ieeexplore.ieee.org/document/8978089/
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition (Dec 2015). https://doi.org/10.48550/arXiv.1512.03385, http://arxiv.org/abs/1512.03385, arXiv:1512.03385 [cs]
11. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023)
12. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLO (Jan 2023), https://github.com/ultralytics/ultralytics
13. Kay, A.: Tesseract: an open-source optical character recognition engine. Linux Journal **2007**(159), 2 (Jul 2007)
14. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A Diagram is Worth a Dozen Images **9908**, 235–251 (2016). https://doi.org/10.1007/978-3-319-46493-0_15, http://link.springer.com/10.1007/978-3-319-46493-0_15, book Title: Computer Vision – ECCV 2016 ISBN: 9783319464923 9783319464930 Place: Cham Publisher: Springer International Publishing
15. Kim, J., Choi, Y., Kahng, M., Kim, J.: FitVid: Responsive and Flexible Video Content Adaptation. In: CHI Conference on Human Factors in Computing Systems. pp. 1–16. ACM, New Orleans LA USA (Apr 2022). https://doi.org/10.1145/3491102.3501948, https://dl.acm.org/doi/10.1145/3491102.3501948
16. Kim, J., Guo, P.J., Cai, C.J., Li, S.W.D., Gajos, K.Z., Miller, R.C.: Data-driven interaction techniques for improving navigation of educational videos. In: Proceedings of the 27th annual ACM symposium on User interface software and technology. pp. 563–572. UIST '14, Association for Computing Machinery, New York, NY, USA (Oct 2014). https://doi.org/10.1145/2642918.2647389, https://doi.org/10.1145/2642918.2647389

17. Li, M., Cui, L., Huang, S., Wei, F., Zhou, M., Li, Z.: TableBank: Table benchmark for image-based table detection and recognition. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 1918–1925. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.236
18. Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., Zhou, M.: DocBank: A Benchmark Dataset for Document Layout Analysis. In: Scott, D., Bel, N., Zong, C. (eds.) Proceedings of the 28th International Conference on Computational Linguistics. pp. 949–960. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.82, https://aclanthology.org/2020.coling-main.82
19. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature Pyramid Networks for Object Detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 936–944. IEEE, Honolulu, HI (Jul 2017). https://doi.org/10.1109/CVPR.2017.106, http://ieeexplore.ieee.org/document/8099589/
20. Liwicki, M., Bunke, H.: Iam-ondb-an on-line english sentence database acquired from handwritten text on a whiteboard. In: Eighth International Conference on Document Analysis and Recognition (ICDAR'05). pp. 956–961. IEEE (2005)
21. Masry, A., Do, X.L., Tan, J.Q., Joty, S., Hoque, E.: ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Findings of the Association for Computational Linguistics: ACL 2022. pp. 2263–2279. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.findings-acl.177, https://aclanthology.org/2022.findings-acl.177
22. Methani, N., Ganguly, P., Khapra, M.M., Kumar, P.: PlotQA: Reasoning over Scientific Plots (Feb 2020), http://arxiv.org/abs/1909.00997, arXiv:1909.00997 [cs]
23. Mukhopadhyay, S., Smith, B.: Passive capture and structuring of lectures. In: Proceedings of the seventh ACM international conference on Multimedia (Part 1) - MULTIMEDIA '99. pp. 477–487. ACM Press, Orlando, Florida, United States (1999). https://doi.org/10.1145/319463.319690, http://portal.acm.org/citation.cfm?doid=319463.319690
24. Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A.S., Staar, P.: DocLayNet: A Large Human-Annotated Dataset for Document-Layout Segmentation. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 3743–3751. ACM, Washington DC USA (Aug 2022). https://doi.org/10.1145/3534678.3539043, https://dl.acm.org/doi/10.1145/3534678.3539043
25. Pisaneschi, L., Gemelli, A., Marinai, S.: Automatic generation of scientific papers for data augmentation in document layout analysis. Pattern Recognition Letters **167**, 38–44 (Mar 2023). https://doi.org/10.1016/j.patrec.2023.01.018, https://www.sciencedirect.com/science/article/pii/S0167865523000247
26. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015), https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf

27. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 102–118. Springer (2016)
28. Tanaka, R., Nishida, K., Nishida, K., Hasegawa, T., Saito, I., Saito, K.: SlideVQA: A Dataset for Document Visual Question Answering on Multiple Images (Jan 2023). https://doi.org/10.48550/arXiv.2301.04883, http://arxiv.org/abs/2301.04883, arXiv:2301.04883 [cs]
29. Wang, J., Min, W., Hou, S., Ma, S., Zheng, Y., Jiang, S.: Logodet-3k: A large-scale image dataset for logo detection. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **18**(1), 1–19 (2022)
30. Xu, C., Wang, R., Lin, S., Luo, X., Zhao, B., Shao, L., Hu, M.: Lecture2Note: Automatic Generation of Lecture Notes from Slide-Based Educational Videos. In: 2019 IEEE International Conference on Multimedia and Expo (ICME). pp. 898–903 (Jul 2019). https://doi.org/10.1109/ICME.2019.00159, iSSN: 1945-788X
31. Yoo, T., Jeong, H., Lee, D., Jung, H.: LectYS: A System for Summarizing Lecture Videos on YouTube. In: 26th International Conference on Intelligent User Interfaces. pp. 90–92. ACM, College Station TX USA (Apr 2021). https://doi.org/10.1145/3397482.3450722, https://dl.acm.org/doi/10.1145/3397482.3450722
32. Zhao, B., Xu, S., Lin, S., Wang, R., Luo, X.: A New Visual Interface for Searching and Navigating Slide-Based Lecture Videos. In: 2019 IEEE International Conference on Multimedia and Expo (ICME). pp. 928–933 (Jul 2019). https://doi.org/10.1109/ICME.2019.00164, iSSN: 1945-788X
33. Zhong, X., Tang, J., Jimeno-Yepes, A.: PubLayNet: Largest Dataset Ever for Document Layout Analysis (Sep 2019). https://doi.org/10.1109/ICDAR.2019.00166