SPaSe – Multi-Label Page Segmentation for Presentation Slides

Monica Haurilet Ziad Al-Halah Rainer Stiefelhagen
Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany
{haurilet, ziad.al-halah, rainer.stiefelhagen}@kit.edu
https://cvhci.anthropomatik.kit.edu/data/SPaSe/

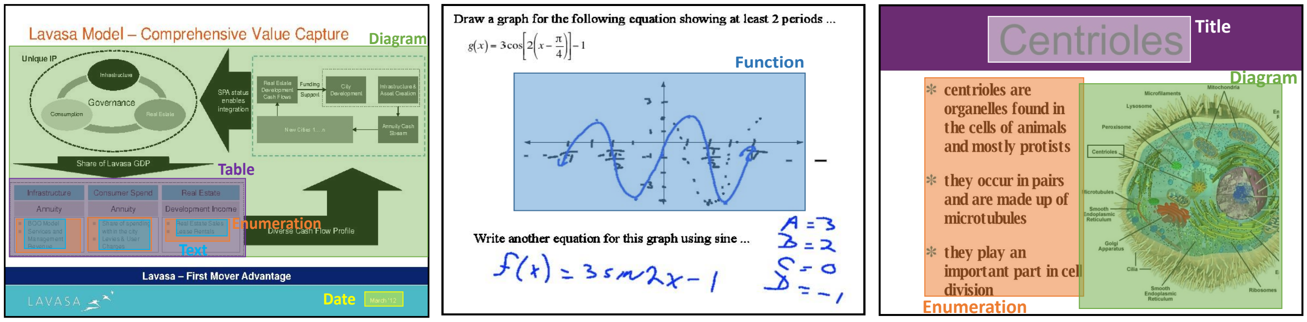

Lavasa Model – Comprehensive Value Capture
Diagram
Unique IP
Infrastructure
Governance
Consumption
Real Estate
SPA status enables integration
Real Estate Development Cash Flows
Funding Support
City Development
Infrastructure & Asset Creation
New Cities 1 … n
Annuity Cash Stream
Share of Lavasa GDP
Table
Infrastructure
Consumer Spend
Real Estate
Annuity
Annuity
Development Income
BOO Model Services and Management Revenue
Share of spending within the city Levies & User Charges
Real Estate Sales Lease Rentals
Enumeration
Diverse Cash Flow Profile
Text
Lavasa – First Mover Advantage
LAVASA
Date
March '12
Draw a graph for the following equation showing at least 2 periods …
$g(x) = 3\cos\left[2\left(x - \frac{\pi}{4}\right)\right] - 1$
Function
Write another equation for this graph using sine …
f(x) = 3sin2x − 1
A = 3
B = 2
C = 0
D = −1
Centrioles
Title
Diagram
centrioles are organelles found in the cells of animals and mostly protists
they occur in pairs and are made up of microtubules
they play an important part in cell division
Enumeration
Microfilaments
Mitochondria
Lysosome
Peroxisome
Centrioles
Microtubules
Smooth Endoplasmic Reticulum
Golgi Apparatus
Cilia
Smooth Endoplasmic Reticulum
Ribosomes

Figure 1: Slide Page Segmentation (SPaSe) dataset contains fine-grained annotations of 25 different classes for 2000 images.

Abstract

*We introduce the first benchmark dataset for slide-page segmentation. Presentation slides are one of the most prominent document types used to exchange ideas across the web, educational institutes and businesses. This document format is marked with a complex layout which contains a rich variety of graphical (*e.g.* diagram, logo), textual (*e.g.* heading, affiliation) and structural components (*e.g.* enumeration, legend). This vast and popular knowledge source is still unattainable by modern machine learning techniques due to lack of annotated data. To tackle this issue, we introduce SPaSe (Slide Page Segmentation), a novel dataset containing in total dense, pixel-wise annotations of 25 classes for 2000 slides. We show that slide segmentation reveals some interesting properties that characterize this task. Unlike the common image segmentation problem, disjoint classes tend to have a high overlap of regions, thus posing this segmentation task as a multi-label problem. Furthermore, many of the frequently encountered classes in slides are location sensitive (*e.g.* title, footnote). Hence, we believe our dataset represents a challenging and interesting benchmark for novel segmentation models. Finally, we evaluate state-of-the-art segmentation networks on our dataset and show that they are suitable for developing deep learning models without any need of pre-training. The annotations will be released to the public to foster further research on this interesting task.*

1. Introduction

In page segmentation, the goal is to extract the semantic components of a document pages (*e.g.* historical documents [3], magazines [12, 4] or scientific papers [11, 36]) represented in a digital format, *i.e.* images. These components are usually related to the layout structure (*e.g.* headers and footnotes) or to the page contents (*e.g.* tables and diagrams). In this work, we approach the page segmentation task as a *pixel-wise* classification task where each pixel in the page is classified into a subset of predefined categories.

Page segmentation is quite relevant to the popular image segmentation task [16] where realistic images from indoor or outdoor environment are segmented into one of the defined object categories like person, bike or building. Nonetheless, there are some key differences between these two segmentation problems. For example, unlike image segmentation where we usually deal with relatively large scale objects like car, road, sky; in page segmentation we need to handle a fine-grained set of classes that usually occupy only a tiny spatial area of the image like footnote, page number or legend (Figure 1 left). Furthermore, the semantics of components in image segmentation is *location invariant* (*e.g.* a car is a car regardless whether it appears on the top, left or right part of the image). However, in page segmentation some components may represent different semantics depending on their spatial location or scale; for example, *Centrioles* in Figure 1 (right) appears as both a *Title* and a *Bullet-Point*. Additionally, we notice a high overlap of pixel labels in page segmentation where a pixel can belong to multiple categories at the same time; for example, in Figure 1 (left) we have pixels that are part of four different classes: Text, Bullet-Point, Table and Diagram.

In this paper, we focus on a special type of documents; namely presentation slides. Slides are perhaps one of the most popular formats to share information and ideas especially for educational and business purposes. For example, the prominent slide sharing service SlideShare [26] claims to have more than 400 thousand presentations uploaded monthly [17] with an estimated 80 million unique visitor per month [26]. Thus, in order to tap to this massive knowledge base it is crucial to enable automatic analysis approaches for this type of document format. Automatic document understanding and retrieval will enable millions of users to have fast and convenient access to the sought information. Moreover, slides are used massively in many educational institutes, hence it is quite important to enable the students with visual impairment to have a convenient and reliable access to this knowledge source. Nonetheless, to the best of our knowledge there is still no publicly available dataset that would enable the vision research community to tackle this important problem. The availability of such a dataset is vital to developing novel and accurate approaches for slides automatic understanding and visual analysis. Additionally a large-scale and diverse dataset is crucial to benchmark and enable modern machine learning technique like deep learning.

We introduce SPaSe, a page segmentation dataset for presentation slides, which augments the publicly available Slideshare-1M dataset [5] with segmentation labels. Our dataset has fine-grained annotations of 25 classes (*e.g.* title, drawing and table) for 2000 slide images. The dataset has a high intra-class variance where, for example, plots and text can be both computer generated and handwritten (Figure 1). Additionally, the collected slides are multilingual where in addition to English there are languages like French, Vietnamese and Romanian. This creates an interesting and challenging benchmark for tasks like text segmentation. While it is common in image segmentation to have a single label annotation per pixel (*e.g.* [16, 13, 25]), this is inadequate to slide segmentation due to the multifacet nature of the objects in the slides as explained earlier. Thus, we provide multi-label annotation of individual pixels to capture the overlapping semantic representations of objects. Furthermore, we define novel evaluation metrics to quantify the performance of multi-label slide segmentation approaches. In a thorough evaluation on SPaSe, we show that our large-scale dataset enables deep learning methods to be trained from scratch without the need of additional data sources. Moreover, we analyze the correlation of the defined categories and location and demonstrate the impact of spatial information on segmentation performance. Finally, we make the annotations publicly available to the research community and hope that our dataset will encourage further research towards developing new and exciting methods for slide segmentation.

2. Related Work

Semantic image segmentation, which deals with segmenting objects in natural images, is a popular and important task in computer vision with a variety of applications in robotics, autonomous cars *etc*. This field shows a rapid advancement in novel machine learning approaches, especially deep learning methods [37, 24, 23, 31, 38]. These advancements were fueled by the availability of publicly available benchmarks [25, 13, 16, 7, 29].

In comparison, page segmentation methods did not benefit as much as semantic segmentation methods. In page segmentation, we have for one the bottom-up [2, 20, 22, 15] and the top-down approaches [19, 18], which use feature engineering and mostly do not necessitate any training data (*e.g.* by using unsupervised approaches like clustering or thresholding). Deep learning models were introduced for both segmenting scientific papers [35, 6, 36] and historical documents [9, 8]. Even though we have some neural network based models, some are shallow containing only few layers [9, 8] and others compensate the lack of data by generating synthetic pages [36]. Most of the available datasets provide either a small number of pages [11, 34, 36, 28] or have a small number of classes [11] (see an overview in Table 1) which make them not adequate for deep learning.

RDCL [4] is a layout recognition dataset, which contains in total 7 training and 70 test pages from magazines and journals. In this dataset various text based classes are labeled: caption, credit, paragraph, page number, heading, drop capital, footer and floating text. However, the number of image-based classes is low and contains solely decoration and images. The DSSE-200 [36] dataset considers six classes: figures, tables, section headings, captions, list and paragraphs, of which it provides bounding box annotations of 200 pages. A larger dataset was introduced by Tao *et al.* [34]. This dataset contains in total 244 pages extracted from 35 English and Chinese e-books. Sectlabel [28] offers bounding box annotations on a variety of text-based classes for pages extracted from the pdf of scientific papers. The CS-150 and CS-Large [11] are publicly available datasets including in total 150 and 3100 pages respectively. While CS-150 only provides annotations of papers released on three conferences, CS-Large contains randomly selected papers from Semantic Scholar that have more than 9 citations. The annotations consist of bounding boxes of text body, figures, tables and the corresponding figure captions.

In comparison to the presented datasets, we provide pixel-wise annotations, where we allow our annotators to provide more than one class annotation per pixel. To the

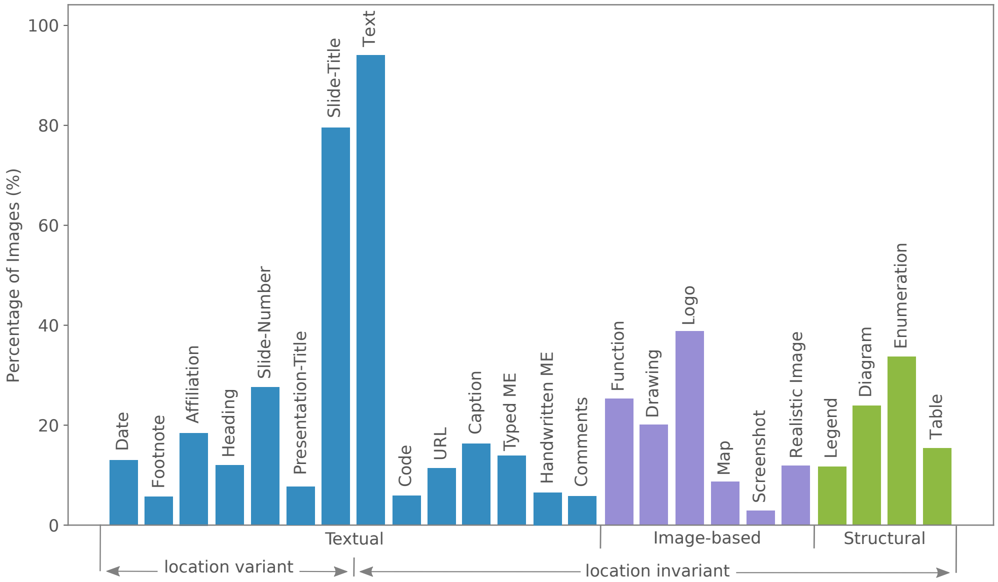


Figure 2: Overview of the class distribution on our dataset. We split the classes by their root-categories: textual, image-based and structural components.

best of our knowledge, SPaSe is the first page segmentation dataset for presentation slides. Page segmentation on slides offers a challenging task to segment regions of complex layouts with a large variety of images and text formats.

3. Slide Page Segmentation Dataset

We introduce the *first* dataset for visual slide segmentation. Presentation slides are one of the main document formats shared across educational and business platforms. Our goal is to enable automatic visual understanding for this popular format by providing a suitable and challenging benchmark. Next, we provide detailed description of the data collection and annotation process (Sec. 3.1) along with an insightful analysis of the dataset features and characteristics (Sec. 3.2).

3.1. Data Collection and Annotation

Categories. We start by defining the set of categories used to annotate the slides' pixels. For that purpose, we collaborate with an academic institute which offers aid services for students with visually impairment. One of their main activities is to provide detailed description of lecture materials like papers, exams and presentation slides. This is usually achieved through manual effort of tens of trained assistances. Theses assistances will go through the large amount of slides and supply structured descriptions that are tailored to the needs of the visually impaired in regards of the slides structure and content. These descriptions represent an excellent data source to identify the most frequent and important object types that are usually encountered in the slides. Based on this data source and discussions with the experts, we identified 25 relevant categories. In addition to the Background class, these categories can be split into 3 main groups (see Figure 2): a) 14 text classes (*e.g*. Title, Pseudocode and Footnote); b) 7 figure classes (*e.g*. Plot, Map and Logo); c) 3 structural classes (*e.g*. Table, Diagram and Enumeration);

Slides. We sample our slide images from the public dataset for image retrieval Slideshare-1M provided by Araujo *et al*. [5]. To ensure the diversity of the collected samples, we select a maximum of one slide per presentation. This will help us to reduce overfitting problems when training segmentation models that might simply memorize specific templates present in the data. Moreover, we restrict the sampled slides to have a minimum of 200 × 350 resolution. In total, we collected 2000 slide images. These are split into 1400 samples for training, 100 for validation, and 500 images for testing.

Annotation. We provide fine-grained annotations at *pixel-level* for our 25 classes. This level of granularity is needed for our type of data since many of the classes represent fine structures in the slides like *Page Number* or *Date*. Using bounding boxes, for example, would not be adequate since many of the background pixels will be wrongly annotated with the foreground class which will significantly increase the noise in the training data. Additionally, we notice that there is a high region overlap between the categories. For example, a table may contain enumerations and a diagram may contain text, drawing or even tables in its nodes (Figure 1 left). This property is especially present in the structure classes. Hence, we decide to use multi-label annotation in our data where a pixel may belong to multiple overlapping categories at the same time. To that end, we develop an annotation tool based on the one used for [7]. Each slide image is first divided into fine superpixels using simple linear iterative clustering (SLIC) [1]. SLIC leverages both visual and spatial information of pixels in a weighted distance measure that controls the size and compactness of the superpixels. Since we have very fine structures in our data, we set the superpixels extracted by SLIC to be relatively small (13000 superpixel per image). Finally, the output of SLIC is used by the annotators to classify each of the superpixels into the 25 classes. To show the performance of this annotation method, we annotate 100 images by 3 different annotators in a similar manner as [7, 32]. We obtain a mean pairwise label agreement of 77% over the annotators, similar to COCO-Stuff [7] with an agreement of 74% and ADE-20K [32] with 67%.

3.2. Dataset Properties

3.2.1 Distribution of Images per Class

In Figure 2, we see the distribution of the classes over the images in our dataset. Most of the classes appear in 10% to 30% of the slides showing a relatively balanced distribution. Two of the most prominent classes in more than 75% of the data are *Text* and *Slide-Title*. Among the least frequent classes (in less than 6% of the data) are *Footnote*, *Code* and *Screenshots*. Additionally, in around 12.3% of the data we encounter handwritten text (*e.g. Comments* and *Handwritten Mathematical Expression (ME)*). This is, for example, one of the unique properties of slides in comparison to other document formats like papers or magazines where text is always typed.

3.2.2 Overlapping Regions

Next, we analyze the frequency of the pairwise occurrences of different classes in our dataset (see Figure 3). In this figure, we only show a subset of the classes that have a strong overlap among each other. Not surprisingly, we have a strong overlap with the text class, since especially structure classes like diagrams and tables contain text. Interestingly, we have more legends assigned to plots than to maps. We also have noticed that handwritten text (comments) are often written on plots and tables to provide additional information for these complicated illustrations. Furthermore, we have enumerations assigned to typed mathematical expressions, which are usually included in math homework slides. Natural (realistic) images and drawings usually do not have an overlap, but drawings are more likely to appear within a diagram than realistic images. A further discussion of co-occurrence between our categories can be found in the supplemental material.

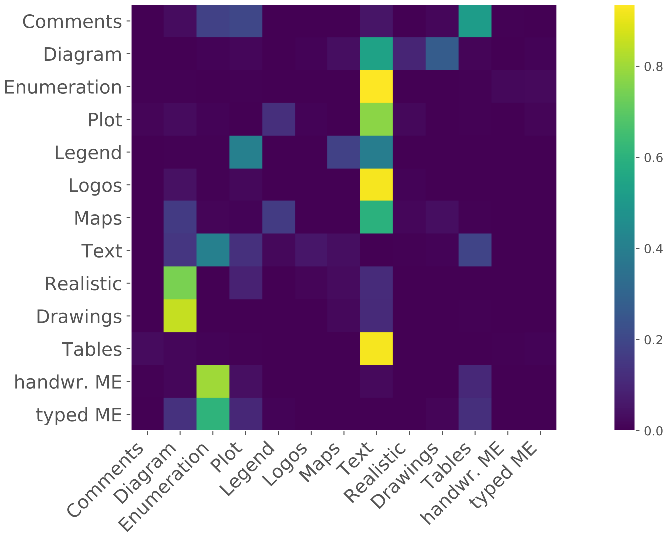


Figure 3: Frequency of pair-wise Pixel Overlap between a subset of the SPaSe classes.

3.2.3 Location Heat Maps

In Figure 4 we show the occurrence frequency in the page of ten of our classes. We see that classes like title, slide number and logo have a strong position prior. Especially we have text that changes its meaning dependent on its size and position like title, slide number and footnote. For example, numbers located in the page body are either mathematical expressions or just belong to the text body. While, we consider this to be page number if it is stand-alone text at the corners of the page. The same we have for slide title at the top and footnotes at the bottom of the image. In the case of slide title we also notice the larger font that is typically used, and thus simply classifying the text at the top as title is not sufficient. Thus, the font size of the surrounding text is also important to be able to classify slide titles.

Interestingly, as we see in the legend heat map, the legends are mostly positioned in the right and bottom side. Also, we are able to recognize the programming code by the short line breaks that we see in most programming languages. Not surprisingly the background is mostly located at the borders of the page as we see in Figure 4a and due to the positioning and the high frequency of titles less at the top than at the bottom of the page.

3.2.4 SPaSe versus other Segmentation Datasets

We compare in Table 1 SPaSe with other publicly available datasets used for page segmentation. We show the document type offered in each dataset with the corresponding number of pages and the number of classes for text, images

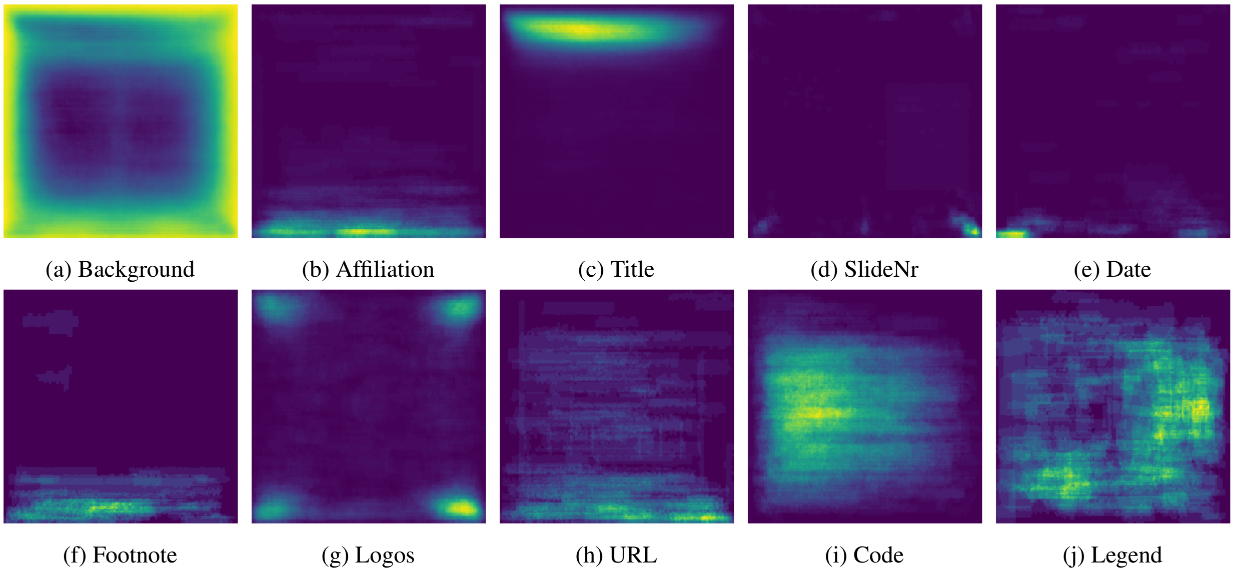


Figure 4: Location Heat Maps

and structural elements (*e.g.* tables, lists). While CS-Large has the largest number of pages, it also has a lowest number of classes. It provides annotations for only 4 classes: figures, tables, image captions and text. In comparison, SPaSe has the second highest number of pages, while providing pixel-wise annotations of more than 20 classes. Moreover, a key feature of our dataset in comparison to the others is that we provide overlapping class annotations of pages. The closest dataset in this regard is the semantic segmentation dataset ADE-20k [32], which allows an overlap between maximum of two subsets of classes (objects and parts of objects). In comparison to them, we allow any possible combination of classes, especially since multiple combination (*i.e.* more than 2) of region classes can occur.

3.2.5 Overlap per Pixel

Another aspect of the dataset is the amount of overlap, as this influences some of our metrics where a prediction is considered correct if and only if all classes of the pixels are classified correctly. We show in Table 2 the number of pixels in the entire dataset for different amount of overlap.

Overlap	1	2	3	4	5
Number pixels	130M	65M	3.6M	180K	2K

Table 2: Number of pixels with an overlap between 1 and 5

As we see, we have 130M pixels (excluding the background class) with no overlap, and thus in these cases one-class segmentation would have sufficed. However, in around 70M pixels this would have failed, as we have an overlap between at least one pair of classes. By far the highest of the overlap is with an overlap of 2, where we have 65M pixels, around a third of non-empty pixels in our dataset. The maximum overlap that SPaSe has is 5, which we have in around 2K pixels.

3.3. Evaluation Metrics

As explained earlier, in our special segmentation task we have multi-label pixel-wise annotations. This is different than the common segmentation problem where classes are assumed to be mutually exclusive [25]. Hence, besides the popular mean Intersection over Union (mIOU) metric [16], we define three additional evaluation metrics for multi-label segmentation, namely: mean balanced accuracy (mbACC), pixel accuracy (pAcc), and pixel intersection over union (pIOU). Let H, W and C be the height, width of the image and the number of categories, respectively. Then, $L^k, P^k \in \{0,1\}^{H\times W\times C}$ are the ground truth annotations and predictions for image k.

Pixel Accuracy (pACC). We define the number of correctly labeled values for a pixel (i,j) in image k as: $t^k_{i,j} = \sum_c 1[L^{k,c}_{i,j} = P^{k,c}_{i,j}]$. Then the pixel accuracy for image k is defined as the percentage of pixels which has an exact match with ground truth annotations:

$$pAcc^k = \frac{1}{H\cdot W}\sum_{i,j} 1[t^k_{i,j} = C]. \tag{1}$$

While this metric gives us an idea of the pixel-wise segmentation accuracy it has the following drawbacks: 1) The

Type	Dataset	# Pages	# Text Cls.	# Image Cls.	# Structure Cls.	Overlapping Segm.
Magazines	RDCL 2017 [12]	70	8	2	0	✗
E-Books	CM [34]	244	12	1	2	✗
Papers	CS-150 [11]	150	2	2	0	✗
	DSSE-200 [36]	200	2	1	2	✗
	SectLabel [28]	347	20	1	2	✗
	CS-Large [11]	3100	2	2	0	✗
Slides	SPaSe (Ours)	2000	14	6	4	✓

Table 1: Comparison of various datasets for page segmentation.

metric harshly punishes partially correct segmentation by assigning a zero accuracy for pixels if one label is misclassified; and 2) the metric might be biased towards the most frequent annotations in case of unbalanced distribution of classes.

Pixel Intersection over Union (pIOU). To tackle the first drawback, we define the pIOU metric, which softens the weight given for incorrect classifications:

$$pIOU^k = \frac{1}{H \cdot W} \sum_{i,j} \frac{n^k_{i,j}}{s^k_{i,j} - n^k_{i,j}}, \tag{2}$$

where $s^k_{i,j} = \sum_c (L^{k,c}_{i,j} + P^{k,c}_{i,j})$ and $n^k_{i,j} = \sum_c 1[L^{k,c}_{i,j} = 1 \wedge P^{k,c}_{i,j} = 1]$. This metric is conceptually similar to the mIOU but for multi-label predictions at the pixel-level.

Mean Balanced Pixel Accuracy (mbAcc). The mbAcc tackles the second drawback through using a class-based weighted accuracy measure:

$$bAcc^k_c = \sum_{\ell \in \{0,1\}} (1 - \alpha^c_\ell) \cdot \sum_{i,j} 1[L^{k,c}_{i,j} = \ell \wedge L^{k,c}_{i,j} = P^{k,c}_{i,j}], \tag{3}$$

where the weights α^c_ℓ is proportional to the number of pixels labeled with class c and label ℓ. That is,

$$\alpha^c_\ell = \frac{1}{T} \cdot \sum_{k,i,j} 1[L^k_{i,j} = \ell],$$

where T is the total number of pixels in the test split. Finally, the mean balanced Accuracy is obtained by averaging the balanced accuracy across all classes:

$$mbAcc^k = \frac{1}{C} \sum_c bAcc^k_c. \tag{4}$$

4. Evaluation

4.1. Methods

Baselines. We consider two simple baselines for the multi-label slide segmentation task: 1) Uniform: a baseline that samples the output class of each pixel from a uniform distribution over all possible classes; and 2) Background: where the output class is set to the most frequent category, *i.e.* the background class.

FCN-8s [27]. The Fully Convolutional Neural Network (FCN) is a deep learning model for semantic segmentation that consists of a VGG-based encoder [33] pretrained on ImageNet [14], and a decoder with multiple upscore layers for segmentation prediction.

FRRN [30]. Unlike previous deep models, the Fully Resolutional ResNet (FRRN) leverages two processing streams. While the first stream processes the input image with progressively larger receptive fields, the second stream leverages residual connections to keep the feature maps at a high resolution.

DeepLab [10]. DeepLab consists of multiple atrous-spatial convolution [21] layers to enlarge the receptive field, but keeping the feature dimensions. These are used in a pyramid pooling to extract features at multiple scales, thus capturing small objects and image context.

Learning Setup. Since for the multi-label segmentation task we need to predict multiple classes per pixel, we replace the softmax output layer in the previous deep models with a sigmoid activation function and train these models using binary cross entropy loss. We train all these models for 50 epochs using the same optimizers as they were originally used for image segmentation in [27, 30, 10]. Then, the model with the highest mIOU over validation set is selected for final evaluation.

4.2. Slide Segmentation

Multi-Label Segmentation. In this experiment, we measure the performance of the variant methods in pixel-wise multi-label slide segmentation. That is, each pixel is classified into a subset of the 25 categories defined in our dataset.

We show in Table 3 the results of the baselines in the first section, while in the second one we show the performance of the deep neural networks. We see that the baselines perform by far worse than the deep learning models; such that

Model	mIOU	pAcc	pIOU	mbAcc
Uniform	1.1	3.4	4.0	50.0
Background	2.5	61.6	61.6	50.0
FCN-8s [27]	20.0	66.2	73.5	62.0
FRRN-A [30]	28.4	69.5	73.8	67.0
FRRN-B [30]	30.9	71.2	75.3	68.5
FRRN-B [30] + Loc.	33.2	73.4	77.2	70.1
DeepLab [10]	34.1	76.5	80.3	71.2
DeepLab [10] + Loc.	**35.8**	**77.4**	**81.2**	**72.6**

Table 3: Multi-label Segmentation Results on our test set.

the neural architectures improve over the baselines between 17.5% (FCN) and 33.3% (DeepLab) in terms of mIOU.

FRRN	mIOU	pAcc	mAcc
Uniform	29.8	50.0	50.0
Background	41.8	83.7	50.0
FCN [27]	80.7	93.7	90.4
FRRN-A [30]	80.4	93.7	88.9
FRRN-B [30]	**83.2**	**94.6**	**91.7**
DeepLab [10]	82.6	94.5	90.5

Table 4: Text Detection

The FCN-8s tackles the downsampling of the feature maps through so called upscore layers, which however, have difficulty in capturing fine-grained structures. Both FRRN versions were able to outperform the mIOU achieved by FNC-8s. The deeper FRRN-B was able to improve the mIOU by 2% in comparison to the shallower FRRN-A. On the other hand, DeepLab achieved the highest performance with 34.1%, further improving over FRRN-B with an additional 3%. Finally, we notice that the difference between FRRN-B and DeepLab in mIOU is smaller than in pIOU and pAcc. This shows that the gain was mostly influenced by improvements to the most frequent classes. In comparison, we notice that FRRN-A performance boost over FCN-8s is mainly due to better segmentation of the less frequent classes as depicted by the higher gain in mbAcc compared to pAcc.

Since convolution layers are partially translation invariant, our models cannot use the location information, which are important in our location variant classes. Thus, we tackle this problem by concatenating a 2-channel map (x and y positions) to our input image with the location of each pixel. This simple modification improved the DeepLab model to an mIOU of 35.8% and FRRN-B to 33.2% (see DeepLab+Loc. and FRRN-B+Loc in Table 3).

FRRN	mIOU	pAcc	mAcc
Uniform	4.6	16.7	16.7
Background	14.3	86.1	16.7
FCN [27]	39.4	87.9	52.1
FRRN-A [30]	44.1	90.8	52.2
FRRN-B [30]	47.2	91.6	56.4
DeepLab [10]	**50.7**	**91.8**	**63.5**

Table 5: Graphics Segmentation

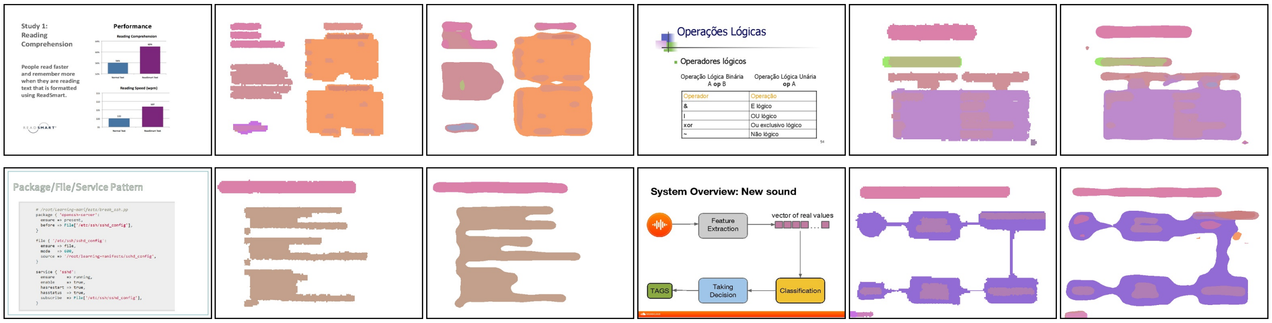
Text and Graphics Segmentation. Next, we explore other popular page segmentation tasks on our dataset: 1) text detection (see Table 4) and 2) graphics segmentation (see Table 5). The text detection task consists of classifying each pixel into two classes: text and non-text. In comparison, graphics segmentation is a more difficult task, where we classify each pixel into five different image classes: drawings, realistic images, plots, maps and logos.

While the baselines have a low performance of only 41.8% for the two-class segmentation task and 14.3% for graphics segmentation, the deep learning models improve results by more than 40% and 30% respectively. We see in Table 4 that FRRN-B has the best results on the text detection task, while DeepLab outperforms it for graphics segmentation. The residual stream architecture of FRRN-B is able to keep the feature maps to a constant size. Thus, it is able to segment fine structured objects like the text class. In comparison, DeepLab uses a pyramid of different sizes of receptive fields. This leads to better segmentation of continuous homogeneous regions that characterize the graphical classes. However, on the other hand, this also leads to worse performance in capturing the fine details of text.

Qualitative Results. Finally, in Figure 5 we show some example slides (left) with both ground truth annotations (middle) and predictions (right) on the multi-label task. We see that DeepLab was able to recognize difficult classes like plot, table, diagram and code correctly. However, it has some difficulties to get the exact borders for fine structures like the arrows of the diagram (see last image in Figure 5). Also, it has some difficulties to segment the right most node at the top of the diagram probably due to the short interruption, especially as the text of the node was detected. The programming code was segmented and classified correctly as can be seen in the second row. The text classes are mostly correctly segmented with some small exceptions.

5. Conclusion

In this work, we introduce the first dataset for pixel-wise slide page segmentation SPaSe. We annotated in total 2000 slide images, which are labeled with 25 different classes including 3 structural and 7 image based classes. Additiona-



Slide GT Prediction Slide GT Prediction

Figure 5: Slide examples along with ground truth annotations and predictions produced by our best multi-label model.

lly, we provide a thorough analysis of the data properties and unique features. In the evaluation, we establish some strong baselines and demonstrate the suitability of our dataset for developing deep learning model from scratch. Moreover, we show that our classes are partially location variant and thus, including pixel spatial location leads to improved segmentation results.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. 4

[2] A. Amin and R. Shiu. Page segmentation and classification utilizing bottom-up approach. *International Journal of Image and Graphics*, 1(02):345–361, 2001. 2

[3] A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher. Icdar 2013 competition on historical book recognition (hbr 2013). In *12th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1459–1463. IEEE, 2013. 1

[4] A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher. Icdar2015 competition on recognition of documents with complex layouts-rdcl2015. In *13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1151–1155. IEEE, 2015. 1, 2

[5] A. Araujo, J. Chaves, H. Lakshman, R. Angst, and B. Girod. Large-Scale Query-by-Image Video Retrieval Using Bloom Filters. *arXiv*, 1604.07939, 2015. 2, 3

[6] T. M. Breuel. Robust, simple page segmentation using hybrid convolutional mdlstm networks. In *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 733–740. IEEE, 2017. 2

[7] H. Caesar, J. Uijlings, and V. Ferrari. Coco-stuff: Thing and stuff classes in context. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 4

[8] K. Chen, M. Seuret, J. Hennebert, and R. Ingold. Convolutional neural networks for page segmentation of historical document images. In *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 965–970. IEEE, 2017. 2

[9] K. Chen, M. Seuret, M. Liwicki, J. Hennebert, and R. Ingold. Page segmentation of historical document images with convolutional autoencoders. In *13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1011–1015. IEEE, 2015. 2

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018. 6, 7

[11] C. Clark and S. Divvala. Pdffigures 2.0: Mining figures from research papers. In *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, pages 143–152. IEEE, 2016. 1, 2, 6

[12] C. Clausner, A. Antonacopoulos, and S. Pletschacher. Icdar2017 competition on recognition of documents with complex layouts-rdcl2017. In *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1404–1410. IEEE, 2017. 1, 6

[13] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3213–3223, 2016. 2

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009. 6

[15] D. Drivas and A. Amin. Page segmentation and classification utilising a bottom-up approach. In *Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 610–614. IEEE, 1995. 2

[16] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 1, 2, 5

[17] C. Gaffney. Linkedin slideshare introduces clipping, saving your favorite content just got easier. 2

[18] J. Ha, R. M. Haralick, and I. T. Phillips. Document page decomposition by the bounding-box project. In *Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 1119–1122. IEEE, 1995. 2

[19] J. Ha, R. M. Haralick, and I. T. Phillips. Recursive xy cut using bounding boxes of connected components. In *Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 952–955. IEEE, 1995. 2

[20] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006. 2

[21] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian. A real-time algorithm for signal analysis with the help of the wavelet transform. In *Wavelets*, pages 286–297. Springer, 1990. 6

[22] F. Lebourgeois, Z. Bublinski, and H. Emptoz. A fast and efficient method for extracting text paragraphs and graphics from unconstrained documents. In *11th IAPR International Conference on Pattern Recognition. Conference B: Pattern Recognition Methodology and Systems*, pages 272–276. IEEE, 1992. 2

[23] C. Liang-Chieh, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *International Conference on Learning Representations (ICLR)*, 2015. 2

[24] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3194–3203, 2016. 2

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision (ECCV)*, pages 740–755. Springer, 2014. 2, 5

[26] LinkedIn. Linkedin slideshare website. 2

[27] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3431–3440, 2015. 6, 7

[28] M.-T. Luong, T. D. Nguyen, and M.-Y. Kan. Logical structure recovery in scholarly articles with rich document features. *Multimedia Storage and Retrieval Innovations for Digital Library Systems*, 270:2, 2012. 2, 6

[29] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy*, pages 22–29, 2017. 2

[30] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe. Full-resolution residual networks for semantic segmentation in street scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6, 7

[31] A. G. Schwing and R. Urtasun. Fully connected deep structured networks. *arXiv preprint arXiv:1503.02351*, 2015. 2

[32] T. Shen, G. Lin, C. Shen, and I. Reid. Learning multi-level region consistency with dense multi-label networks for semantic segmentation. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017. 4, 5

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[34] X. Tao, Z. Tang, C. Xu, and Y. Wang. Logical labeling of fixed layout pdf documents using multiple contexts. In *2014 11th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 360–364. IEEE, 2014. 2, 6

[35] C. Tensmeyer and T. Martinez. Document image binarization with fully convolutional neural networks. *arXiv preprint arXiv:1708.03276*, 2017. 2

[36] X. Yang, E. Yumer, P. Asente, M. Kraley, D. Kifer, and C. L. Giles. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 6

[37] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[38] Z. Zhang, A. G. Schwing, S. Fidler, and R. Urtasun. Monocular object instance segmentation and depth ordering with cnns. *arXiv preprint arXiv:1505.03159*, 2015. 2