WiSe – Slide Segmentation in the Wild

Monica Haurilet Alina Roitberg Manuel Martinez Rainer Stiefelhagen
Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany
{haurilet, alina.roitberg, manuel.martinez, rainer.stiefelhagen}@kit.edu
https://cvhci.anthropomatik.kit.edu/data/WiSe/

***Abstract*—We address the task of segmenting presentation slides, where the examined page was captured as a live photo during lectures. Slides are important document types used as visual components accompanying presentations in a variety of fields ranging from education to business. However, automatic analysis of presentation slides has not been researched sufficiently, and, so far, only preprocessed images of already digitalized slide documents were considered. We aim to introduce the task of analyzing unconstrained photos of slides taken during lectures and present a novel dataset for Page *Se*gmentation with slides captured in the *Wi*ld (WiSe). Our dataset covers pixel-wise annotations of 25 classes on 1300 pages, allowing overlapping regions (i.e., multi-class assignments). To evaluate the performance, we define multiple benchmark metrics and baseline methods for our dataset. We further implement two different deep neural network approaches previously used for segmenting natural images and adopt them for the task. Our evaluation results demonstrate the effectiveness of the deep learning-based methods, surpassing the baseline methods by over 30%. To foster further research of slide analysis in unconstrained photos, we make the WiSe dataset publicly available to the community.**

***Keywords*-Page Segmentation, Presentation Slides, Dataset**

I. Introduction

Presentation slides are highly useful for introducing a topic to an audience in an intuitive way. The importance of slides has grown rapidly from being a supplement to lecturers' speech and printed papers to being the main pathway for dissemination of knowledge, e.g., in university classes. A downside of this medium is the difficult automated analysis and information extraction, as slides are more visual and vary greatly in their structure, layout and relations of the entities, which can become very complex. While understanding and automatically converting the knowledge in a digital form is more difficult for slides than for conventional textbooks, solving this problem would make information more accessible, having a substantial impact for people struggling with visual impairment or blindness.

Page segmentation is an essential prerequisite step for most of document understanding tasks, which identifies distinct portions of the document and assigns them a semantic meaning. Problems in this task vary greatly in terms of the segment types: ranging from simple binarization which classifies a pixel to page or non-page; to *semantic* page segmentation with a high number of diverse content classes, such as heading, title, URLs and tables.

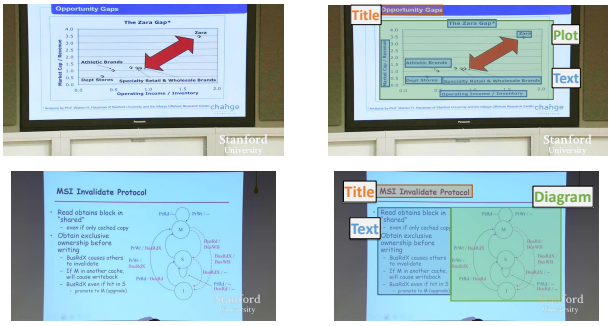


Figure 1: Example of slides captured in the wild (left) with the corresponding class annotations (right).

While multiple works targeted the problem of page segmentation in conventional documents, the research of slide analysis has been sparse so far. Compared to previously published methods, which focus on papers [1], [2], magazines [3], e-books [4] etc., page segmentation of slides encounters a far higher complexity of page structure with various possible visual classes (e.g., drawings and handwritten text).

Although page segmentation on presentation slides has been recently addressed by Haurilet et al. [5] for the first time, the authors use images of digital presentation slides, which were automatically extracted from PDFs. However, slide documents are not always offered to the viewers beforehand. Besides, the listener might not know, which of the slides is currently presented making such presentations very hard to follow for visually impaired people. It is unclear, how such methods perform in an uncontrolled setting, where the slide image does not originate from a preprocessed document, but, for example, from a display photo taken during a lecture. A model for segmenting such "slides in the wild" should be able to additionally handle problems related to natural images, such as noise, varying illumination conditions, and different rotation and scaling (see Figure 1).

In this work, we tackle the problem of slide segmentation in the wild, where the slide does not originate from a digital document but is contained in a part of a *natural* image. Our overall application goal is a model, which allows a person to take a live photo of a slide, e.g., from a display during

a lecture, and then provides the digitalized version of the contained information.

The *WiSe* dataset for Slide ***Se***gmentation in the ***Wi***ld contains covers pixel-wise annotations for 25 different classes on 1300 pages captured during lectures. Since a pixel can belong to multiple independent classes (see example in Figure 1 with an overlap between *text* and *plot*), we enable the overlapping annotations, i.e., a pixel can have multiple labels. This poses an additional challenge, as the single-class constraint that is usually present in semantic segmentation, does not hold anymore. We employ and systematically compare multiple neural architectures in our setting, and report promising results for unconstrained slide segmentation, which, to the best of our knowledge, is addressed for the first time.

Our work aims to bring the task of slide segmentation to a realistic scenario of unconstrained photos and has the following major contributions:

1) We introduce a novel task of page segmentation on slides captured during lectures and present the $WiSe$ dataset, containing pixel-wise annotations with 25 classes for 1300 naturalistic slide images.
2) We evaluate two baseline methods as well as popular deep learning methods for segmentation on the WiSe dataset and show that the deep networks are able to achieve strong results on various metrics proposed for overlapping page segmentation.
3) To foster further research in this task we make the WiSe dataset publicly available to the community.

II. Related Work

Page Segmentation Datasets. In the field of document analysis, various datasets for page segmentation and word spotting were proposed in recent years (overview in Table I). As we see, there are datasets ranging from documents with a simple layout like papers published in a small set of conferences to difficult word spotting on street images.

RDCL [3] consists in total of 7 images in the training set and 70 test images from magazines and journals. In this dataset, a high variety of semantic text classes are provided, including caption, page number, heading and footer. A larger dataset was introduced in [4] collected from 35 English and Chinese e-books.

Page segmentation of scientific papers is a very popular topic as we see by the high number of different datasets proposed for this task (see Table I). Both SectLabel and DSSE-200 offer coarse bounding box annotations on 350 pages, and 200 pages, respectively. CS-150 is a publicly available dataset containing 150 pages extracted from papers published at three different conferences. In comparison, CS-large includes 20 times more pages extracted from randomly selected papers from Semantic Scholar. The annotations are non-overlapping bounding boxes of four classes: text body and image captions, the figure and table class.

Dataset	# Pages	# T-Cls.	# I+S-Cls.	Wild	Pxw.	Ovl.
Magazines						
RDCL17 [3]	77	10	2	✗	✓	✗
E-Books						
CM [4]	244	12	3	✗	✓	✗
Papers						
CS-150 [1]	150	2	2	✗	✗	✗
DSSE-200 [2]	200	2	3	✗	✗	✓
SectLabel [7]	347	20	3	✗	✗	✓
CS-Large [1]	3100	2	2	✗	✗	✗
Street-View						
SVT [6]	350	1	0	✓	✗	✓
Presentation Slides						
SPaSe [5]	2000	14	10	✗	✓	✓
WiSe (ours)	1335	14	10	✓	✓	✓

Table I: Overview of datasets for page segmentation. We show the number of text, graphical and structural classes for each dataset and if the images were captured by a camera (i.e. are captured in the 'wild'). We point out which datasets contain pixel-wise segmentation and have overlapping annotations (i.e. enabling the possibility of a pixel belonging to multiple classes).

The SVT [6] dataset contains 350 images captured in street view. The annotations are in the form of bounding boxes of the words found in natural images like street names and traffic signs (i.e. has a single class). In comparison, SPaSe [5] contains pixel-wise, overlapping annotations of 25 semantic classes on presentation slides. The images in this dataset are difficult since these consist of a high variety of layouts, fonts and figure types.

We also consider segmenting presentation slides like SPaSe, however, in our case the slides were captured by a camera during lectures and not directly extracted from their digital format. Thus, the problem of page segmentation of slides in the wild combines the difficulty of both SVT and SPaSe, as its pages has a varied and difficult layout while also including noise, rotation and illumination variance encountered in natural images. WiSe includes a high number of classes in total 14 semantic text classes (e.g. title, text body, mathematical expressions), 7 image classes (e.g. natural images, drawings) and 3 structural classes (tables, diagrams and enumerations). With the rise of deep learning, the necessity of large-scale datasets has grabbed the attention of the computer vision community. Thus, we made sure that we have a large number of non-overlapping slides, totaling to over 1.3K pages.

Page Segmentation. Most methods for page image segmentation can be categorized into bottom-up [8]–[11] (that combine small regions like super-pixel or even pixel into semantic meaningful regions) and top-down approaches [12], [13] (that split up the page in multiple regions). These methods mostly make use of feature engineering and do not necessitate any training data by applying methods like clustering or thresholding. In comparison, deep networks

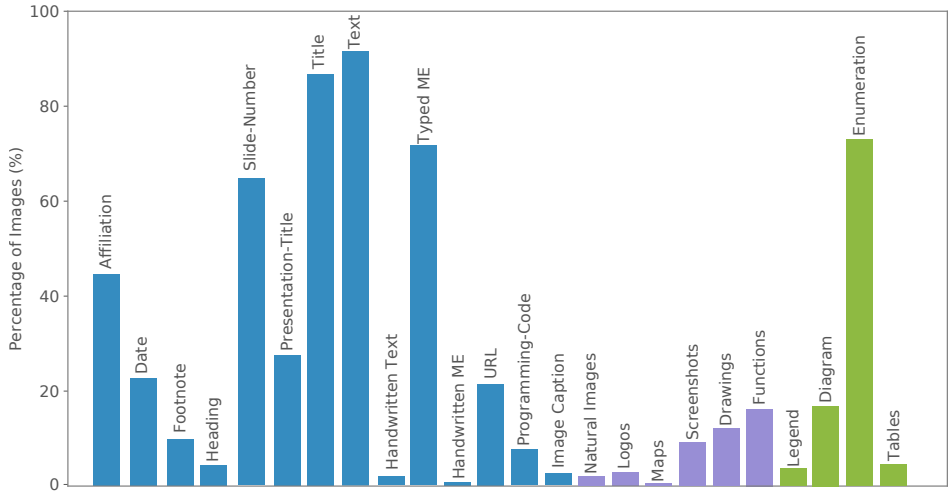


Figure 2: Overview of the class distribution in the WiSe dataset. We group the classes into three groups: text-based, graphical and structural regions. Our dataset covers 25 distinct classes allowing overlapping regions.

were successfully applied for segmenting pages of scientific papers [2], [14], [15] and historical documents [16], [17] achieving state-of-the-art results. However, deep neural networks necessitate a large amount of data in order to obtain this high performance. In page segmentation, the proposed approaches address the data scarcity, by either using shallow networks containing only few layers [16], [17], or by compensating the lack of data by generating synthetic pages [2].

Models applied on the WiSe dataset have to tackle both the difficult layouts of presentation slides and have to deal with difficult conditions of natural images, including strong variations in rotation, translation and illumination (see Figure 1). Nonetheless, the models are able to obtain a high performance on WiSe even without making use of pre-trained models, as it often is the case in semantic segmentation approaches [18], [19].

Segmentation on Natural Images. Semantic segmentation deals with the classification of each pixel of a natural image into a semantic class. Page segmentation is closely related to semantic segmentation in the sense that it assigns semantically similar classes into the same group. Deep models tackling this task usually have an hourglass-shape with an encoder and a decoder module. First, the model *encodes* the image by using a CNN with pooling or convolution layers with a stride larger than one. This ensures the rapidly increase of the receptive field, while removing unnecessary information in the space coordinate. At the end of the encoder only a small version of the image remains i.e. the height and width of the final tensor is by far smaller than the original image. As page segmentation requires that the input image has the same height and width as the predicted segmentation, a decoder is used to transform the feature maps back to the initial size. The output of the final layer has the same height and width as the input image, while the number of channels is equal to the number of classes (i.e. each pixel in the output models a probability distribution). We make use of the popular hour-glass neural networks (FCN [18] and DeepLab [20]) on our dataset and show that these models are able to provide good results on our semantic page segmentation task.

III. The WiSe Dataset

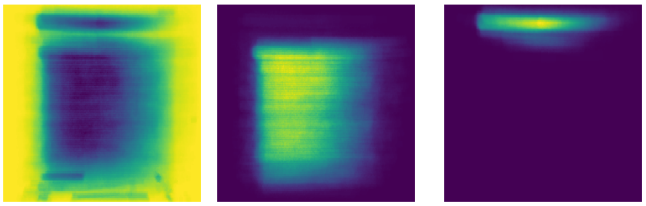
A. Annotation Protocol

In total, we annotated 1300 slides which we split into 300 images for testing, 100 for validation and the remaining 900 slides are kept for training. To help reduce redundancy and, thus, possible over-fitting of our models, we made sure that there are no identical slides in the dataset. We extracted the slides in our dataset from the publicly available Class-X lecture dataset provided by Araujo et al. [21], which we manually label using pixel-level annotations for our 25 selected classes. Fine-grained annotations are especially important in case of slides as we have graphics and structures that often do not have a quadratic shape (see examples in Figure 1). We used the annotation tool provided by [22] that was used for annotating objects in natural images.

In comparison to semantic segmentation on natural images, where it is common to label each pixel with a single label, in case of pages in documents we deal with regions of multiple distinct classes (e.g., a region can be *text*, *diagram* and part of an *enumeration* at the same time). Thus, we enable our annotators the possibility to label any pixel with multiple classes offering more flexibility in the annotation process.

B. Dataset Properties

Next, we look into the collected data by analyzing the class distribution and frequency of overlapping regions.



(a) Background (b) Enumeration (c) Title

Figure 3: The distribution of three of our classes on the page. We show that some classes (e.g., title) are strongly dependent on the location i.e., most labels are in a specific region within the page.

Distribution of Images per Class. In Figure 2, we show the distribution of the different classes in our dataset by grouping them into textual, graphical and structure based regions. As we see, the *text* label is by far the most frequent class with more than 90% of the available slides contain some type of text. The text label is followed mostly by more specific text classes like *title*, *slide number* and *printed mathematical expressions* which occur in over 60% of our images. Not surprisingly classes like *maps* and *handwritten mathematical* expressions are not very common, but still are present in some slides.

Location Heat-Maps. In comparison to semantic segmentation on natural images where the class of an object does not change dependent on the location in the image, in case of pages some classes are strongly location variant. Even though photos of slides captured using a camera have strong tilts and translations, we still see that the slides are mostly centered in the image, as some components are clustered in some regions in the figure (see Figure 3). For instance, large text on top of the page is very frequent the title of the slide, while document components like diagrams and enumerations are located in the center of the page, in comparison to footnotes, which are mostly at the bottom. Finally, the background class can be seen in the first image showing that most content is localized in the center with a small interruption due to the title class at the top of the page.

Overlapping Regions. As regions in our dataset can belong to multiple classes at the same time (e.g., drawings can be nodes in a diagram and therefore these pixels belong to both classes) we enabled our annotation tool to label a pixel with multiple classes. Even though most pixels are classified with a single label, we have over 47 Million pixels that are labeled with at least two classes, showing that this overlapping property is important for slide segmentation.

IV. Segmentation Methods

Baselines. We evaluate two baselines on the slide segmentation task: 1) *Uniform*, which picks each class with equal probability; and 2) *Background*, which chooses the most frequent category as the output class, i.e. the background class.

FCN [18]. The Fully Convolutional Network (FCN) is a neural network for semantic segmentation that comprises of an encoder consisting of several convolutional layers [23] pre-trained on ImageNet [24] (i.e. the features are continuously down-sampled by maxpooling layers), and a decoder with multiple upscale layers to resize the features to the initial size of the input image.

DeepLab [20]. DeepLab contains multiple dilated convolution layers [25] to enlarge the receptive field, but at the same time keep the feature dimensions. The pyramid pooling then extracts features at multiple scales, and thus captures small objects and image context.

V. Evaluation

A. Learning Setup

We split our dataset into 900 training images, 100 for validation and 300 images are used for testing. Since we deal with multi-label segmentation, we normalize our predictions with sigmoid activation opposed to the more popular softmax normalization for semantic segmentation. All models are trained at most 50 epochs with the same optimizer and hyperparameters as they were originally used for the segmentation task on natural images. For testing, we choose the models with the highest mIOU over the validation set.

Model	mIOU	pAcc	mAcc
Baseline Methods			
Uniform	29.8	50.0	50.0
Background	42.4	84.9	50.0
Neural Networks			
FCN-8s [18]	83.7	95.2	91.4
DeepLab [20]	**84.4**	**95.5**	**91.3**

Table II: Text Segmentation Results on our test set.

B. Text Segmentation

We analyze the first task on the WiSe dataset is segmenting the text from the presentation slides. This task is very similar to the popular text spotting on natural images (e.g. street name, logo on clothing). Table II illustrates the results of our baseline methods and the deep learning networks on this task. We show the mean Intersection Over Union (mIOU), pixel Accuracy (pAcc) – accuracy over all pixels in the image and mean Accuracy (mAcc) that calculates the mean accuracy over all classes. These are popular metrics for non-overlapping semantic segmentation [18].

Not surprisingly, since the text segmentation uniform baseline only chooses between two classes, it is able to

Image Ground Truth Prediction

Figure 4: Example of pixel-wise labeled slides: Raw images from lectures collected by [21] (left) extended with corresponding manual annotations in our WiSe dataset (center) and the predictions of the DeepLab Model (right).

achieve a pixel and mean Accuracy of 50%. The prior baseline, that always predicts the most frequent class (i.e. background) is able to achieve a higher pAcc of 83%. While FCN is able to improve the baseline methods by over 40%, DeepLab achieves the best performance of a mIOU of 84.4%.

C. Binarization

We show in Table III the results of our models on the binarization task, which consists of determining if a pixel belongs to the page or background class. Not surprisingly, the pixel Accuracy for the background-only baseline is smaller in the binarization task in comparison to the text segmentation, since the background class is more frequent in the text segmentation class. The DeepLab architecture is able to outperform the FCN network by 2% in mIOU, while improving the baseline methods by over 50%.

Model	mIOU	pAcc	mAcc
Baseline Methods			
Uniform	32.9	50.0	50.0
Background	38.1	76.1	50.0
Neural Networks			
FCN-8s [18]	84.3	93.7	91.4
DeepLab [20]	**86.3**	**94.5**	**92.8**

Table III: Binarization Results on our test set, where we aim to classify each pixel into two classes: page or non-page.

D. Semantic Slide Segmentation

The most difficult of the proposed settings is the semantic slide segmentation where our models aim to assign each pixel to one or multiple labels from the 25 available classes. Since we are performing page segmentation with overlapping regions, we use the same metrics as in [5]. As can be seen in Table IV, the baselines perform poorly especially in mIOU with a performance of 3% and 0.2%, respectively. By using as input solely the slide image in RGB format, DeepLab is able to achieve an accuracy of 35.8% improving the baseline methods by over 30%. Since some components in the page are strongly location variant (e.g., title, footnote) and fully convolutional networks are translation invariant, we conducted an experiment where we obtain as input additionally to the RGB values of each pixel its location in the page (DeepLab+L). This new setting is able to achieve an improvement over the RGB-only model of 1.4% in mIOU.

Model	mIOU	pAcc	pIOU	mbAcc
Baseline Methods				
Uniform	0.2	0.0	0.0	50.0
Background	3.0	76.1	76.1	50.0
Neural Networks				
FCN-8s [18]	18.3	81.7	59.8	59.8
DeepLab [20]	35.8	88.3	**90.4**	72.2
DeepLab [20]+L	**37.2**	**88.5**	**90.4**	**72.8**

Table IV: Results of the semantic segmentation on our test set. We group our approaches into baseline methods and deep learning approaches.

E. Qualitative Results

In Figure 4, we show example slides (left) with their corresponding manually annotated segmentation (center) and prediction from the DeepLab model (right). As we see, the

model is able to correctly predict difficult classes like enumeration and diagrams. Even though the model sometimes has problems to recognize the correct edges of the current segment (e.g., the contours of the diagram) it is able to produce a smoother boundary in comparison to the manual annotation (e.g., text- and enumeration-labels). Moreover, the model was able to recognize the document components in frontal slides (top) as well as in tilted ones (bottom).

VI. Conclusion

In this work, we introduced the WiSe dataset for presentation slides segmentation captured during lectures. We annotated in total 1300 images with fine-grained labels of 25 different classes: 3 structural, 7 image-based and 14 textual-classes. Furthermore, we provide an in-depth analysis of our collected data and discuss its interesting properties. In the evaluation, we compare baseline methods to deep learning based approaches that were originally designed for semantic segmentation on natural images. Finally, we show that the deep learning models are able to achieve a strong performance on all three tasks: text segmentation, binarization and semantic segmentation, improving baseline methods by over 30% in mIOU.

References

[1] C. Clark and S. Divvala, "Pdffigures 2.0: Mining figures from research papers," in *Joint Conference on Digital Libraries (JCDL)*, 2016.

[2] 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," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] C. Clausner, A. Antonacopoulos, and S. Pletschacher, "Icdar2017 competition on recognition of documents with complex layouts-rdcl2017," in *International Conference on Document Analysis and Recognition (ICDAR)*, 2017.

[4] X. Tao, Z. Tang, C. Xu, and Y. Wang, "Logical labeling of fixed layout pdf documents using multiple contexts," in *IAPR International Workshop on Document Analysis Systems (DAS)*, 2014.

[5] M. Haurilet, Z. Al-Halah, and R. Stiefelhagen, "Spase - multi-label page segmentation for presentation slides," *IEEE Winter Conference on Applications of Computer Vision*, 2018.

[6] K. Wang and S. Belongie, "Word spotting in the wild," in *European Conference on Computer Vision (ECCV)*, 2010.

[7] 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*, 2012.

[8] A. Amin and R. Shiu, "Page segmentation and classification utilizing bottom-up approach," *International Journal of Image and Graphics*.

[9] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*.

[10] F. Lebourgeois, Z. Bublinski, and H. Emptoz, "A fast and efficient method for extracting text paragraphs and graphics from unconstrained documents," in *International Conference on Pattern Recognition. Conference B: Pattern Recognition Methodology and Systems*, 1992.

[11] D. Drivas and A. Amin, "Page segmentation and classification utilising a bottom-up approach," in *International Conference on Document Analysis and Recognition (ICDAR)*, vol. 2. IEEE, 1995, pp. 610–614.

[12] 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)*, 1995.

[13] ——, "Recursive xy cut using bounding boxes of connected components," in *International Conference on Document Analysis and Recognition (ICDAR)*, 1995.

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

[15] T. M. Breuel, "Robust, simple page segmentation using hybrid convolutional mdlstm networks," in *International Conference on Document Analysis and Recognition (ICDAR)*, 2017.

[16] 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)*, 2015.

[17] K. Chen, M. Seuret, J. Hennebert, and R. Ingold, "Convolutional neural networks for page segmentation of historical document images," in *International Conference on Document Analysis and Recognition (ICDAR)*, 2017.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[19] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015.

[20] 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," *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.

[21] A. Araujo, J. Chaves, H. Lakshman, R. Angst, and B. Girod, "Large-Scale Query-by-Image Video Retrieval Using Bloom Filters," *arXiv*, vol. 1604.07939, 2016.

[22] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[25] 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*, 1990.