NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

Document Layout Analysis via Machine Learning Technique

Liang Xu Chao
U1920092B

Supervisor: Dr Loke Yuan Ren

School of Computer Science and Engineering
2023

Table of Content

Abstract

Document layout analysis (DLA) is to understanding the physical structure of a document and identifying and classifying the different regions and components of a document image. There are many different layout formats for documents, and a universal model to understand all of them is not yet available. DLA usually consists of pre-processing, model training, post-processing, and performance evaluation. This project will follow these four steps and build a model that can detect 11 different components in digital, scanned, or photographed documents.

Deep learning techniques will be used for model training and prediction. This project will use the Mask R-CNN approach and the experimental results will be discussed, and the final model achieved an average Intersection over Union (IoU) of 86% on digital documents while also providing insightful detection on photos of documents.

To further enhance prediction accuracy on photo documents, this project will introduce a second model that will be trained on manually labeled images to first find the bounding box of the entire document in the image and then apply the prediction on the bounding box area.

Acknowledgements

I would like to express my deepest appreciation to my supervisor, professor Loke Yuan Ren for guiding and keeping me stay on the right track in this final year project. I have pushed myself to achieve better result based on his suggestions. Without his help, this project would not be possible.

Liang Xu Chao

List of Figure

List of Table

1. Introduction

Document Layout Analysis (DLA) can be considered as a form of object detection for images. However, it is a unique challenge due to the varied and complex layouts found in different types of documents. DLA involves categorizing and detecting different regions and then applying different methods for detecting the content based on the type of region, such as text, table or figure.

1.1 Motivation

Existing solutions for DLA often perform well on one type of digital document but struggle when presented a layout that is completely different. In addition, these solutions often work poorly on documents which capture as a photo. Thus it motivates me to to explore train a model that can handle as many layouts as possible, regardless of whether they are digital or captured in a photo.

1.2 Objectives and Scope

The objective of this project is to build a generic model for document layout detection, which will require improvements in three key areas.

1.2.1 Data Preparation

To handle a wide range of document layouts, a substantial dataset of labeled images will be required. This dataset should include all major document components.

1.2.2 Training Technique

This project will explore different training tools and techniques to identify the most effective approach.

1.2.3 Prediction Technique

To improve the accuracy of predictions, this project will train a model to predict only on the document itself. Once the document has been predicted, the figure can be extracted and the generic layout prediction model can be applied to predict the layout.

2. Literature Review

The Literature Review will cover the relevant DLA concepts, review existing solutions and analyse available dataset for model training.

2.1 DLA concepts

DLA can be classified as a task of object detection for images. It is a changeling topic due to the complex layouts from different types of documents. Solutions for document analysis strategy often follow the Bottom-Up or Top-Down method. Bottom-Up strategy first locate local information and connected component to get the words then merge into line then merge into paragraph then into column. Top-Down strategy will start with analyse the global information of the document like the black and white stripes. It uses the information to break the document into different region and splitting them into paragraph then line then words.

To perform DLA usually will require following steps:

Pre-processing

Pre-processing is an important step in document layout detection that involves enhancing the quality of the input image through techniques like image enhancement, noise reduction, and image binarization. Image enhancement adjusts the contrast, brightness, and sharpness of the input image. Noise reduction removes noise and artifacts, while image binarization converts the image to a binary format, making it easier to detect and segment different components of the document.

Model training

Selecting the appropriate algorithm for training is crucial in achieving good results in image recognition. Machine learning models, specifically convolutional neural networks (CNNs), are commonly used for document layout detection. CNNs use a sliding window approach to scan the input image and classify each patch of the image as a different component of the document. CNNs have demonstrated high performance in detecting and classifying document components. To choose a correct algorithm for training is important to achive good result in image recongition.

Post-processing

Post-processing techniques are used to refine the output of the machine learning model by removing false positives and improving the accuracy of the detected components. Post-processing techniques involve filtering, clustering, and merging of detected components. Filtering techniques are used to remove false positives by applying various rules and heuristics to the detected components. Clustering techniques group similar components together, while merging techniques combine adjacent components that belong to the same document component.

2.2 Existing solutions

2.2.1 Detection models

Rule-Based Model

One of the commonly used methods for DLA is the rule-based approach. In this approach, the rules are defined based on the characteristics of each document type to identify and extract the layout components. Xu et al. (2019) proposed a rule-based method for the detection of text regions in historical documents. The attainment of a F1 score of 0.93 on the ICDAR 2017 dataset through the proposed method indicates the efficiency of the rule-based technique in identifying and extracting layout components in certain types of documents.

However, the rule-based approach has limitations as it heavily relies on the predefined rules and may not be suitable for complex layout structures. Therefore, researchers have proposed machine learning-based approaches that can automatically learn the layout features and predict the components of the documents.

Conditional Random Fields (CRF)

Another machine learning-based approach for DLA is the Conditional Random Fields (CRF). CRF is a probabilistic graphical model that can model the dependencies among the layout components. Chen et al. (2016) proposed a CRF-based approach for DLA that can detect text, image, and table regions in business documents. The effectiveness of the CRF-based technique in detecting layout components across diverse document types is supported by the attainment of a F1 score of 0.83 on the ICDAR 2013 dataset through the proposed method.

Convolutional Neural Network (CNN)

CNN is a deep learning technique that has demonstrated outstanding performance in various computer vision tasks, including DLA. Gupta et al. (2016) proposed a CNN-based approach for DLA that can detect text, image, and table regions in scientific articles. The effectiveness of the proposed method in detecting diverse layout structures in scientific articles is demonstrated by achieving a F1 score of 0.89 on the PubLayNet dataset.

Fully Convolutional Network (FCN)

FCN is a deep learning technique that can perform pixel-level prediction of the layout components. Hu et al. (2017) proposed an FCN-based approach for DLA that can detect text, image, and table regions in scientific articles. The FCN-based approach proposed in the study achieved an F1 score of 0.91 on the PubLayNet dataset, indicating its effectiveness in handling complex document layouts.

U-Net

U-Net is another deep learning technique that can perform pixel-level prediction of the layout components. Ronneberger et al. (2015) proposed a U-Net-based approach for DLA that can detect text regions in forms.

Faster R-CNN

Faster R-CNN is an object detection architecture consisting of a Region Proposal Network (RPN) and a Fast R-CNN network. The RPN generates object proposals by sliding a window over a convolutional feature map and predicting objectness scores and bounding box coordinates. These proposals are then fed into the Fast R-CNN network, which classifies the proposals and refines their bounding boxes. Faster R-CNN has achieved state-of-the-art performance on benchmark datasets and is known for its ability to generate high-quality object proposals efficiently.

Mask R-CNN

Mask R-CNN is an advanced architecture of convolutional neural network (CNN) primarily used for performing image segmentation tasks. It is built upon Faster R-CNN, a well-known object detection model. Compared to Faster R-CNN, Mask R-CNN has an additional branch that predicts segmentation masks for each region of interest (RoI) in the image, in addition to the existing branches for classification and bounding box regression.

Mask R-CNN follows a similar two-stage detection process as Faster R-CNN. In the first stage, a set of candidate RoIs is generated using a Region Proposal Network (RPN). In the second stage, these RoIs are fed into a series of fully connected layers to perform classification and bounding box regression.

Apart from the classification and regression branches, Mask R-CNN also features a third branch dedicated to segmentation masks. This mask branch takes each RoI and produces a binary mask that outlines the object's shape within the RoI. The mask branch is a fully convolutional network that works on a coarse feature map of the RoI, and generates a mask with the same resolution as the input image.

Mask R-CNN has achieved remarkable performance on various benchmark datasets for instance segmentation tasks such as COCO, Cityscapes, and Pascal VOC. Its success is attributed to its ability to perform both object detection and instance segmentation simultaneously in a single framework, which simplifies the training and inference processes, and facilitates end-to-end learning of both tasks.

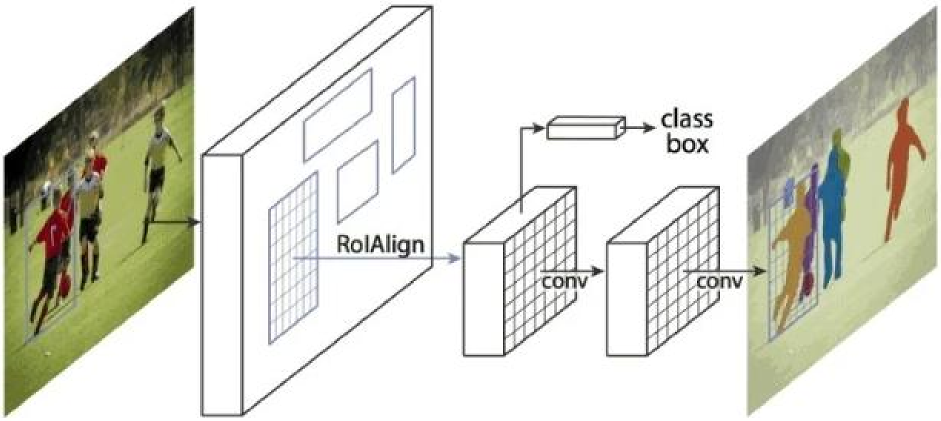


Figure 1: Mask R-CNN architecture

YOLO

YOLO (You Only Look Once) is a popular real-time object detection system that uses a single convolutional neural network (CNN) to simultaneously predict object classes and bounding boxes. Unlike other object detection systems that apply a sliding window approach to detect objects at multiple locations and scales, YOLO divides the input image into a grid of cells and predicts the object class probabilities and bounding boxes for each cell.

2.2.2 Training tools

TensorFlow Object Detection API

This is an open-source framework for building, training, and deploying object detection models. It provides pre-trained models for a wide range of object detection tasks and supports several popular object detection architectures such as Faster R-CNN, SSD, and YOLO.

PyTorch

This is another open-source deep learning framework that supports building custom object detection models. PyTorch offers a flexible and easy-to-use interface for designing and training deep neural networks and has become increasingly popular in recent years.

Detectron2

This is a modular object detection library built on PyTorch that provides a comprehensive set of pre-trained models and tools for training custom models.

Detectron2 supports several state-of-the-art object detection architectures such as Faster R-CNN, Mask R-CNN, and RetinaNet.

Keras

This is a high-level neural network API that can run on top of TensorFlow, Theano, and other popular deep learning libraries. Keras provides an easy-to-use interface for building and training deep neural networks and includes pre-trained models for several popular object detection tasks.

Caffe

The Berkeley Vision and Learning Center developed Caffe which is a deep learning framework. Caffe is particularly well-suited for image and video recognition tasks and supports several popular object detection architectures such as Faster R-CNN and SSD.

Darknet

This is an open-source neural network framework written in C and CUDA that supports a wide range of deep learning tasks, including object detection. Darknet includes an implementation of YOLO, a popular object detection architecture known for its speed and accuracy.

Detectron2

Detectron2 is considered as a next-generation object detection tool which introduce by Facebook' AI research group. It is open-source in github based on the caffe2 framework which now is part of Pytorch. Detectron2 supports a number of computer vision research projects and production applications in Facebook.

2.3 Document Dataset

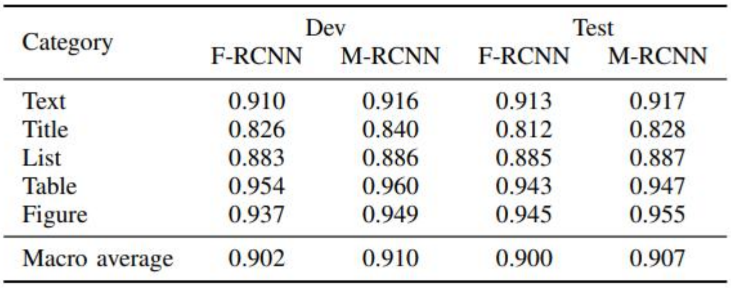
There are three focus on finding dataset for this project:

1. Should include majority of layout categories
2. Should include decent amount of training images
3. Must include accurate label data

2.3.1 PubLayNet

The PubLayNet dataset is designed for scientific papers and articles and includes 5 layout categories: text, list, figure, table, and title. It contains 335,703 training images, 11,245 validation images, and 11,405 test images. The dataset has become a popular choice for pre-training models in document layout detection, such as LayoutLMv3 and Layout Parser.

Table 1: MAP @ IOU [0.50:0.95] of the F-RCNN and M-RCNN models

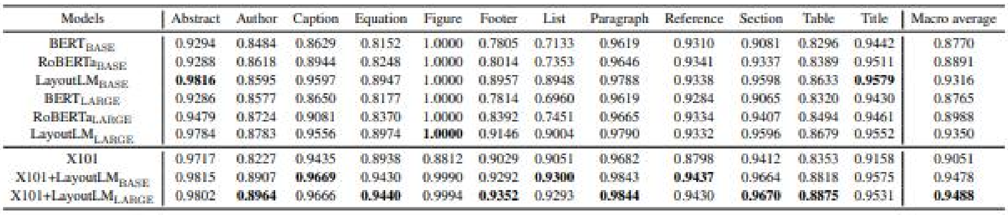


Category	Dev		Test	
	F-RCNN	M-RCNN	F-RCNN	M-RCNN
Text	0.910	0.916	0.913	0.917
Title	0.826	0.840	0.812	0.828
List	0.883	0.886	0.885	0.887
Table	0.954	0.960	0.943	0.947
Figure	0.937	0.949	0.945	0.955
Macro average	0.902	0.910	0.900	0.907

2.3.2 DocBank

The DocBank dataset contains 500,000 images that are categorized into 12 layout categories, including Abstract, Author, Caption, Equation, Figure, Footer, List, Paragraph, Reference, Section, Table, and Title. The dataset is focused on scientific papers and supports both text-based and image-based models. It is also expanding with automatic annotation methods. The availability of a large number of images and categories in DocBank makes it a valuable resource for research in document layout detection and related fields.

Table 2: The performance of BERT, RoBERTa, LayoutLM and Faster R-CNN on the DocBank test set



Models	Abstract	Author	Caption	Equation	Figure	Footer	List	Paragraph	Reference	Section	Table	Title	Macro average
$\text{BERT}_{\text{BASE}}$	0.9294	0.8484	0.8629	0.8152	1.0000	0.7805	0.7133	0.9619	0.9310	0.9081	0.8296	0.9442	0.8770
$\text{RoBERTa}_{\text{BASE}}$	0.9288	0.8618	0.8944	0.8248	1.0000	0.8014	0.7353	0.9646	0.9341	0.9337	0.8389	0.9511	0.8891
$\text{LayoutLM}_{\text{BASE}}$	**0.9816**	0.8595	0.9597	0.8947	1.0000	0.8957	0.8948	0.9788	0.9338	0.9598	0.8633	**0.9579**	0.9316
$\text{BERT}_{\text{LARGE}}$	0.9286	0.8577	0.8650	0.8177	1.0000	0.7814	0.6960	0.9619	0.9284	0.9065	0.8320	0.9430	0.8765
$\text{RoBERTa}_{\text{LARGE}}$	0.9479	0.8724	0.9081	0.8370	1.0000	0.8392	0.7451	0.9665	0.9334	0.9407	0.8494	0.9461	0.8988
$\text{LayoutLM}_{\text{LARGE}}$	0.9784	0.8783	0.9556	0.8974	**1.0000**	0.9146	0.9004	0.9790	0.9332	0.9596	0.8679	0.9552	0.9350
X101	0.9717	0.8227	0.9435	0.8938	0.8812	0.9029	0.9051	0.9682	0.8798	0.9412	0.8353	0.9158	0.9051
$\text{X101+LayoutLM}_{\text{BASE}}$	0.9815	0.8907	**0.9669**	0.9430	0.9990	0.9292	**0.9300**	0.9843	**0.9437**	0.9664	0.8818	0.9575	0.9478
$\text{X101+LayoutLM}_{\text{LARGE}}$	0.9802	**0.8964**	0.9666	**0.9440**	0.9994	**0.9352**	0.9293	**0.9844**	0.9430	**0.9670**	**0.8875**	0.9531	**0.9488**

After reviewing the labeled data in DocBank, it was found that the dataset does not provide segmentation coordinates, indicating that the Mask R-CNN approach is not applicable to this dataset. This means that the instance segmentation task, which requires identifying the precise location and boundaries of objects within an image, cannot be performed with Mask R-CNN on this dataset. However, other methods such as object detection or classification may still be applicable depending on the specific task and goals of the project.

2.3.3 DocLayNet

DocLayNet is a dataset used for document layout analysis, containing 80,863 human-annotated images from diverse data sources, providing a wide range of layout formats. The dataset includes 11 layout categories, which are caption, footnote, formula, list-item, page-footer, page-header, picture, section-header, table, text, and title. The dataset is intended to be used for training and evaluating document layout analysis algorithms and models.

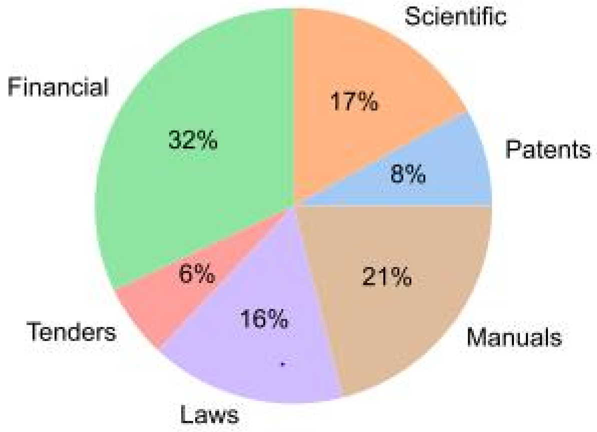


Figure 2: Distribution of DocLayNet pages across document categories.

DocLayNet dataset has specific guidelines for annotation to ensure consistency and meaningfulness in the dataset. These guidelines include:

1. List items are considered as individual object instances, unlike PubLayNet and DocBank where grouped list paragraphs are considered as one list element.

2. A list-item is a paragraph with hanging indentation, and even single-line elements can qualify as list-items if the neighbor elements expose hanging indentation. Bullet or enumeration symbols are not required.

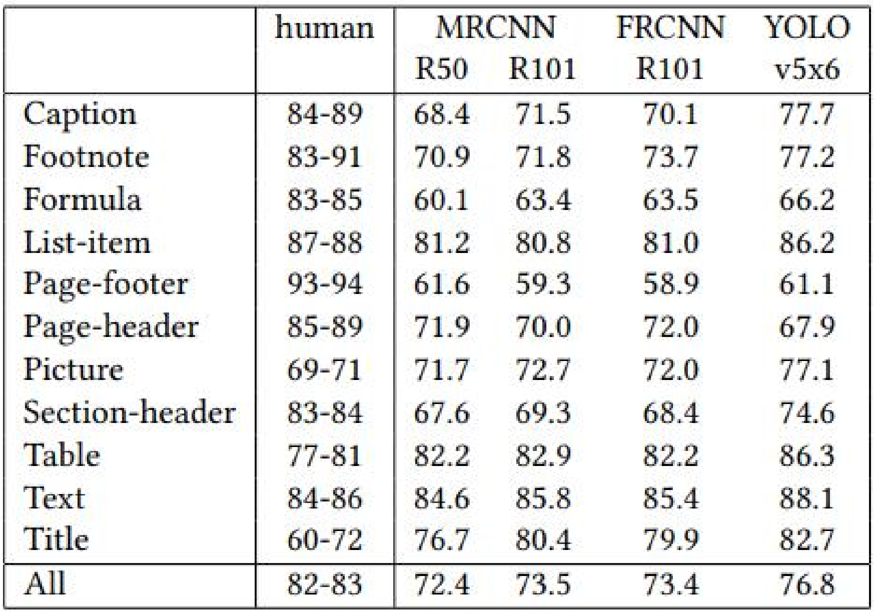
3. All captions are mapped to exactly one table or figure.

4. Figures that are connected will be considered as one whole figure.

5. Numbers are included in the formula object.

6. Italic and bold text alone in a new line are considered as section headers, but not at the beginning of the paragraph.

Table 3: Prediction performance (mAP@0.5-0.95) of object detection networks on DocLayNet test set



	human	MRCNN		FRCNN	YOLO
		R50	R101	R101	v5x6
Caption	84-89	68.4	71.5	70.1	77.7
Footnote	83-91	70.9	71.8	73.7	77.2
Formula	83-85	60.1	63.4	63.5	66.2
List-item	87-88	81.2	80.8	81.0	86.2
Page-footer	93-94	61.6	59.3	58.9	61.1
Page-header	85-89	71.9	70.0	72.0	67.9
Picture	69-71	71.7	72.7	72.0	77.1
Section-header	83-84	67.6	69.3	68.4	74.6
Table	77-81	82.2	82.9	82.2	86.3
Text	84-86	84.6	85.8	85.4	88.1
Title	60-72	76.7	80.4	79.9	82.7
All	82-83	72.4	73.5	73.4	76.8

2.3.4 WarpDoc

The purpose of the WarpDoc dataset is to train and evaluate machine learning algorithms for document image analysis tasks, including text detection and recognition. The dataset includes photo-taking document images captured from different angles, but it does not come with annotations provided by the source.

3. Implementation

This chapter will cover the design concept, resource planning and training implementation.

3.1 Design concept

To improve the accuracy of document layout prediction for complex and varied document images, this project proposes a novel two-step approach. The first step involves detecting the bounding box of the whole document, which is then cropped from the original image. The second step applies the prediction using a generic document layout model on the cropped bounding box.

In cases where no document is detected, the original image will be used for layout detection. This approach provides a more efficient way to predict document layouts for images with complex backgrounds or other non-document elements.

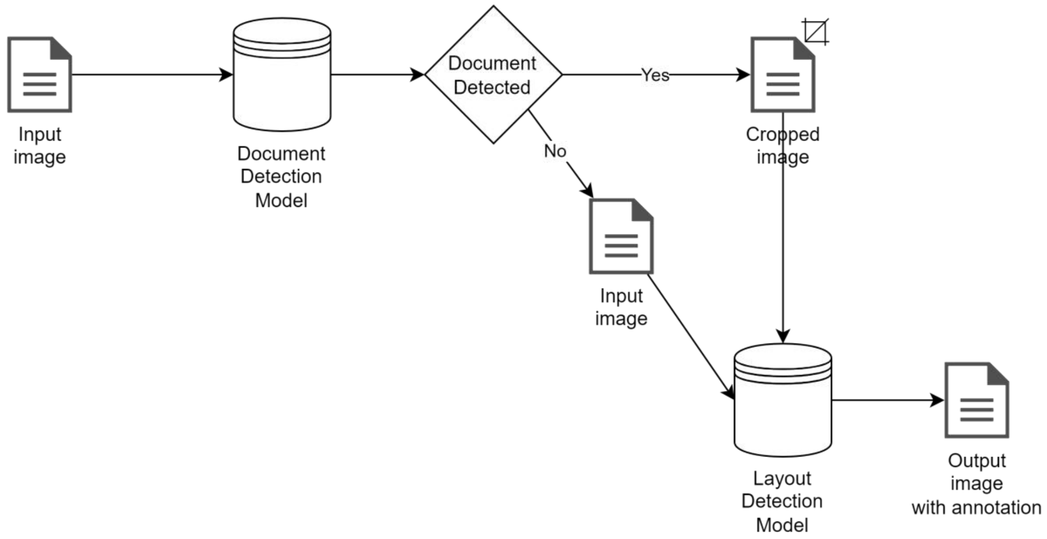


Figure 3: Document layout prediction work flow

3.2 Resource planning

3.2.1 Hardware

Device	Windows PC
RAM	16GB
CPU	Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz
GPU	RTX 2060 @ 6GB RAM

3.2.2 Software tools

Based on the research finding, Detectron2 library will be used in the model training as it is considered a relative new library and many base-line models are provided by the Facebook official. Installation prerequisite as follow:

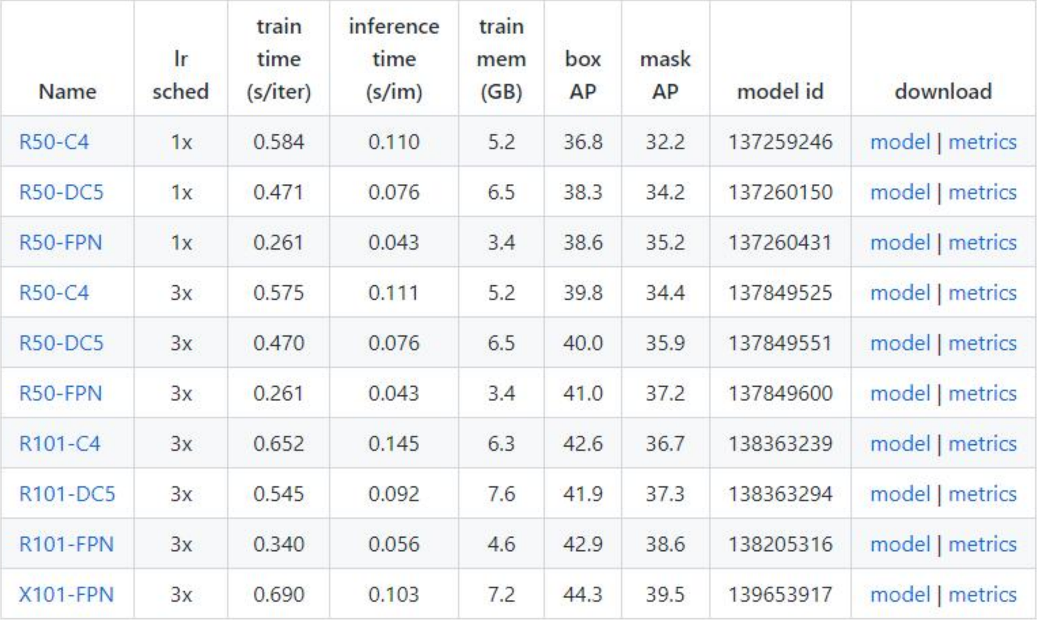
Python 3.7 or higher
PyTorch 1.8 or higher and torchvision match the PyTorch version
CUDA toolkit match PyTorch version

3.2.3 Choice of model architecture

To achieve precise detection of document layouts, this project will utilize the Mask R-CNN method for model training. This approach is capable of predicting not only the bounding box but also the mask area of the component. This is helpful when the input image is in photo taken form and it maybe distorted or rotated as the bounding box area may include redundant information.

To facilitate the training process, the Detectron2 library will be used, which offers various baseline models for Mask R-CNN architecture. Specifically, the R101-FPN baseline model will be utilized as it provides a good balance between training time and precision. This model has been pre-trained on a large-scale dataset and can detect objects at different scales, which is crucial for accurately detecting the various components of a document layout. Additionally, the use of this baseline model will reduce the need for extensive training and tuning, enabling the project to focus more on optimizing the model for document layout detection. Overall, the combination of the Mask R-CNN approach and the R101-FPN baseline model is expected to result in a highly accurate and efficient model for document layout detection.

Table 4: Base Lines model provided by Detectron2 with Mask R-CNN



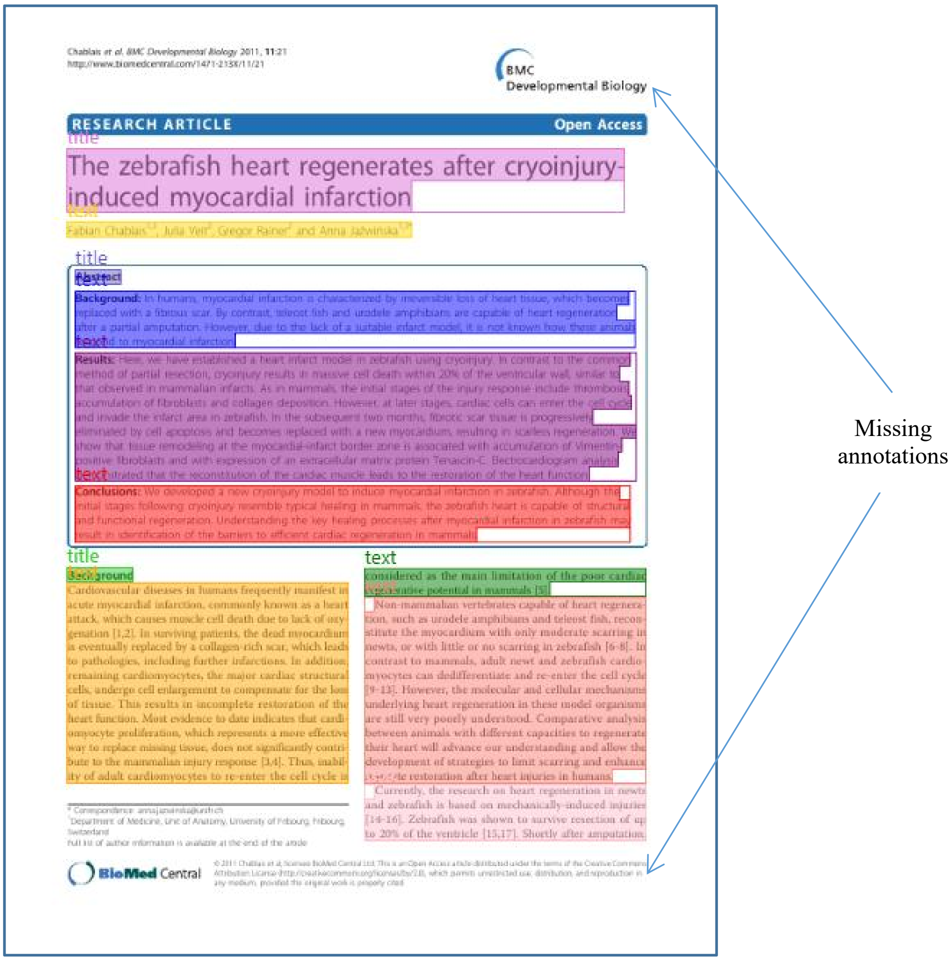
Name	lr sched	train time (s/iter)	inference time (s/im)	train mem (GB)	box AP	mask AP	model id	download
R50-C4	1x	0.584	0.110	5.2	36.8	32.2	137259246	model | metrics
R50-DC5	1x	0.471	0.076	6.5	38.3	34.2	137260150	model | metrics
R50-FPN	1x	0.261	0.043	3.4	38.6	35.2	137260431	model | metrics
R50-C4	3x	0.575	0.111	5.2	39.8	34.4	137849525	model | metrics
R50-DC5	3x	0.470	0.076	6.5	40.0	35.9	137849551	model | metrics
R50-FPN	3x	0.261	0.043	3.4	41.0	37.2	137849600	model | metrics
R101-C4	3x	0.652	0.145	6.3	42.6	36.7	138363239	model | metrics
R101-DC5	3x	0.545	0.092	7.6	41.9	37.3	138363294	model | metrics
R101-FPN	3x	0.340	0.056	4.6	42.9	38.6	138205316	model | metrics
X101-FPN	3x	0.690	0.103	7.2	44.3	39.5	139653917	model | metrics

3.2.4 Choice of dataset

Layout detection Model

The project has opted to use the DocLayNet dataset for training and testing data due to its reliable annotations, diverse document layouts, and reasonable number of images. This dataset's variety will enable the model to learn and detect various document layouts accurately. In contrast, the PubLayNet and DocBank models have limited effectiveness in detecting non-standard formats of research papers or articles.

Furthermore, the dataset annotation in PubLayNet may not encompass all of the details presented in an image. An example of this is presented below.



Chablais *et al. BMC Developmental Biology* 2011, **11**:21
http://www.biomedcentral.com/1471-213X/11/21

BMC Developmental Biology

RESEARCH ARTICLE Open Access

The zebrafish heart regenerates after cryoinjury-induced myocardial infarction

Fabian Chablais[1,2], Julia Veit[2], Gregor Rainer[2] and Anna Jaźwińska[1,2*]

Abstract

Background: In humans, myocardial infarction is characterized by irreversible loss of heart tissue, which becomes replaced with a fibrous scar. By contrast, teleost fish and urodele amphibians are capable of heart regeneration after a partial amputation. However, due to the lack of a suitable infarct model, it is not known how these animals respond to myocardial infarction.

Results: Here, we have established a heart infarct model in zebrafish using cryoinjury. In contrast to the common method of partial resection, cryoinjury results in massive cell death within 20% of the ventricular wall, similar to that observed in mammalian infarcts. As in mammals, the initial stages of the injury response include thrombosis, accumulation of fibroblasts and collagen deposition. However, at later stages, cardiac cells can enter the cell cycle and invade the infarct area in zebrafish. In the subsequent two months, fibrotic scar tissue is progressively eliminated by cell apoptosis and becomes replaced with a new myocardium, resulting in scarless regeneration. We show that tissue remodeling at the myocardial-infarct border zone is associated with accumulation of Vimentin-positive fibroblasts and with expression of an extracellular matrix protein Tenascin-C. Electrocardiogram analysis demonstrated that the reconstitution of the cardiac muscle leads to the restoration of the heart function.

Conclusions: We developed a new cryoinjury model to induce myocardial infarction in zebrafish. Although the initial stages following cryoinjury resemble typical healing in mammals, the zebrafish heart is capable of structural and functional regeneration. Understanding the key healing processes after myocardial infarction in zebrafish may result in identification of the barriers to efficient cardiac regeneration in mammals.

Background

Cardiovascular diseases in humans frequently manifest in acute myocardial infarction, commonly known as a heart attack, which causes muscle cell death due to lack of oxygenation [1,2]. In surviving patients, the dead myocardium is eventually replaced by a collagen-rich scar, which leads to pathologies, including further infarctions. In addition, remaining cardiomyocytes, the major cardiac structural cells, undergo cell enlargement to compensate for the loss of tissue. This results in incomplete restoration of the heart function. Most evidence to date indicates that cardiomyocyte proliferation, which represents a more effective way to replace missing tissue, does not significantly contribute to the mammalian injury response [3,4]. Thus, inability of adult cardiomyocytes to re-enter the cell cycle is considered as the main limitation of the poor cardiac regenerative potential in mammals [5].

Non-mammalian vertebrates capable of heart regeneration, such as urodele amphibians and teleost fish, reconstitute the myocardium with only moderate scarring in newts, or with little or no scarring in zebrafish [6-8]. In contrast to mammals, adult newt and zebrafish cardiomyocytes can dedifferentiate and re-enter the cell cycle [9-13]. However, the molecular and cellular mechanisms underlying heart regeneration in these model organisms are still very poorly understood. Comparative analysis between animals with different capacities to regenerate their heart will advance our understanding and allow the development of strategies to limit scarring and enhance cardiomyocyte restoration after heart injuries in humans.

Currently, the research on heart regeneration in newts and zebrafish is based on mechanically-induced injuries [14-16]. Zebrafish was shown to survive resection of up to 20% of the ventricle [15,17]. Shortly after amputation,

* Correspondence: anna.jazwinska@unifr.ch
[1]Department of Medicine, Unit of Anatomy, University of Fribourg, Fribourg, Switzerland
Full list of author information is available at the end of the article

BioMed Central

© 2011 Chablais et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Figure 4: PubLayNet dataset sample

The absence of labels on the text and logo situated at the top and bottom of the document could result in inaccurate predictions.

The annotations in the DocBank dataset may not be entirely trustworthy, as some of the labels do not correspond accurately to the actual content. Samples of this are presented below.

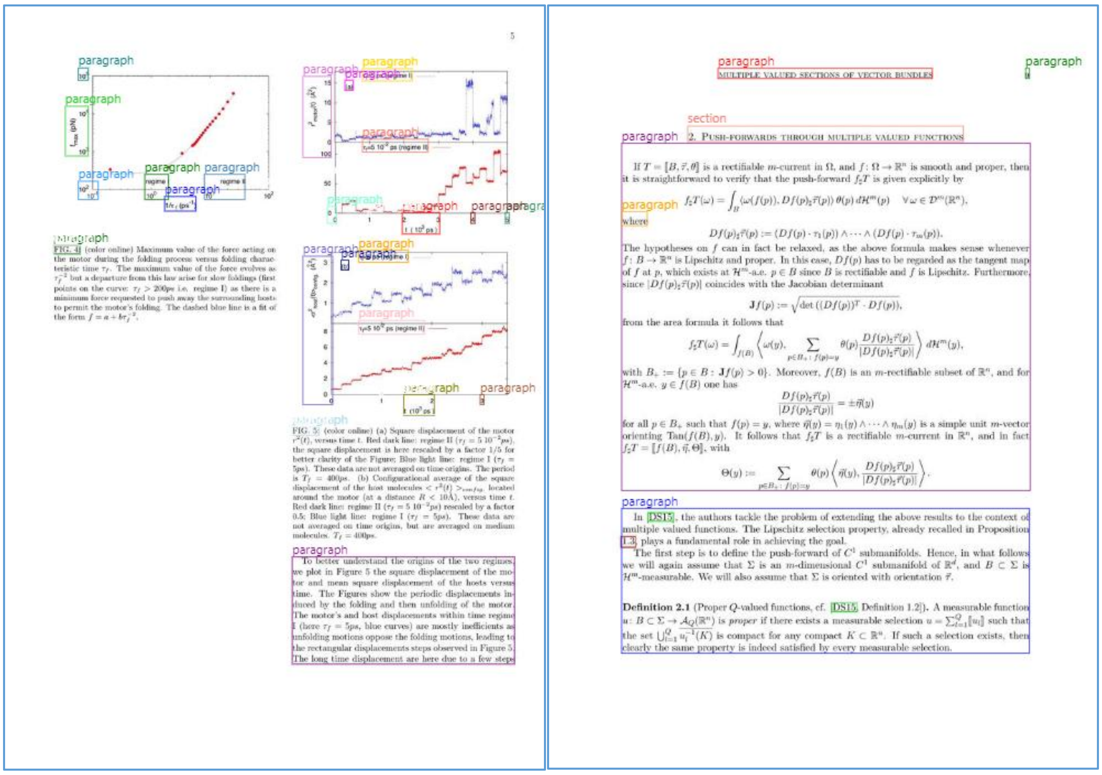


FIG. 4: (color online) Maximum value of the force acting on the motor during the folding process versus folding characteristic time τ_f. The maximum value of the force evolves as τ_f^{-2} but a departure from this law arise for slow foldings (first points on the curve: $\tau_f > 200$ps i.e. regime I) as there is a minimum force requested to push away the surrounding hosts to permit the motor's folding. The dashed blue line is a fit of the form $f = a + b\tau_f^{-2}$.

FIG. 5: (color online) (a) Square displacement of the motor $r^2(t)$, versus time t. Red dark line: regime II ($\tau_f = 5\ 10^{-2}ps$), the square displacement is here rescaled by a factor 1/5 for better clarity of the Figure; Blue light line: regime I ($\tau_f =$ 5ps). These data are not averaged on time origins. The period is T_f = 400ps. (b) Configurational average of the square displacement of the host molecules $< r^2(t) >_{conf/ag}$, located around the motor (at a distance $R <$ 10Å), versus time t. Red dark line: regime II ($\tau_f = 5\ 10^{-2}$ps) rescaled by a factor 0.5; Blue light line: regime I (τ_f = 5ps). These data are not averaged on time origins, but are averaged on medium molecules. T_f = 400ps.

To better understand the origins of the two regimes, we plot in Figure 5 the square displacement of the motor and mean square displacement of the hosts versus time. The Figures show the periodic displacements induced by the folding and then unfolding of the motor. The motor's and host displacements within time regime I (here τ_f = 5ps, blue curves) are mostly inefficients as unfolding motions oppose the folding motions, leading to the rectangular displacements steps observed in Figure 5. The long time displacement are here due to a few steps

2. PUSH-FORWARDS THROUGH MULTIPLE VALUED FUNCTIONS

If $T = [\![B, \vec{\tau}, \theta]\!]$ is a rectifiable m-current in Ω, and $f : \Omega \to \mathbb{R}^n$ is smooth and proper, then it is straightforward to verify that the push-forward $f_\sharp T$ is given explicitly by

$$f_\sharp T(\omega) = \int_B \langle \omega(f(p)), Df(p)_\sharp \vec{\tau}(p) \rangle\, \theta(p)\, d\mathcal{H}^m(p) \quad \forall \omega \in \mathcal{D}^m(\mathbb{R}^n),$$

where

$$Df(p)_\sharp \vec{\tau}(p) := (Df(p) \cdot \tau_1(p)) \wedge \cdots \wedge (Df(p) \cdot \tau_m(p)).$$

The hypotheses on f can in fact be relaxed, as the above formula makes sense whenever $f : B \to \mathbb{R}^n$ is Lipschitz and proper. In this case, $Df(p)$ has to be regarded as the tangent map of f at p, which exists at $\mathcal{H}^m$-a.e. $p \in B$ since B is rectifiable and f is Lipschitz. Furthermore, since $|Df(p)_\sharp \vec{\tau}(p)|$ coincides with the Jacobian determinant

$$\mathbf{J}f(p) := \sqrt{\det\left((Df(p))^T \cdot Df(p)\right)},$$

from the area formula it follows that

$$f_\sharp T(\omega) = \int_{f(B)} \left\langle \omega(y), \sum_{p \in B_+ : f(p) = y} \theta(p) \frac{Df(p)_\sharp \vec{\tau}(p)}{|Df(p)_\sharp \vec{\tau}(p)|} \right\rangle d\mathcal{H}^m(y),$$

with $B_+ := \{p \in B : \mathbf{J}f(p) > 0\}$. Moreover, $f(B)$ is an m-rectifiable subset of $\mathbb{R}^n$, and for $\mathcal{H}^m$-a.e. $y \in f(B)$ one has

$$\frac{Df(p)_\sharp \vec{\tau}(p)}{|Df(p)_\sharp \vec{\tau}(p)|} = \pm \vec{\eta}(y)$$

for all $p \in B_+$ such that $f(p) = y$, where $\vec{\eta}(y) = \eta_1(y) \wedge \cdots \wedge \eta_m(y)$ is a simple unit m-vector orienting $\mathrm{Tan}(f(B), y)$. It follows that $f_\sharp T$ is a rectifiable m-current in $\mathbb{R}^n$, and in fact $f_\sharp T = [\![f(B), \vec{\eta}, \Theta]\!]$, with

$$\Theta(y) := \sum_{p \in B_+ : f(p) = y} \theta(p) \left\langle \vec{\eta}(y), \frac{Df(p)_\sharp \vec{\tau}(p)}{|Df(p)_\sharp \vec{\tau}(p)|} \right\rangle.$$

In [DS15], the authors tackle the problem of extending the above results to the context of multiple valued functions. The Lipschitz selection property, already recalled in Proposition 1.3, plays a fundamental role in achieving the goal.

The first step is to define the push-forward of C^1 submanifolds. Hence, in what follows we will again assume that Σ is an m-dimensional C^1 submanifold of $\mathbb{R}^d$, and $B \subset \Sigma$ is $\mathcal{H}^m$-measurable. We will also assume that Σ is oriented with orientation $\vec{\tau}$.

Definition 2.1 (Proper Q-valued functions, cf. [DS15, Definition 1.2]). A measurable function $u : B \subset \Sigma \to \mathcal{A}_Q(\mathbb{R}^n)$ is *proper* if there exists a measurable selection $u = \sum_{i=1}^{Q} [\![u_i]\!]$ such that the set $\overline{\bigcup_{i=1}^{Q} u_i^{-1}(K)}$ is compact for any compact $K \subset \mathbb{R}^n$. If such a selection exists, then clearly the same property is indeed satisfied by every measurable selection.

Figure 5: DocBank dataset sample

The annotations are significantly inaccurate, rendering them is not suitable for training purposes. Upon reviewing a selection of samples at random, it became apparent that incorrect annotations are a frequent occurrence. Therefore, the DocBank dataset will not be used as a training dataset.

As a result, the DocLayNet dataset was deemed preferable due to its inclusion of a wide variety of document types, including manuals, financial reports, legal documents, and more. This diverse range of document layouts can aid in training the model to recognize a broader spectrum of layouts accurately. Utilizing the DocLayNet dataset will enable the model to perform more effectively in real-world scenarios, where document layouts and formats can vary considerably.

Document detection Model

Manual labeling photo taking images from WarpDoc dataset.

3.3 Training implementation

3.3.1 Data pre-processing

Layout detection Model

To enhance the model's ability to detect document layouts in photo-taking form, the original dataset, which consists only of digital documents, may not be sufficient. However, manually labeling and photographing at least 10,000 different documents is impractical. Instead, the training data can be augmented by transforming various samples randomly in each training iteration to simulate different textures and layouts.

To achieve this, images can be randomly transformed by flipping 90 degrees, rotating between -15 and 15 degrees, and applying saturation, brightness, contrast, or lighting adjustments. Each transformation will have an equal probability of 0.2. This approach will help to increase the diversity and variability of the training dataset, allowing the model to learn and generalize better to new and unseen document layouts in photo-taking form. Further experimentation can be done to determine the optimal transformation combinations and probabilities for the dataset augmentation.

Below will be the illustration on the image transformation:

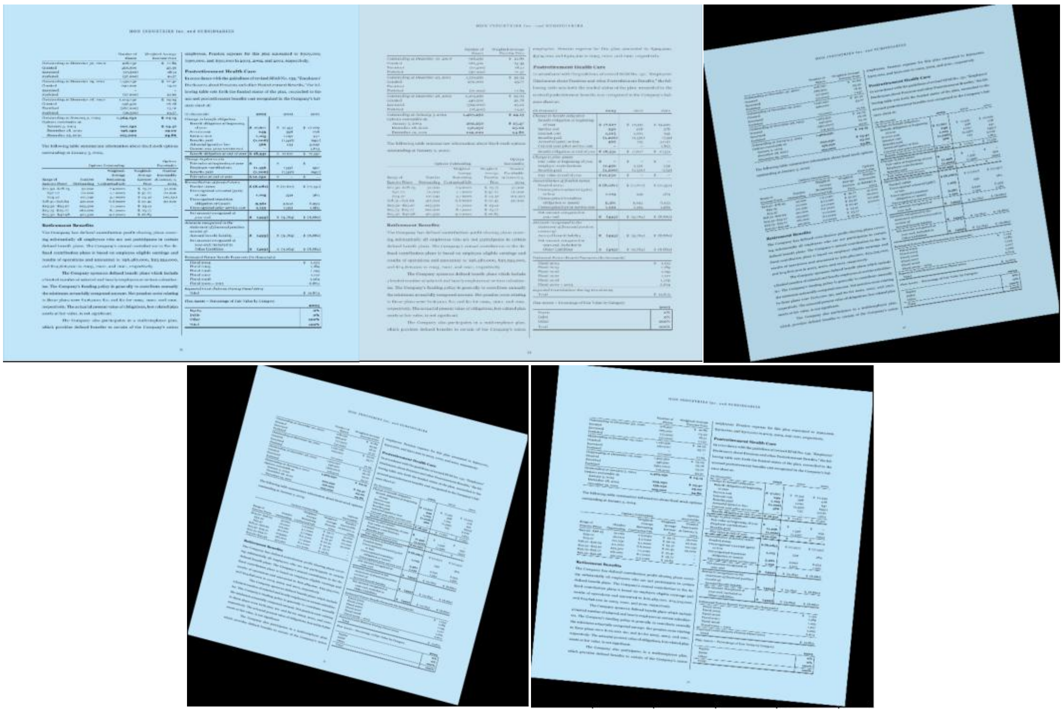


Figure 6: Image augmentation samples for DocLayNet dataset

Document detection Model

To overcome hardware limitations, image transformations during training will cause out-of-memory errors. Therefore, the WarpDoc dataset is transformed beforehand.

This dataset consists of photo-taking images, so only rotation transformations are applied. The following is an illustration of the image transformation:

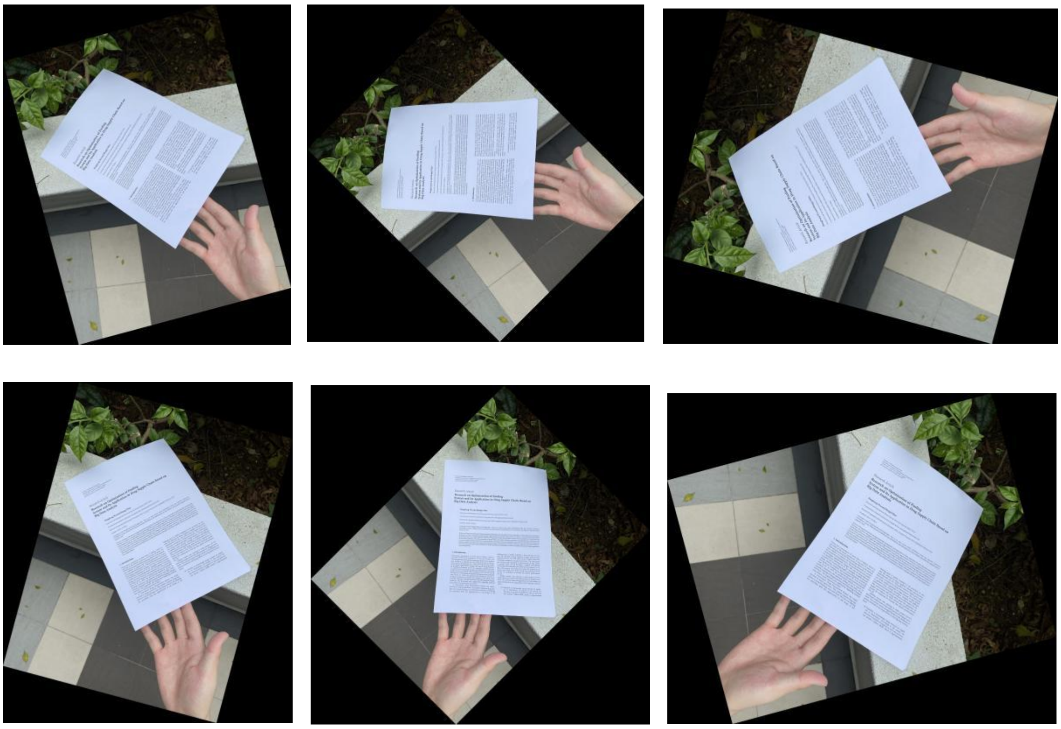


Figure 7: Image augmentation samples for WarpDoc dataset

3.3.2 Model Training

Layout detection Model

Due to the limited hardware resource, the model was break down to batches and each batch was evaluate using the test set and sample of photo taking images.

The training set contains 69375 images and below is total number of item in 11 categories.

Table 4: DocLayNet train set category summary

Category	Item count
Caption	19218
List-item	161818
Picture	39667
Text	431251

Footnote	5619
Page-footer	61313
Section-header	118590
Title	4437
Formula	21167
Page-header	47973
Table	30070
Total	941123

All the training configuration is set based on guideline with adjustment on hardware constraint and custom implementation.

Table 5: Training parameter for Layout Model

Parameter	Value	Remark
Base line model	mask_rcnn_R_101_FPN_3x	
Number of Class	11	The total number categories will able to detect in the model
Image per Iteration	1	Due to hardware constraint, each iteration will only take 1 input image
Iteration	Per training >= 60000 Final iteration = 1460000	Due to hardware constraint, the training process will be break down to many batches. In order to fit as many training data in each batch, 60000 iteration is set as only 1 image is able to fit in 1 iteration. Based on the evaluation data, the model will not improve significantly on detecting all categories after 140000. Thus, it is stopped in iteration 1460000

Learning Rate	0.00025	

Document detection Model

Table 6: Training parameter for Document Model

Parameter	Value	Remark
Base line model	mask_rcnn_R_101_FPN_3x	
Number of Class	1	Only 1 category which is ‘document’.
Image per Iteration	1	Due to hardware constraint, each iteration will only take 1 input image
Iteration	80000	First batch training is set to 80000 and the model will not improve significant over 50000. Thus, stop in first batch.
Learning Rate	0.00025	

4. Result summary

This chapter will discuss on the performance of the model and how the prediction process mentioned in Chapter 3.1 help to improve the accuracy on detecting layout from photo taking images.

4.1 Evaluation Metric

Intersection over Union (IoU) is measure the overlap of the predicted area and the ground truth area.

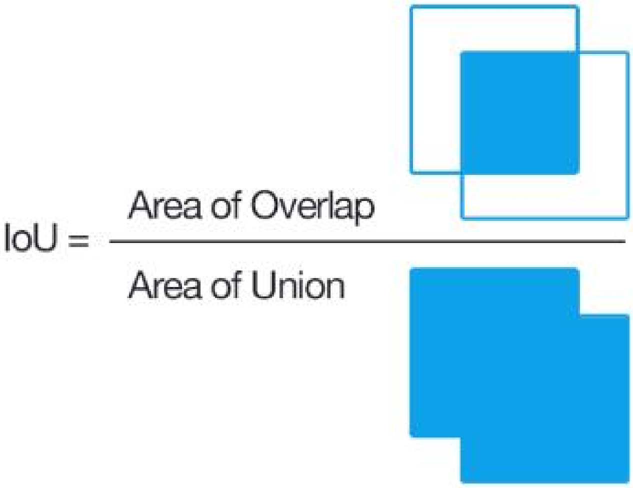


Figure 7: Intersection over Union

The evaluation result metric will use the IoU value to calculate the Average Precision (AP) for each category. The AP will use the 10 threshold IoU value: 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95 and average the output value.

4.2 Evaluation Result

Layout detection Model

The test set contains 6489 images and below is total number of item in 11 categories.

Table 7: DocLayNet test set category summary

Category	AP (over 10 threshold) bbox
Caption	1763
List-item	13320

Picture	2775
Text	49186
Footnote	312
Page-footer	5571
Section-header	15744
Title	299
Formula	1894
Page-header	6683
Table	2269
Total	99816

The Mask R-CNN model generates two types of prediction outputs: bounding boxes (bbox) and mask segmentation.

Table 8: Overall Average Precision

AP (over 10 threshold) bbox	AP50 bbox	AP75 bbox	AP (over 10 threshold) segmentation	AP50 segmentation	AP75 segmentation
63.598	86.445	71.208	64.655	86.420	72.469

Table 9: Average Precision on the 11 categories.

Category	AP (over 10 threshold) bbox	AP (over 10 threshold) segmentation
Caption	71.970	72.100
List-item	64.659	67.131
Picture	77.225	77.699
Text	77.603	79.288

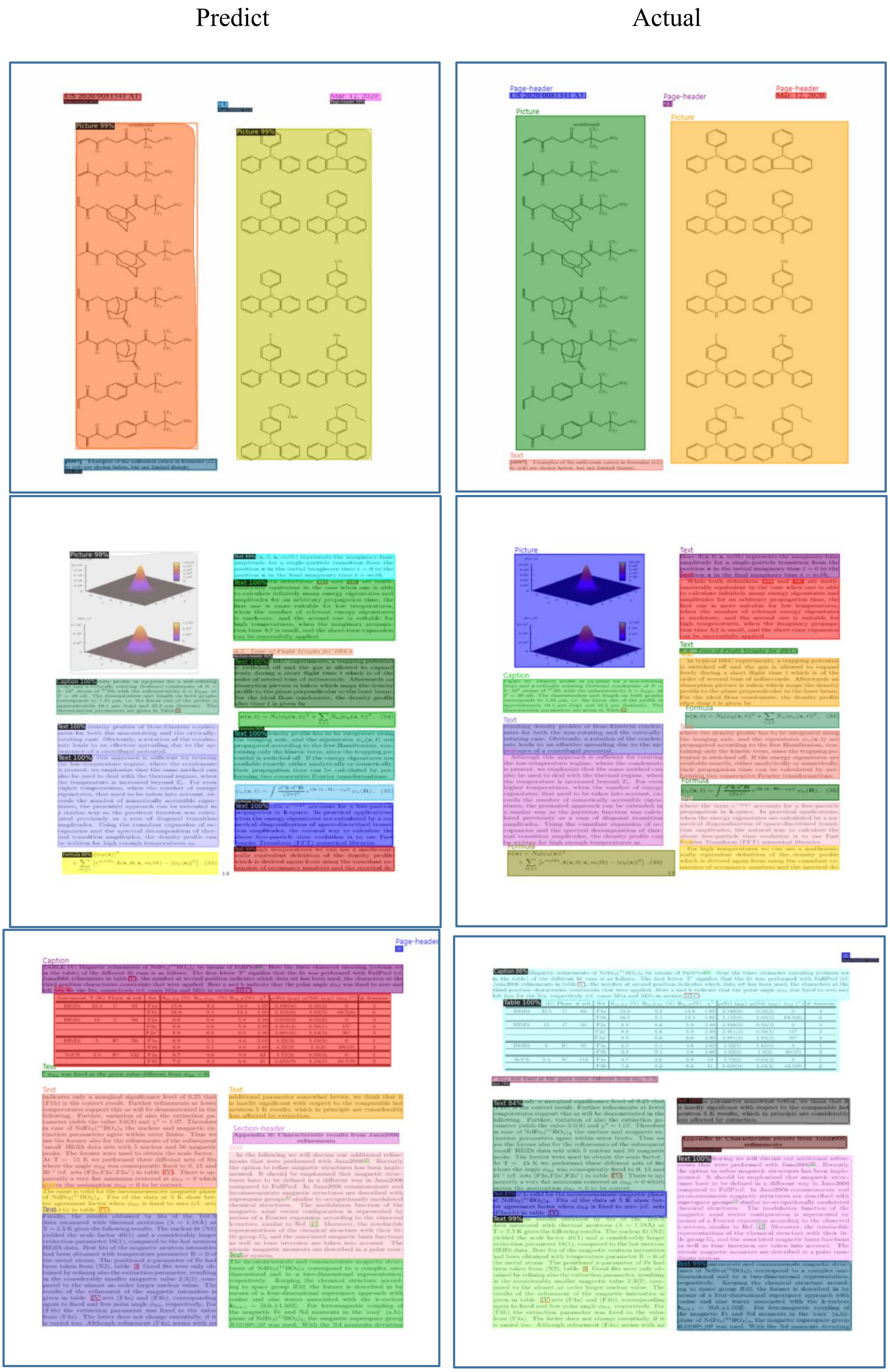
Footnote	55.656	55.761
Page-footer	48.761	50.034
Section-header	57.685	58.545
Title	53.664	55.513
Formula	52.739	53.937
Page-header	62.205	62.780
Table	77.415	78.413

4.3 Prediction experiment

4.3.1 Prediction on Digital document

DocLayNet

Below are some examples of the model's predictions on the digital document test set, which includes all 11 categories.


Predict
Actual

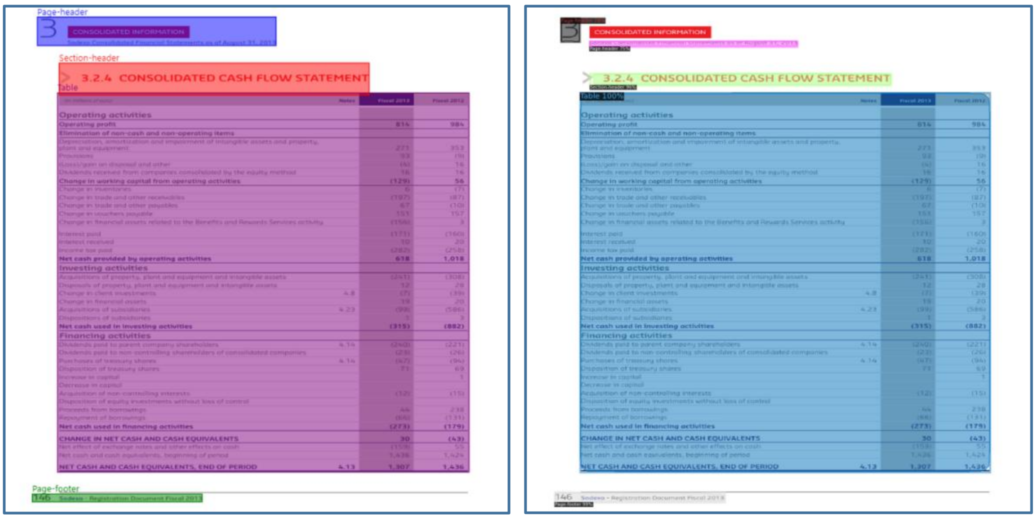
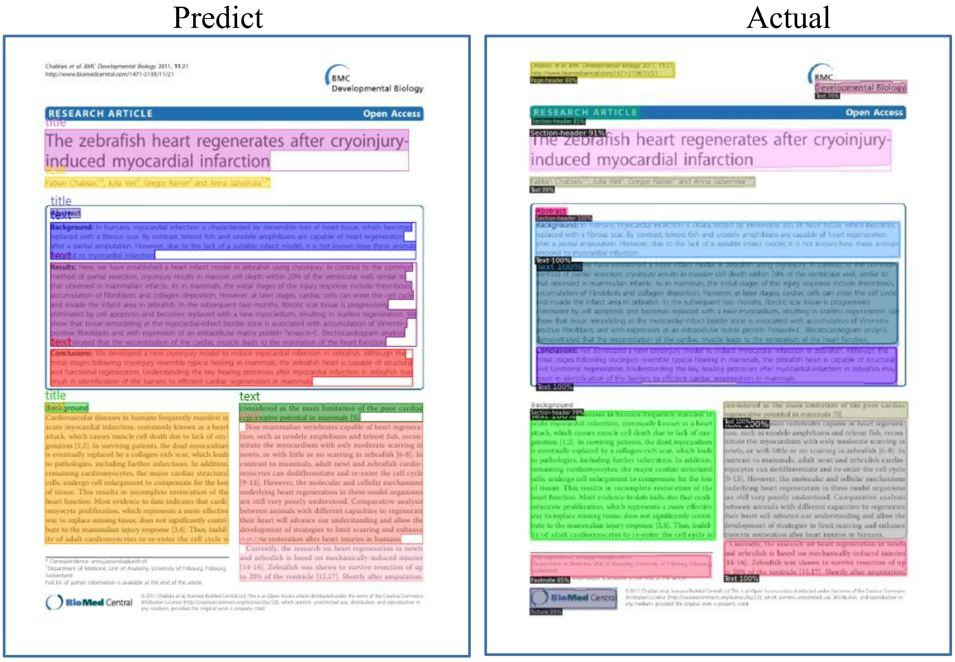
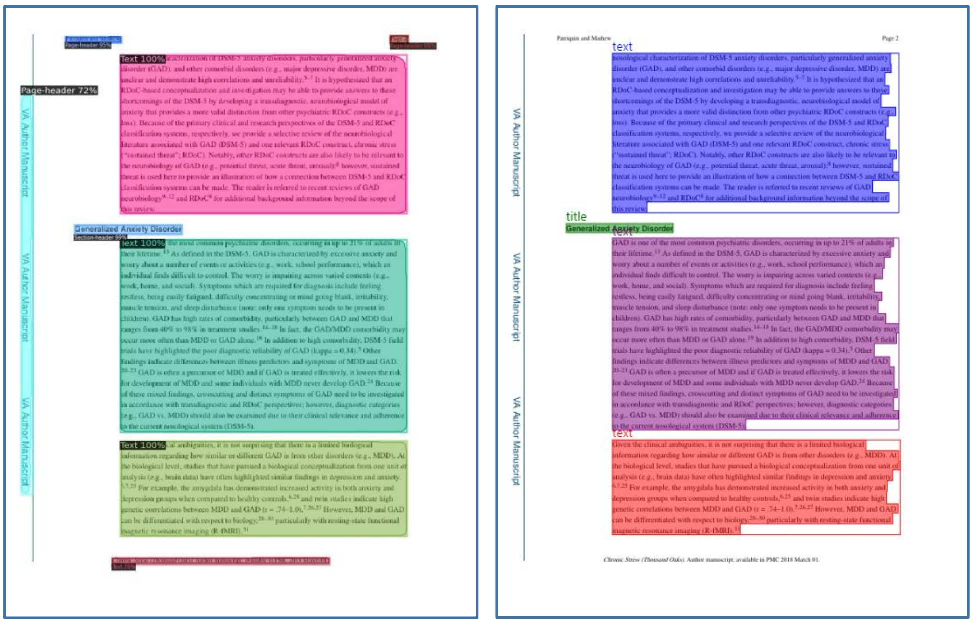


Figure 8: Prediction on DocLayNet test set

The model demonstrates strong performance in predicting layout from digital documents within the DocLayNet test set. In order to assess the model's ability to detect layout in a variety of digital documents, it was also tested on the DocBank and PubLayNet datasets. However, due to differences in the number of categories and annotation styles among these three datasets, it is not possible to calculate mAP. Instead, the results must be analyzed through observation. Sample predictions and actual annotations comparison are provided below.

PubLayNet





Patriquin and Mathew

Page 2

VA Author Manuscript

nosological characterization of DSM-5 anxiety disorders, particularly generalized anxiety disorder (GAD), and other comorbid disorders (e.g., major depressive disorder, MDD) are unclear and demonstrate high correlations and unreliability.[5–7] It is hypothesized that an RDoC-based conceptualization and investigation may be able to provide answers to these shortcomings of the DSM-5 by developing a transdiagnostic, neurobiological model of anxiety that provides a more valid distinction from other psychiatric RDoC constructs (e.g., loss). Because of the primary clinical and research perspectives of the DSM-5 and RDoC classification systems, respectively, we provide a selective review of the neurobiological literature associated with GAD (DSM-5) and one relevant RDoC construct, chronic stress ("sustained threat"; RDoC). Notably, other RDoC constructs are also likely to be relevant to the neurobiology of GAD (e.g., potential threat, acute threat, arousal);[8] however, sustained threat is used here to provide an illustration of how a connection between DSM-5 and RDoC classification systems can be made. The reader is referred to recent reviews of GAD neurobiology[9–12] and RDoC[8] for additional background information beyond the scope of this review.

Generalized Anxiety Disorder

GAD is one of the most common psychiatric disorders, occurring in up to 21% of adults in their lifetime.[13] As defined in the DSM-5, GAD is characterized by excessive anxiety and worry about a number of events or activities (e.g., work, school performance), which an individual finds difficult to control. The worry is impairing across varied contexts (e.g., work, home, and social). Symptoms which are required for diagnosis include feeling restless, being easily fatigued, difficulty concentrating or mind going blank, irritability, muscle tension, and sleep disturbance (note: only one symptom needs to be present in children). GAD has high rates of comorbidity, particularly between GAD and MDD that ranges from 40% to 98% in treatment studies.[14–18] In fact, the GAD/MDD comorbidity may occur more often than MDD or GAD alone.[19] In addition to high comorbidity, DSM-5 field trials have highlighted the poor diagnostic reliability of GAD (kappa = 0.34).[5] Other findings indicate differences between illness predictors and symptoms of MDD and GAD.[20–23] GAD is often a precursor of MDD and if GAD is treated effectively, it lowers the risk for development of MDD and some individuals with MDD never develop GAD.[24] Because of these mixed findings, crosscutting and distinct symptoms of GAD need to be investigated in accordance with transdiagnostic and RDoC perspectives; however, diagnostic categories (e.g., GAD vs. MDD) should also be examined due to their clinical relevance and adherence to the current nosological system (DSM-5).

Given the clinical ambiguities, it is not surprising that there is a limited biological information regarding how similar or different GAD is from other disorders (e.g., MDD). At the biological level, studies that have pursued a biological conceptualization from one unit of analysis (e.g., brain data) have often highlighted similar findings in depression and anxiety.[6,7,25] For example, the amygdala has demonstrated increased activity in both anxiety and depression groups when compared to healthy controls,[6,25] and twin studies indicate high genetic correlations between MDD and GAD (r = .74–1.0).[7,26,27] However, MDD and GAD can be differentiated with respect to biology,[28–30] particularly with resting-state functional magnetic resonance imaging (R-fMRI).[31]

Chronic Stress (Thousand Oaks). Author manuscript; available in PMC 2018 March 01.

Figure 9: Prediction on PubLayNet dataset

DocBank

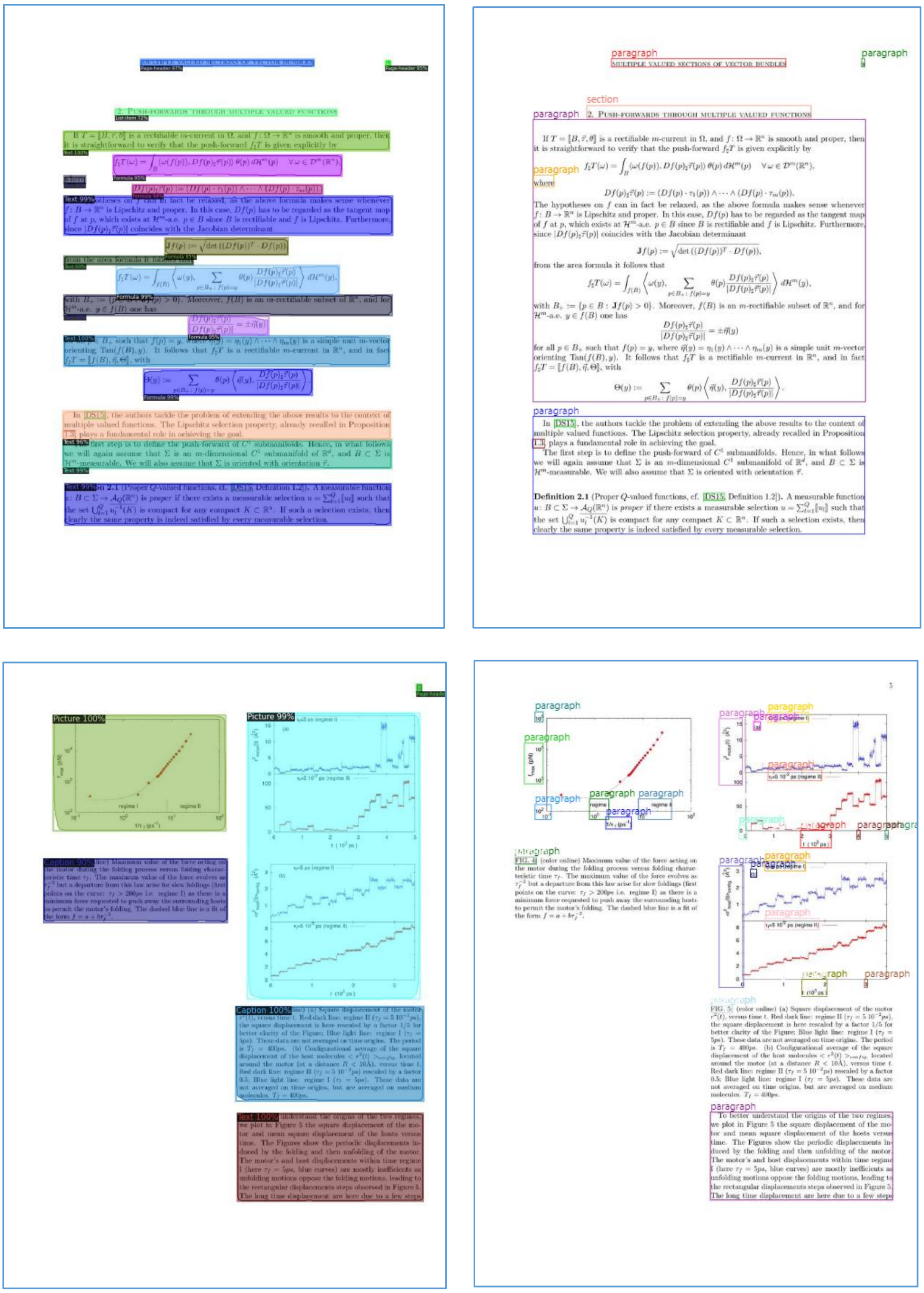


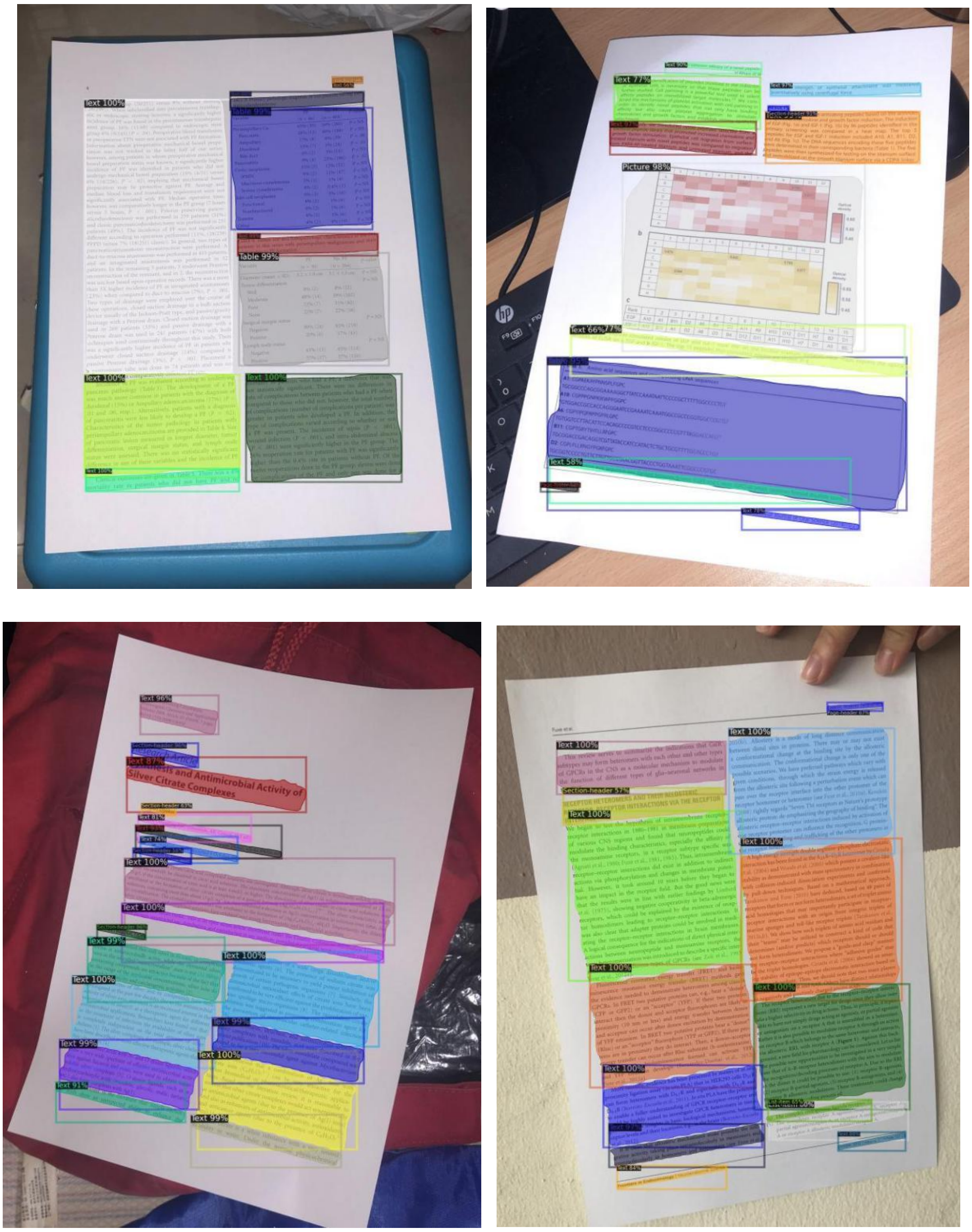
Figure 10: Prediction on DocBank dataset

The prediction results accurately identify the layout contents and contain more information than the annotations from PubLayNet. Furthermore, the predictions for the DocBank dataset offer more precise and detailed information compared to the actual annotations.

4.3.2 Prediction on Document in Photo Taking Images

As mentioned in Chapter 3.1, this project included a prediction model for detecting the entire document and then perform the layout detection.

Layout Detection Samples



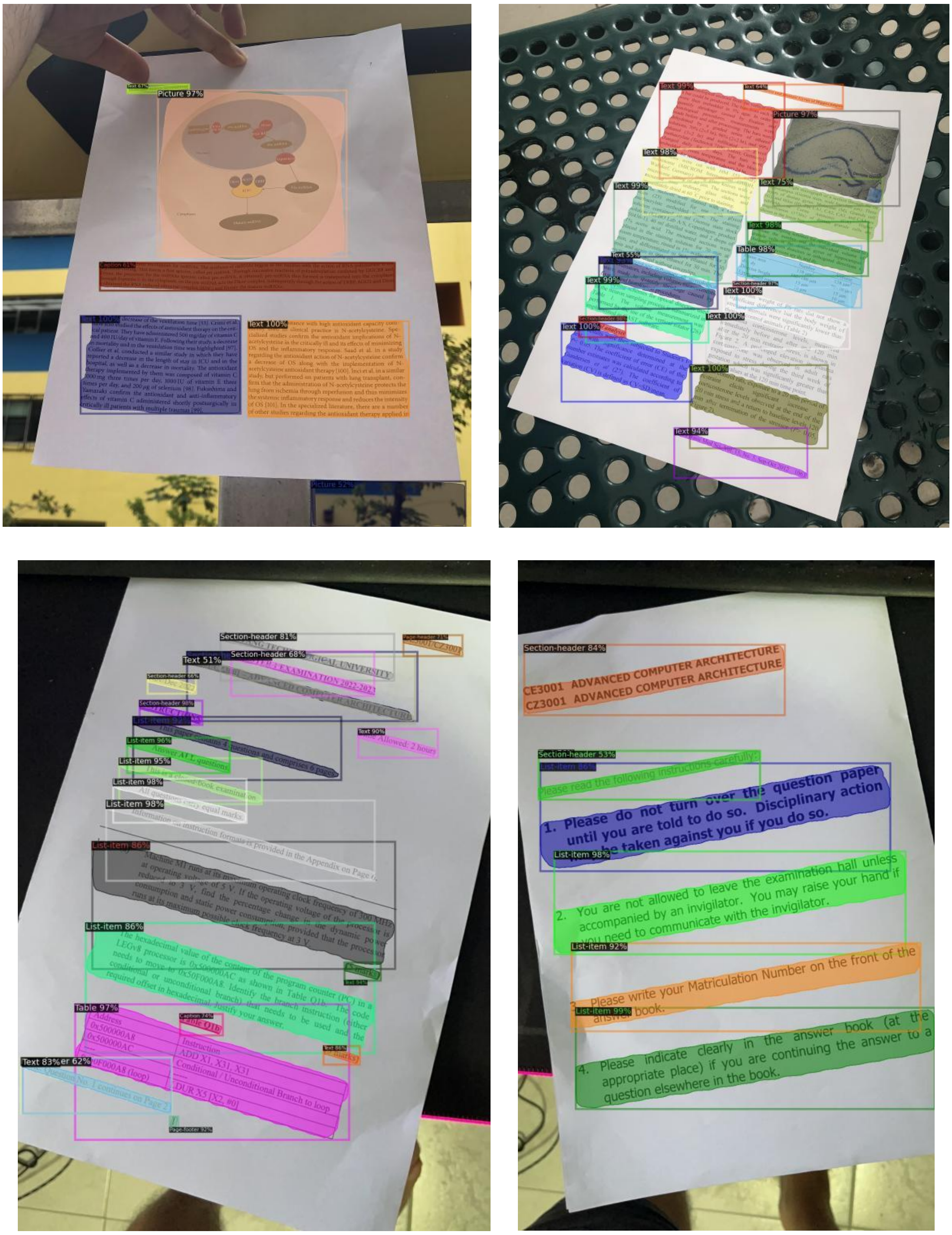
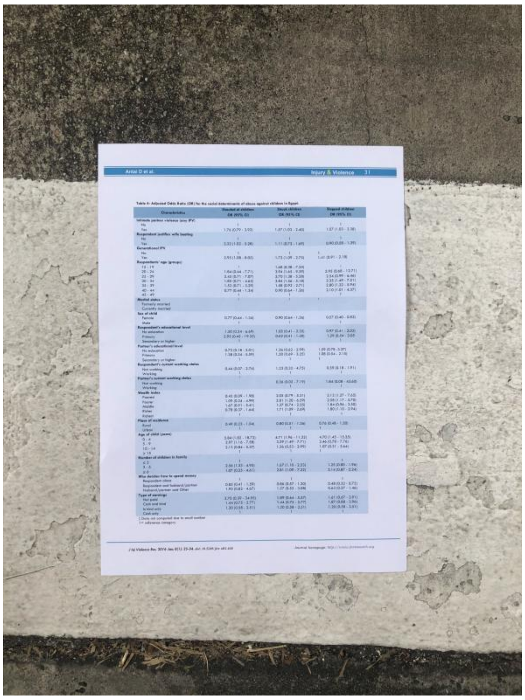


Figure 11: Prediction on DocBank dataset

The below is an example show the detection of the document and the improvements when using the layout model to predict on the cropped document image.

Original Image



Detection for Document

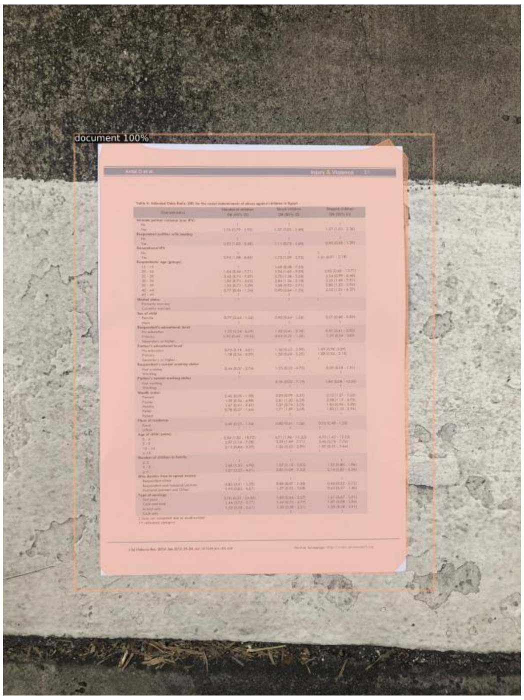
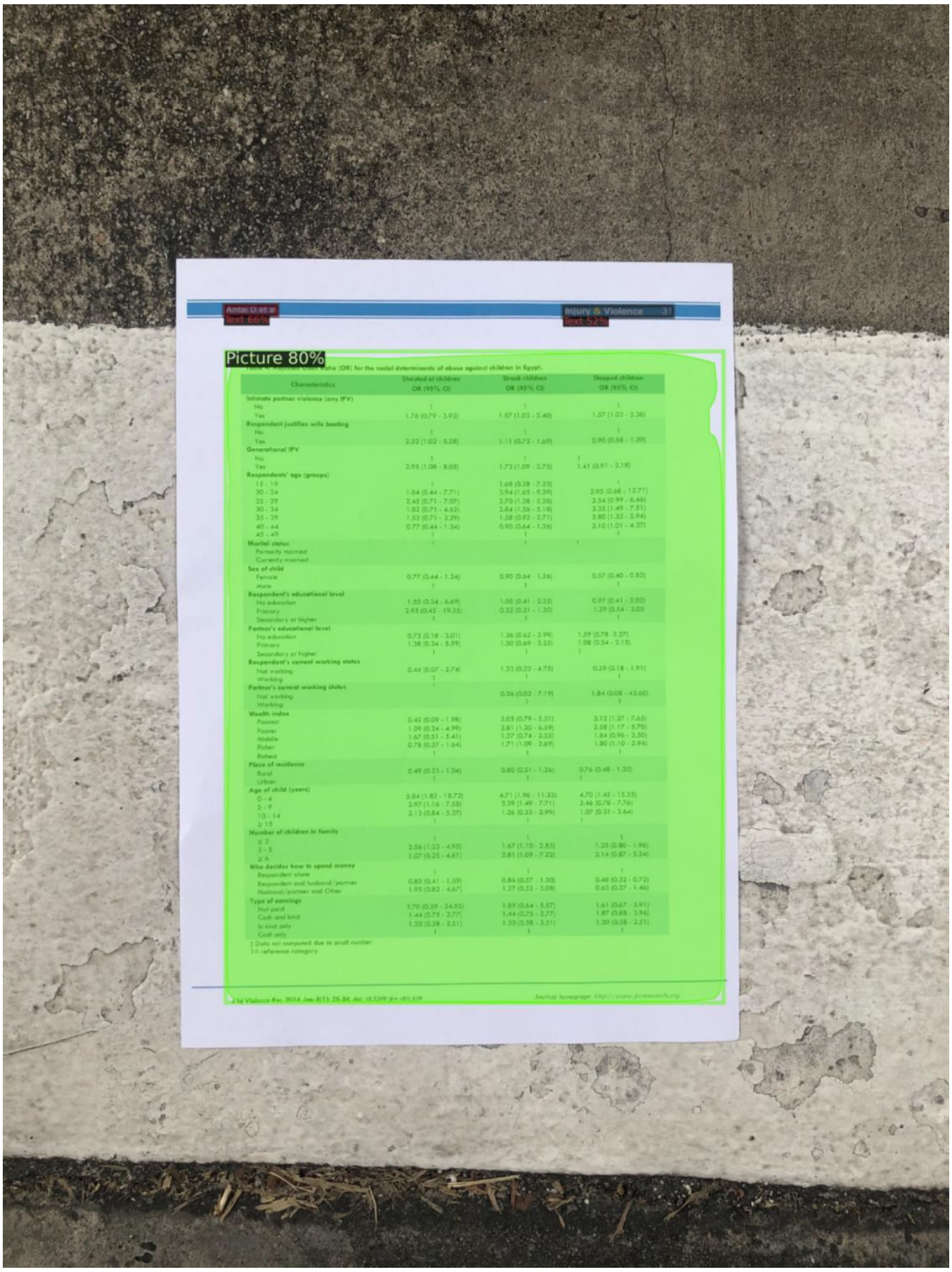


Figure 12: Prediction using the Document Model

Prediction on original image using the Layout Model



Antai D et al

Injury & Violence 31

Ratio (OR) for the social determinants of abuse against children in Egypt.

Characteristics	Shouted at children OR (95% CI)	Struck children OR (95% CI)	Slapped children OR (95% CI)
Intimate partner violence (any IPV)			
No	1	1	1
Yes	1.76 (0.79 - 3.92)	1.57 (1.03 - 2.40)	1.57 (1.03 - 2.38)
Respondent justifies wife beating			
No	1	1	1
Yes	2.32 (1.02 - 5.28)	1.11 (0.73 - 1.69)	0.90 (0.58 - 1.39)
Generational IPV			
No	1	1	1
Yes	2.95 (1.08 - 8.05)	1.73 (1.09 - 2.75)	1.41 (0.91 - 2.18)
Respondents' age (groups)			
15 - 19	1	1.68 (0.38 - 7.35)	1
20 - 24	1.84 (0.44 - 7.71)	3.94 (1.65 - 9.39)	2.95 (0.68 - 12.71)
25 - 29	2.45 (0.71 - 7.07)	2.70 (1.38 - 5.28)	2.54 (0.99 - 6.46)
30 - 34	1.82 (0.71 - 4.62)	2.84 (1.56 - 5.18)	3.35 (1.49 - 7.51)
35 - 39	1.53 (0.71 - 3.29)	1.58 (0.92 - 2.71)	2.80 (1.32 - 5.94)
40 - 44	0.77 (0.44 - 1.34)	0.90 (0.64 - 1.26)	2.10 (1.01 - 4.37)
45 - 49	1	1	1
Marital status	‡	‡	‡
Formerly married			
Currently married			
Sex of child			
Female	0.77 (0.44 - 1.34)	0.90 (0.64 - 1.26)	0.57 (0.40 - 0.82)
Male	1	1	1
Respondent's educational level			
No education	1.50 (0.34 - 6.69)	1.05 (0.41 - 2.35)	0.97 (0.41 - 2.02)
Primary	2.95 (0.45 - 19.35)	0.52 (0.21 - 1.30)	1.29 (0.54 - 3.05)
Secondary or higher	1	1	1
Partner's educational level			
No education	0.73 (0.18 - 3.01)	1.36 (0.62 - 2.99)	1.59 (0.78 - 3.27)
Primary	1.38 (0.34 - 5.59)	1.50 (0.69 - 3.25)	1.08 (0.54 - 2.15)
Secondary or higher	1	1	1
Respondent's current working status			
Not working	0.44 (0.07 - 2.74)	1.23 (0.32 - 4.75)	0.59 (0.18 - 1.91)
Working	1	1	1
Partner's current working status	‡		
Not working		0.36 (0.02 - 7.19)	1.84 (0.08 - 43.60)
Working		1	1
Wealth index			
Poorest	0.42 (0.09 - 1.98)	2.05 (0.79 - 5.31)	3.12 (1.27 - 7.65)
Poorer	1.09 (0.24 - 4.99)	2.81 (1.20 - 6.59)	2.58 (1.17 - 5.70)
Middle	1.67 (0.51 - 5.41)	1.37 (0.74 - 2.55)	1.84 (0.96 - 3.50)
Richer	0.78 (0.37 - 1.64)	1.71 (1.09 - 2.69)	1.80 (1.10 - 2.94)
Richest	1	1	1
Place of residence			
Rural	0.49 (0.23 - 1.04)	0.80 (0.51 - 1.26)	0.76 (0.48 - 1.20)
Urban	1	1	1
Age of child (years)			
0 - 4	5.84 (1.82 - 18.72)	4.71 (1.96 - 11.32)	4.70 (1.45 - 15.25)
5 - 9	2.97 (1.16 - 7.58)	3.39 (1.49 - 7.71)	2.46 (0.78 - 7.76)
10 - 14	2.13 (0.84 - 5.37)	1.26 (0.53 - 2.99)	1.07 (0.31 - 3.64)
≥ 15	1	1	1
Number of children in family			
≤ 2	1	1	1
3 - 5	2.56 (1.33 - 4.95)	1.67 (1.10 - 2.53)	1.25 (0.80 - 1.96)
≥ 6	1.07 (0.25 - 4.61)	2.81 (1.09 - 7.22)	2.14 (0.87 - 5.24)
Who decides how to spend money			
Respondent alone	1	1	1
Respondent and husband/partner	0.85 (0.41 - 1.59)	0.86 (0.57 - 1.30)	0.48 (0.32 - 0.72)
Husband/partner and Other	1.95 (0.82 - 4.67)	1.27 (0.52 - 3.08)	0.63 (0.27 - 1.46)
Type of earnings			
Not paid	3.70 (0.39 - 34.95)	1.89 (0.64 - 5.57)	1.61 (0.67 - 3.91)
Cash and kind	1.44 (0.75 - 2.77)	1.44 (0.75 - 2.77)	1.87 (0.88 - 3.96)
In kind only	1.30 (0.58 - 2.51)	1.20 (0.58 - 2.51)	1.30 (0.58 - 2.51)
Cash only	1	1	1

‡ Data not computed due to small number
1 = reference category

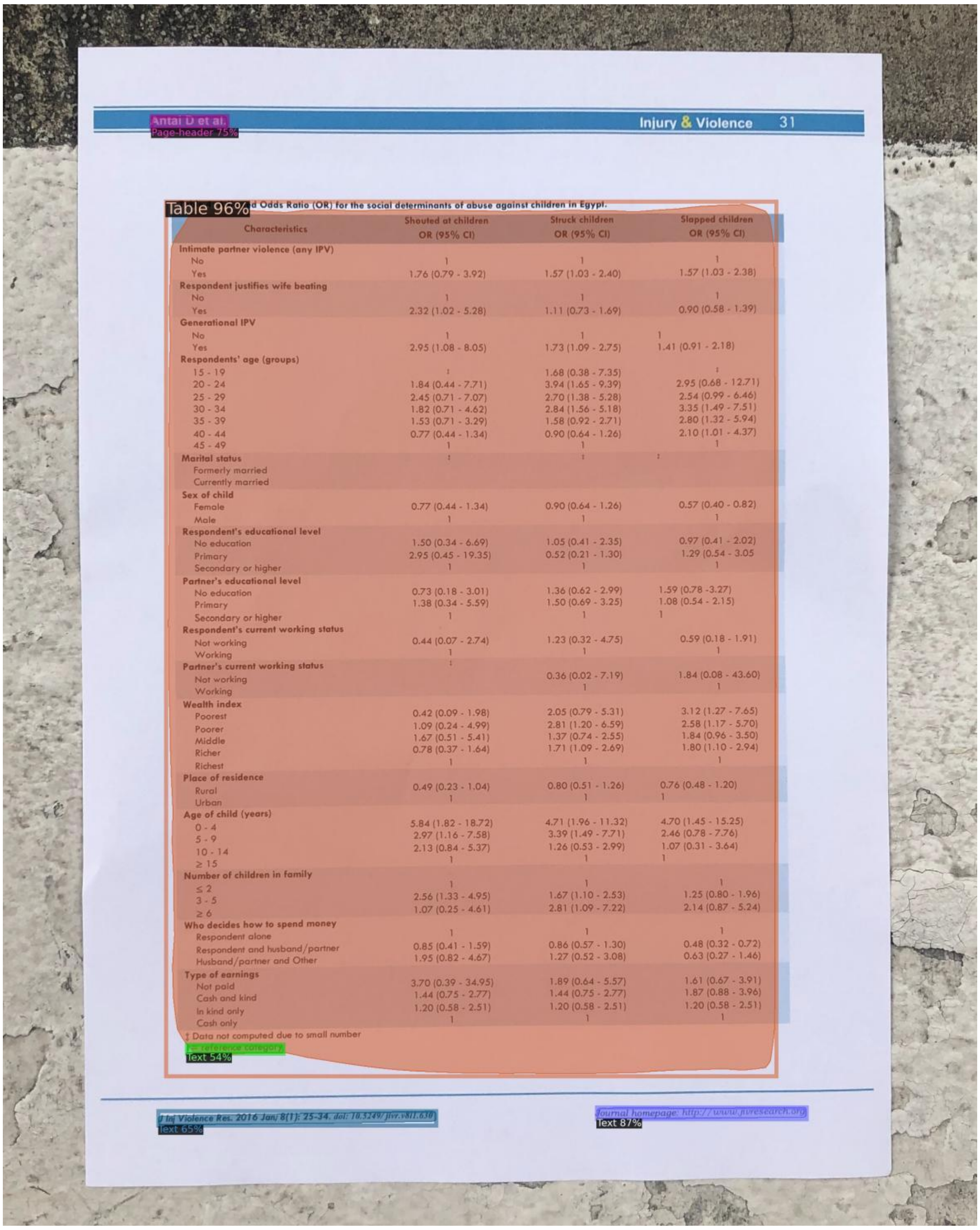
Inj Violence Res. 2014 Jan; 6(1): 25-34. doi: [illegible]

Journal homepage: http://www.jivresearch.org

Figure 13: Prediction using the Layout Model on original photo taking document

The Layout Model is capable of generating predictions for the image, but its accuracy is limited as it classifies table content under the Picture category and includes the text at the bottom as part of the Picture.

Prediction on Cropped Document


Page-header 75%
Table 96%
Text 54%
Text 65%
Text 87%

Antai D et al. Injury & Violence 31

d Odds Ratio (OR) for the social determinants of abuse against children in Egypt.

Characteristics	Shouted at children OR (95% CI)	Struck children OR (95% CI)	Slapped children OR (95% CI)
Intimate partner violence (any IPV)			
No	1	1	1
Yes	1.76 (0.79 - 3.92)	1.57 (1.03 - 2.40)	1.57 (1.03 - 2.38)
Respondent justifies wife beating			
No	1	1	1
Yes	2.32 (1.02 - 5.28)	1.11 (0.73 - 1.69)	0.90 (0.58 - 1.39)
Generational IPV			
No	1	1	1
Yes	2.95 (1.08 - 8.05)	1.73 (1.09 - 2.75)	1.41 (0.91 - 2.18)
Respondents' age (groups)			
15 - 19	‡	1.68 (0.38 - 7.35)	‡
20 - 24	1.84 (0.44 - 7.71)	3.94 (1.65 - 9.39)	2.95 (0.68 - 12.71)
25 - 29	2.45 (0.71 - 7.07)	2.70 (1.38 - 5.28)	2.54 (0.99 - 6.46)
30 - 34	1.82 (0.71 - 4.62)	2.84 (1.56 - 5.18)	3.35 (1.49 - 7.51)
35 - 39	1.53 (0.71 - 3.29)	1.58 (0.92 - 2.71)	2.80 (1.32 - 5.94)
40 - 44	0.77 (0.44 - 1.34)	0.90 (0.64 - 1.26)	2.10 (1.01 - 4.37)
45 - 49	1	1	1
Marital status	‡	‡	‡
Formerly married			
Currently married			
Sex of child			
Female	0.77 (0.44 - 1.34)	0.90 (0.64 - 1.26)	0.57 (0.40 - 0.82)
Male	1	1	1
Respondent's educational level			
No education	1.50 (0.34 - 6.69)	1.05 (0.41 - 2.35)	0.97 (0.41 - 2.02)
Primary	2.95 (0.45 - 19.35)	0.52 (0.21 - 1.30)	1.29 (0.54 - 3.05
Secondary or higher	1	1	1
Partner's educational level			
No education	0.73 (0.18 - 3.01)	1.36 (0.62 - 2.99)	1.59 (0.78 -3.27)
Primary	1.38 (0.34 - 5.59)	1.50 (0.69 - 3.25)	1.08 (0.54 - 2.15)
Secondary or higher	1	1	1
Respondent's current working status			
Not working	0.44 (0.07 - 2.74)	1.23 (0.32 - 4.75)	0.59 (0.18 - 1.91)
Working	1	1	1
Partner's current working status	‡		
Not working		0.36 (0.02 - 7.19)	1.84 (0.08 - 43.60)
Working		1	1
Wealth index			
Poorest	0.42 (0.09 - 1.98)	2.05 (0.79 - 5.31)	3.12 (1.27 - 7.65)
Poorer	1.09 (0.24 - 4.99)	2.81 (1.20 - 6.59)	2.58 (1.17 - 5.70)
Middle	1.67 (0.51 - 5.41)	1.37 (0.74 - 2.55)	1.84 (0.96 - 3.50)
Richer	0.78 (0.37 - 1.64)	1.71 (1.09 - 2.69)	1.80 (1.10 - 2.94)
Richest	1	1	1
Place of residence			
Rural	0.49 (0.23 - 1.04)	0.80 (0.51 - 1.26)	0.76 (0.48 - 1.20)
Urban	1	1	1
Age of child (years)			
0 - 4	5.84 (1.82 - 18.72)	4.71 (1.96 - 11.32)	4.70 (1.45 - 15.25)
5 - 9	2.97 (1.16 - 7.58)	3.39 (1.49 - 7.71)	2.46 (0.78 - 7.76)
10 - 14	2.13 (0.84 - 5.37)	1.26 (0.53 - 2.99)	1.07 (0.31 - 3.64)
≥ 15	1	1	1
Number of children in family			
≤ 2	1	1	1
3 - 5	2.56 (1.33 - 4.95)	1.67 (1.10 - 2.53)	1.25 (0.80 - 1.96)
≥ 6	1.07 (0.25 - 4.61)	2.81 (1.09 - 7.22)	2.14 (0.87 - 5.24)
Who decides how to spend money			
Respondent alone	1	1	1
Respondent and husband/partner	0.85 (0.41 - 1.59)	0.86 (0.57 - 1.30)	0.48 (0.32 - 0.72)
Husband/partner and Other	1.95 (0.82 - 4.67)	1.27 (0.52 - 3.08)	0.63 (0.27 - 1.46)
Type of earnings			
Not paid	3.70 (0.39 - 34.95)	1.89 (0.64 - 5.57)	1.61 (0.67 - 3.91)
Cash and kind	1.44 (0.75 - 2.77)	1.44 (0.75 - 2.77)	1.87 (0.88 - 3.96)
In kind only	1.20 (0.58 - 2.51)	1.20 (0.58 - 2.51)	1.20 (0.58 - 2.51)
Cash only	1	1	1

‡ Data not computed due to small number

= reference category

J Inj Violence Res. 2016 Jan; 8(1): 25-34. doi: 10.5249/jivr.v8i1.630

Journal homepage: http://www.jivresearch.org

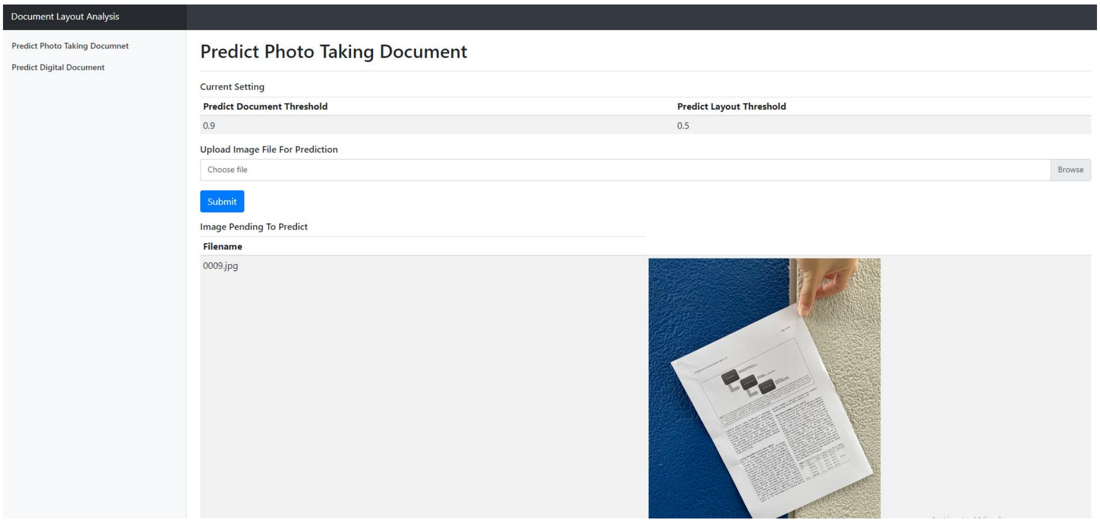
Figure 14: Prediction using the Document Model on predicted document area

The model's performance is improved when predicting on the cropped document, as it is able to correctly detect the table and label the text in the footer.

To sum up, the Layout Model attains a mean average precision of 64% across the 10 IoU thresholds ranging from 0.5 to 0.95, and at 0.5 IoU, the model yields an average precision of 86% using the test dataset. Additionally, with the aid of the Document Model, the Layout Model can also make reliable predictions for the 11 categories components in the layout of photo-taking document images.

4.4 Demo application

This project also included a simple demo application which developed using the Python Django framework. This application allow user to upload picture and check the prediction result.



Predict Photo Taking Document

Current Setting

Predict Document Threshold	Predict Layout Threshold
0.9	0.5

Upload Image File For Prediction

Image Pending To Predict

Filename
0009.jpg

Figure 15: Demo application

5. Conclusion

Project Summary

The primary objective of this project was to develop an accurate document layout detection model using deep learning techniques, specifically the Mask R-CNN approach. The final model achieved an overall average precision of 86%, with each category having a minimum average precision of 50%. The project aimed to build a model which capable of detecting layouts in various forms of documents, including digital, scanned, or photo form documents. The selection of dataset, training, and prediction techniques was made to achieve this goal. Results of the experiments are shared and discussed for future improvements and research in the field of document layout detection.

Project Limitations

One of the primary limitations of the project was the difficulty in adding photo-taking images to the dataset due to a lack of human resources available for labeling the samples for training. Although the model is designed to detect any type of document, it is more suitable for research papers or articles. The model may require additional fine-tuning for other types of documents, which can be further explored in future studies.

Future Enhancements

To enhance the accuracy and efficiency of document layout detection, future studies may consider introducing new models, each targeting a different document type such as magazines, exam papers, or comic books. The application of model assembly methods can be utilized to provide the most accurate result by utilizing the strengths of each model. Moreover, exploring the use of transfer learning techniques and the inclusion of more diverse datasets can also improve the model's performance and extend its applicability.

Reference

[1] Viso.ai. "Mask R-CNN." [Online]. Available: https://viso.ai/deep-learning/mask-r-cnn/#:~:text=%2DR%2DCNN.-What%20is%20Mask%20R%2DCNN%3F,Region%2DBased%20Convolutional%20Neural%20Network. [Accessed: Feb. 27, 2023].

[2] IBM Developer. "PubLayNet." [Online]. Available: https://developer.ibm.com/exchanges/data/all/PubLayNet/. [Accessed: Feb. 27, 2023].

[3] C. Van Beusekom and D. Keysers, "A Survey of Document Layout Analysis," ACM Journal on Computing and Cultural Heritage, vol. 14, no. 4, pp. 1-31, 2021. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3534678.3539043. [Accessed: Feb. 27, 2023].

[4] DocBank, "DocBank," [Online]. Available: https://doc-analysis.github.io/docbank-page/. [Accessed: Feb. 27, 2023].

[5] SG-VILab. "WarpDoc." [Online]. Available: https://sg-vilab.github.io/event/WarpDoc/. [Accessed: Feb. 27, 2023].

[6] V7Labs. "Mean Average Precision (mAP) in Object Detection." [Online]. Available: https://www.v7labs.com/blog/mean-average-precision#:~:text=Average%20Precision%20is%20calculated%20as,mAP%20varies%20in%20different%20contexts.. [Accessed: Feb. 27, 2023].

[7] COCO. "Detection Evaluation." [Online]. Available: https://cocodataset.org/#detection-eval. [Accessed: Feb. 27, 2023].

[8] Appen. "PubLayNet Data Preview." [Online]. Available: https://dax-cdn.cdn.appdomain.cloud/dax-PubLayNet/1.0.0/data-preview/index.html. [Accessed: Feb. 27, 2023].

[9] A. Littlepain, "Simple Understanding of Mask RCNN," Medium, 20-May-2020. [Online]. Available: https://alittlepain833.medium.com/simple-understanding-of-mask-rcnn-134b5b330e95. [Accessed: Feb. 27, 2023].

[10] Wu, Y., Kirillov, A., Massa, F., Lo, W. Y., & Girshick, R. (2019). Detectron2. [Online]. Available: https://github.com/facebookresearch/detectron2.

[11] X. Xu, J. Zhang and Y. Huang, "A Rule-Based Method for Text Region Detection in Historical Documents," 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1676-1681, doi: 10.1109/ICDAR.2019.00263.

[12] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 675-678).