applied sciences

MDPI

Article

Fi-Fo Detector: Figure and Formula Detection Using Deformable Networks

Junaid Younas [1,2,*], Shoaib Ahmed Siddiqui [1,2], Mohsin Munir [1,2], Muhammad Imran Malik [3,4], Faisal Shafait [3,4], Paul Lukowicz [1,2] and Sheraz Ahmed [2]

1 Department of Computer Science, Technical University, 67663 Kaiserslautern, Germany; shoaib_ahmed.siddiqui@dfki.de (S.A.S.); mohsin.munir@dfki.de (M.M.); paul.lukowicz@dfki.de (P.L.)
2 German Research Institute for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany; sheraz.ahmed@dfki.de
3 School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), H-12, Islamabad 44000, Pakistan; malik.imran@seecs.edu.pk (M.I.M.); faisal.shafait@seecs.edu.pk (F.S.)
4 Deep Learning Laboratory, National Centre of Artificial Intelligence (NCAI), H-12, Islamabad 44000, Pakistan
* Correspondence: junaid.younas@dfki.de

Received: 17 August 2020; Accepted: 11 September 2020; Published: 16 September 2020

Abstract: We propose a novel hybrid approach that fuses traditional computer vision techniques with deep learning models to detect figures and formulas from document images. The proposed approach first fuses the different computer vision based image representations, i.e., color transform, connected component analysis, and distance transform, termed as Fi-Fo image representation. The Fi-Fo image representation is then fed to deep models for further refined representation-learning for detecting figures and formulas from document images. The proposed approach is evaluated on a publicly available ICDAR-2017 Page Object Detection (POD) dataset and its corrected version. It produces the state-of-the-art results for formula and figure detection in document images with an f1-score of 0.954 and 0.922, respectively. Ablation study results reveal that the Fi-Fo image representation helps in achieving superior performance in comparison to raw image representation. Results also establish that the hybrid approach helps deep models to learn more discriminating and refined features.

Keywords: figure detection; formula detection; Fi-Fo image representation; deep learning; deformable convolution; computer vision

1. Introduction

Digitization of document images is a growing need for commercial and non-commercial entities, for example, banks, industries, educational institutes, and libraries. Aside from record-keeping, it significantly improves the availability of data just at a click and/or a tap from anywhere in the world, at any time. These digitized documents can be processed in an automated fashion given that the information contained in those documents can be extracted reliably. Reliable extraction of information from documents has been a major focus of the document analysis community for decades [1–4].

Figures are an integral part of a range of different types of documents as they portray the maximum amount of information in the least amount of space/time. Formulas, on the other hand, are the best way to express these relations symbolically leveraging the power of mathematics at its core. Detection of formulas and figures from document images is a challenging task as document images are composed of multi-level information. The information encapsulated in a document includes title, author details, corresponding text, figures, formulas, and many other related objects.

Figure detection from document images is a challenging and crucial task. Figure detection is a prefatory step in document image processing systems, enabling these systems to discriminate between textual and non-textual regions present in a document. Figures that are usually present in document images include layout design, block diagrams, natural images, and plots/graphs. Decorative graphics, i.e., long lines and "rules", are not considered as figures in this work. Similarly, formulas are presented as a two-dimensional arrangement, with distinctive structural features as compared to the plain text, which is one-dimensional. They can portray complex inter-relationship between different entities in a concise form. The advantages of the ability to recognize formulas are twofold: (i) Formula detection eases the dissemination and retrieval of mathematical knowledge from document images and (ii) enhances the performance of text recognition systems like optical character recognition systems (OCRs) as the conventional text processing pipeline should not be executed on those regions producing counterproductive transcriptions [5,6]. Figure and formula detection from document images is a challenging task as figures and formulas are usually spread widely across the document images at varying locations. Likewise, figure and formula appearance rely massively on the document format, style layout, orientation, aspect ratio, and other factors. Therefore, it is not easy to detect figures and formulas directly from document images, which could be a potential reason why existing commercial and open-source tools lack support for this functionality. Moreover, as table detection is already available in commercial OCRs [5], for example, Tesseract, Abby, therefore, tables are not considered for evaluation in this work.

Significant efforts have been made in the past to segment out the different page-objects in document images. Most of the early approaches heavily relied on heuristics—which are task-specific—and thus fail to generalize to novel scenarios [7,8]. Deep-learning based models have been leveraged for this segmentation in the more recent past [2,9–12]. All of these methods involve a significant amount of pre or post-processing based on hand-designed heuristics. A recent attempt has been made by Siddiqui et al. [3] to incorporate deformable CNNs for the analysis of document images. However, the potential of deep-learning methods in combination with traditional computer vision approaches has not been well explored in this context.

In this work, we present a generic, data-driven, and end-to-end method, *Fi-Fo detector*, based on deformable Feature Pyramid Network (FPNs) [13] for detection of figures and formulas in document images. The Fi-Fo detector successfully identifies figures and formulas occurring at different scales, orientations, and aspect ratios. Furthermore, the Fi-Fo detector also leverages the potential of combination of computer vision techniques to further boost the capabilities of the deep model. We particularly leverage a novel combination of traditional approaches, which includes inverse distance transform and connected components analysis (CCA) along with the gray-scale version of the raw input image to further strengthen the capabilities of the model as color features are not particularly useful in telling these page-objects apart. We evaluated the proposed method on the publicly available ICDAR-2017 POD competition dataset. The major contributions of this research work are two-fold:

- A hybrid approach that glues traditional computer vision approaches with deep-learning for refined representation learning to detect heterogeneous objects in document images, figures and formulas in particular.
- Refinement of the ICDAR-2017 POD dataset to eliminate disproportions and confusions.
- Ablation study of the proposed method on a large publicly available dataset ICDAR-2017 POD to justify the efficacy of the proposed approach.

The rest of the paper is structured as follows. Section 2 summarizes the recent developments and previous state-of-the-art systems in the domain of page-object detection. Section 3 covers the Fi-Fo Detector and its components, along with a detailed analysis of the methodology. Section 4 presents an overview of the dataset, training details, evaluation protocol followed by the presentation of the obtained results, along with a brief discussion. Section 5 concludes and provides an outlook for future work.

2. Related Work

Document image processing is an interesting topic among the computer vision research community. Significant progress has been made in this domain, including heuristic-based, convolutional neural network (CNN) based, statistics-based-like conditional random fields (CRFs) and graph trees, and\or a combination of these methods [7,8,14–16]. Heuristics include color-based features, shape-based features, geometric features, and keypoint descriptors. Deep-learning based approaches use CNNs [16], fully convolutional networks (FCNs) [2], region proposal networks (RPNs) [11], and deformable CNNs [3]. Tasks performed on document images include (but are not limited to) textual and non-textual region discrimination, graphics, and page object detection, which includes text, formulas, and figures.

Ha et al. [7] presented a fundamental image segmentation method based on the recursive X-Y cut algorithm. Their method used a projection profile based on the spatial configuration of connected components (CCs), which extracts columns from document page images. It might appear a simple method today but has opened up a new research direction in page object segmentation and detection decades back.

Chiu et al. [17] presented an OCR based picture detection method from document images. OCR is applied to detect text regions followed by segmentation methods to mask them out, and finally, non-textual regions are clustered together. Segmentation is further improved using caption information in post-processing. This approach depends not only on the performance of the OCR but also on the subsequent steps, which are to be executed precisely to achieve better results.

Kavasidis et al. [16] presented a saliency-based CNN for table and chart detection from digitized images. They applied saliency detection on input images to preserve the contextual information. FCN is used as a base detector followed by fully-connected CRFs for localizing tables and charts. They evaluated the presented method on the extended version of the ICDAR-2013 dataset. This approach is not only multi-step, but an extended version of the used dataset is not publicly available to draw comparisons.

Yi et al. [10] presented a page object detection method using region proposal CNNs, followed by a custom algorithm to refine proposed regions, and a CNN classifier for object category classification. It first pre-processes the input image by applying a component-based region proposal algorithm customized for document images, which extracts the rough region proposals at the initial stage and prunes them later. The refined region proposals are fed to the CNN model for classification. The results of CNN models are finally post-processed by a dynamic algorithm to optimize the detected region proposals. They evaluated their system on a private dataset and considered four objects, i.e., text-lines, figures, formulas, and tables.

Iwatsuki et al. [14] presented a CRF based method to extract formulas and mathematical zones from PDF documents. Their method uses layout features like font, style, and linguistic features such as *n-gram* context to build their CRF model. Phong et al. [15] developed a new method for the detection of mathematical expressions. They used OCR to analyze layout, text-lines, and expressions. Features are extracted from expressions using the Fast Fourier Transform and mean square error. SVM classifiers were applied on extracted features to classify mathematical expressions and formulas.

Gao et al. [18] presented a combination of a convolutional neural network and recurrent neural network (RNN) models to detect formulas from PDF documents. A combination of CNN and RNN models enables this method to preserve both character and visual features for formula detection. They applied a bottom-up and top-down strategy to generate formula region candidates, followed by feature extraction networks (CNNs and RNNs) and post-processing for refined formula region. This work, however, can only be applied to PDF documents, which is a considerable limitation when we talk about document images.

Deep learning, in the recent past, has become the center of attention for research in the document analysis community. Gilani et al. [11] and Schreiber et al. [12] leveraged Faster-RCNN for table

detection. Siddiqui et al. [3] additionally equipped the Faster-RCNN model with deformable property to gain significant improvements over prior state-of-the-art.

NLPR-PAL [19] were the winners of the ICDAR-2017 competition on POD. They presented a multi-stage approach for classification of figures, formulas, and tables from document images using the connected components of the input image, SVM classifiers, CNN based CRFs, Faster-RCNN, and normal CNNs. Lastly, final results are achieved by integrating the intermediate results of these stages.

A recent state-of-the-art POD system is proposed by Li et al. [9]. They proposed a hybrid model, which is a combination of deep structured prediction and supervised clustering for page object detection. First of all, they extract columns and then line regions of document images. They used conditional random fields formulated by a convolution neural network with unary and pairwise potentials to classify and cluster primitive region proposals from line regions. After classification, the same class clusters are merged to get page objects. Their presented approach is comprised of partially trainable networks with heuristics driven pre-processing and post-processing heads, which might limit its application in a generic scenario.

Saha et al. [20] presented the most recent method for page object detection in document images. Their approach is based on mask-RCNN for figure, formula, and table detection in document images. Although the presented approach has no strings (pre and\or post-processing) attached to it, they did not include compared their approach against the state-of-the-art results [9]. Moreover, they used different evaluation metrics rather than following the standards introduced in the ICDAR-2017 POD competition for page object detection. Thus their approach is not considered for comparison in this work.

Recently, systems have been presented for parsing, classifying, and localizing figures from PDF documents. Siegel et al. [21] presented a method to parse figures from PDF documents, parsed figures are then classified using graph-based CNNs. They also present a figure classification dataset namely "FigureSeer". Clark et al. [22] present another method "PDFFigures 2.0" to parse and classify figures from PDF documents along with a new dataset. Siegel et al. [23] present "DeepFigures" a deep neural method for detecting figures from PDF documents. These methods deal with PDF documents but not document images as in our case, so are out of scope for comparison in the presented work.

In this paper, we will be mainly comparing against the state-of-the-art system by Li et al. [9] and with NLRP-PAL [19], i.e., the winner of the ICDAR-2017 competition on page object detection from document images.

3. Fi-Fo Detector

3.1. System Overview

We use Fi-Fo image representation (instead of the raw image) as input to the network. The Fi-Fo image representation comprises of distance transform, connected-component analysis, and the original gray-scale image. We stack these three representations together before feeding them to the network. Fi-Fo detector is powered by a deformable FPN, which uses ResNet-101 as the backbone of the Fi-Fo detector to extract figures and formulas from document images. We will now discuss each of the individual components in the overall pipeline in detail, as shown in Figure 1.

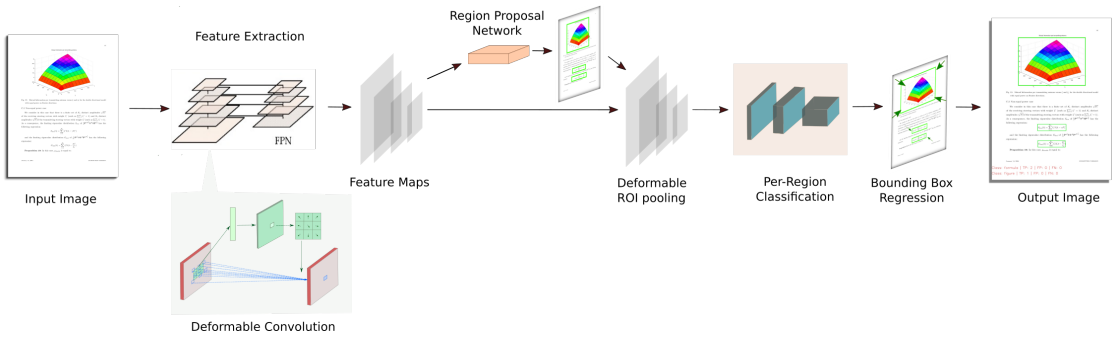


Figure 1. Proposed Fi-Fo detector outline based on a deformable Feature Pyramid Network (FPN).

3.2. Fi-Fo Image Representation

Deep neural networks dominate the counterparts when it comes to natural scene image processing whether its classification, segmentation, or object detection. However, document images are very different from natural scene images. Page objects appear at varying positions and differ among the documents depending on the document's format. Therefore, it is very important to preserve the contextual information, aiding the classifier in learning the desired patterns more efficiently. Fi-Fo image representation not only transforms the document image to appear as close as a natural scene image but also preserves the original image information in the form of a gray-scale image. Image transformation has already been used for page-object detection and segmentation in the past. Ha et al. [7] used vertical projection profiles to extract column regions and draw bounding boxes around connected components. Bukhari et al. [24] used connected components for document image segmentation, while Gilani et al. [11] used distance-based profiles, which were fed to the final classifier to extract the table structure from document images.

We use color transform, connected component analysis, and distance transform to generate Fi-Fo image representations. Gray-scale image retains the original information of the input image in a single channel. Connected component analysis is applied horizontally to identify regions in the image. Distance transform conserves the precise distance between page objects and blank regions. Additionally, we also reverse the distance transform, i.e., the maximum value occurs at the textual regions and diffuses smoothly as a function of the distance to the textual regions. We stack these representations together to feed them to the network. Figure 2 shows the Fi-Fo image representation with intermediate information in every channel.

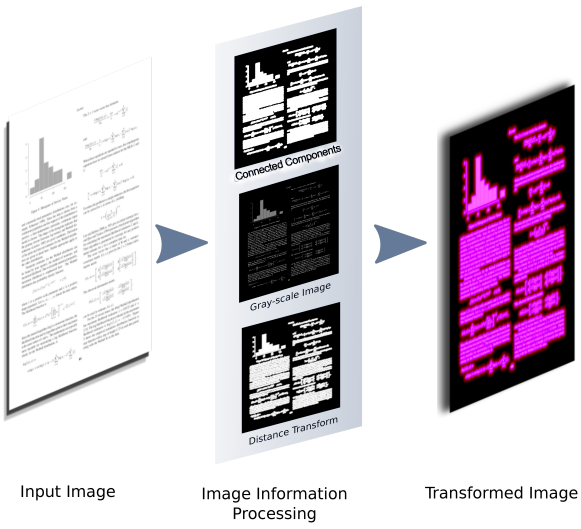


Figure 2. Fi-Fo image representation.

3.3. Fi-Fo Architecture

3.4. ResNet-101

We used pretrained ResNet-101 [25] as the backbone of the Fi-Fo detector. ResNet-101, as the name implies, consists of 101 convolution layers stacked together as 33 residual blocks, where each block is made up of three convolutional layers. As our focus is on the implementation of deformable CNNs for document images, regular ResNet-101 is transformed into its deformable variant. To achieve deformable functionality in ResNet-101, regular higher level convolution layers namely $res(5a, 5b, 5c)\_branch2b$ are replaced with their deformable counterparts. To benefit from transfer learning, we initialized deformable layers with zero offsets, making deformable convolutional layers equivalent to their non-deformable counterparts.

3.5. Deformable Convolutional Network (DCN)

The proposed method is based on deformable convolution networks (DCN) [26,27]. Convolutional neural networks learn the relevant feature representation depending on the task at hand [28]. These features are extracted in every layer using filters. Filters in lower convolutional layers usually capture textures and preliminary objects, which includes gradients, textures, materials, and colors, whereas filters in higher convolutional layers describe more abstract objects and their parts [29]. In traditional CNNs, a convolutional layer samples the input feature-maps at fixed locations, which the subsequent layers carry forward, resulting in a fixed and known geometric transformation. Using the fixed grid for the detection of objects occurring at different scales and different transformations is not ideal. Deformable convolutional networks address these constraints of traditional CNNs by introducing two additional modules to existing deep neural networks, namely (i) the deformable convolution and (ii) deformable RoI-pooling. Regular convolutional layers are augmented with a $2D$-offset convolutional layer to form the deformable convolution layer. Regular convolution operates on a uniform grid as its receptive field, whereas deformable convolution leverages the additional offset layers to augment the uniform grid conditioned on the input. The adaptive receptive field allows filters in convolution layers to adapt to different scales and transformation. Since objects like figures and formulas appear at vastly different scales, the deformable property significantly helps in coping up with these intense input variations. Mathematical formulation of deformable convolution is explained in [26] as:

$$y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \times x(p_0 + p_n + \Delta(p_n)) \tag{1}$$

where $\mathcal{R}$ defines the offsets from the point under consideration (p_0) in a regular-grid pattern, x represents the input, y represents the output feature map while w represents the filter weights. Considering a 3×3 convolutional layer, the set $\mathcal{R} = \{(-1,-1),(-1,0),(-1,1),(0,-1),(0,0),(0,1),(1,-1),(1,0),(1,1)\}$ defines this regular-grid comprising of the nine positions within the receptive field of the filter. In deformable convolution, sampling is done on irregular locations determined by the offset. The offset is defined as $\Delta(p_n)$, which augments the predefined offsets to arbitrarily deform the receptive field of the filter. Both the features as well as the offsets are learned by back-propagation of gradients. Since these offsets are fractional, they are implemented via bilinear interpolation. For simplicity, let us consider $p = p_0 + p_n + \Delta(p_n)$. Hence, the operation can be represented as:

$$x(p) = \sum_q G(q,p) \times x(q) \quad \text{where} \quad G(q,p) = g(q_x, p_x) \times g(q_y, p_y) \tag{2}$$

where q enumerates all the possible spatial locations on the feature map x, G is the bilinear interpolation kernel and g is defined to be $g(a,b) = max(0, 1-a-b)$. Region proposals are an integral part of object detection methods, achieved by using RoI-pooling, which converts an arbitrary sized input region into fixed-size feature representation. Regular RoI-pooling divides the RoI into $k \times k$ spatial bins. Similar to

the deformable convolution, deformable RoI-pooling introduces additional offsets that are added to spatial bins. This can be mathematically written as Equation (3).

$$y(i,j) = \sum_{p \in bin(i,j)} \frac{x(p_0 + p + \Delta p_i j)}{n_{ij}} \tag{3}$$

where $bin(i,j)$ defines a bin over spatial locations for feature aggregation ($\lfloor i\frac{w}{k} \rfloor <= p_x < \lceil (i+1)\frac{w}{k} \rceil$, $\lfloor j\frac{h}{k} \rfloor <= p_y < \lceil (j+1)\frac{h}{k} \rceil$), and n_{ij} represents the number of items in $bin(i,j)$. We refer readers to [26,27] for a comprehensive introduction to the deformable convolutional layers.

3.6. Network Architecture

The proposed Fi-Fo detector is based on deformable FPN [13], which integrates features from multiple scales within a single forward pass, transforming it into a faster variant of multi-scale detection. This makes it capable of better handling objects of small sizes. As a comparison, we also include results from deformable Faster-RCNN [30], and deformable RFCN [31], which were the most dominant architectures before FPN. All these models are augmented with deformable convolutions along with the replacement of a conventional RoI-pooling layer with deformable RoI-pooling.

The deformable convolutions generate offsets for every location in feature maps explicitly, making the process a memory-intensive operation. Therefore, all the models used for our experiments are built upon the *ResNet*-101, converted to a deformable network by replacing three higher level traditional layers into deformable counterparts to aid multi-scale feature extraction. This adaption enables us to leverage a deformable *ResNet*-101 as the base model for all the models used in our experiments.

The performance of deep neural networks rely heavily on the amount of data available for training, making them data-hungry [32]. Since the initial layers of the network are generic feature extractors, therefore, the initial layers trained on a large corpus of images are adapted as the feature extractor in our case, which are fine-tuned for the document analysis task during the course of training. This is commonly referred to as transfer learning in the literature, where the learned knowledge is transferred from one problem to another [3].

4. Experimental Results

4.1. Dataset

We used the publicly available ICDAR-2017 Page-Object Detection (POD) competition dataset [19], referred to as ICDAR-2017 POD, to benchmark the performance of our model. ICDAR-2017 POD was released recently for a competition focused on figure, formula, and table detection from document images. The dataset is comprised of page document images from 1500 scientific papers available at CiteSeer (http://csxstatic.ist.psu.edu/). This dataset is comprised of 2417 English document images, segregated into 1600 train and 817 test document images. The dataset exhibits high variability in terms of format and page layout. Page layout styles include single column, double column, and multi-column pages. There are various kinds of formulas, figures, tables, and other page objects spread across the document images.

The page-objects include textual content, page title, captions, headings, etc., but only figures, tables, and formulas were annotated for the task. Every document image is accompanied by a corresponding *.xml* file in PASCAL-VOC format annotated ground-truth. Page objects are annotated by rectangular co-ordinates to generate bounding boxes. Figure 3 shows some images from the ICDAR-2017 POD competition dataset along with the corresponding ground-truth information.

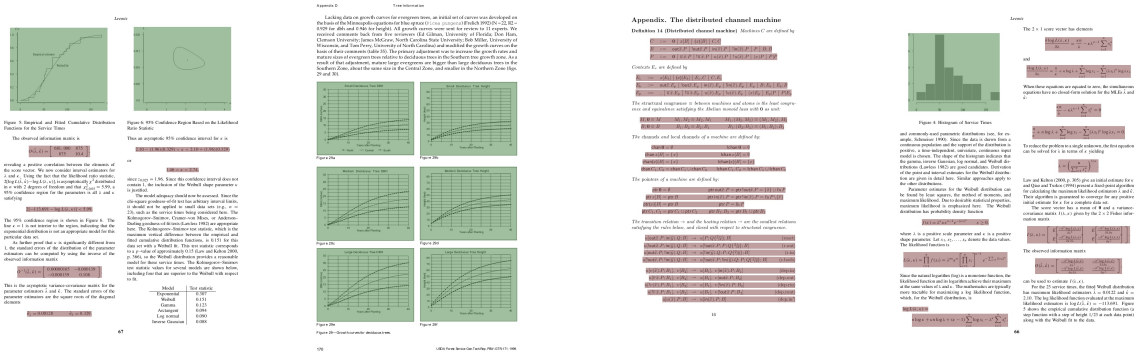


Figure 3. Annotated examples from ICDAR-2017 Page Object Detection (POD) dataset (formulas are labeled as brown color and figures as green color).

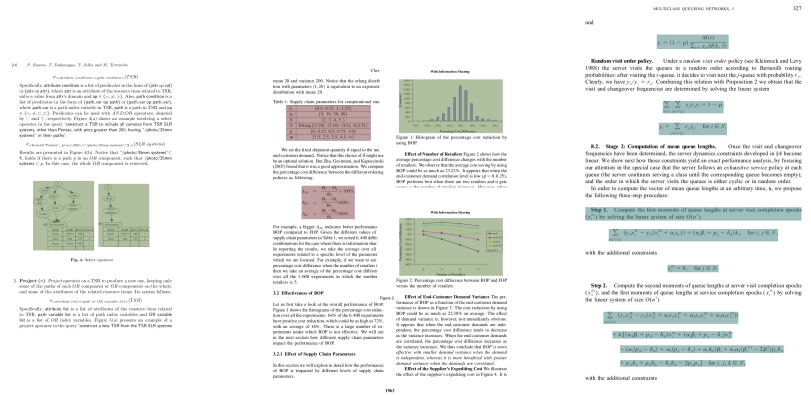
4.1.1. Faulty Annotations

Initial experiments led us to the discovery of the problems in annotations of the ICDAR-2017 POD dataset. Some of these problems are highlighted in Figure 4, where green, blue, and brown color annotate figures, formulas, and tables, respectively. There were missing annotations, which include formulas, figures, and tables, as shown in Figure 4a. There were occasions where page objects were mislabeled, i.e., formulas annotated as tables, text annotated as formulas, and so on. As shown in Figure 4c, text-lines are labeled as formulas which can confuse the system to distinguish text from formulas. Similarly, in Figure 4b, a block of formulas is labeled as a table, again creating vagueness for the system. There were inconsistencies in the figure's annotation, as shown in Figure 4d. Common problems with figure annotation include over-segmentation, under-segmentation, and inconsistent labeling for multi-panel figures, i.e., at some instances, objects enclosed inside a solid boundary are annotated as single figure considering outer bounding box, while for other similar instances, the outer boundary is neglected and enclosed objects are annotated individually. These irregularities in original annotations resulted in penalizing the overall performance of the system.

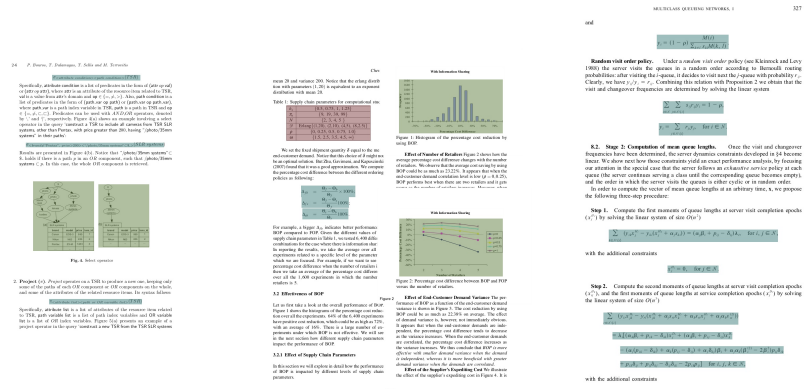
4.1.2. ICDAR-2017 POD (Corrected)

While evaluating the ICDAR-2017 POD dataset, the Fi-Fo detector highlighted the missing, confused, and irregular labels. As the Fi-Fo detector is a data-driven approach, it relies not only on the quantity of data but also the quality of data is very important. Confusions in the given data penalize the system resulting in degrading of performance, as is the case in the ICDAR-2017 POD dataset. Therefore, we decided to inspect the entire dataset manually to update missing annotations and fine-tune the confusing ones. While updating the ICDAR-2017 POD dataset, we did not add or remove any image from the dataset. Rather, we only updated the annotations to minimize the ambiguities and inconsistencies present in the dataset following the existing labeling conventions, as shown in Figure 4e. In the ICDAR-2017 POD dataset, some instances of decorative graphics are annotated as figures, and the rest were ignored. So, the only thing we completely changed is the removal of decorative graphics as figures from the original dataset. Figure 5c is a perfect example to demonstrate where the header-rule line was annotated as a figure, but the footer-rule line was ignored. All these problems contribute to the performance degradation of any data-driven system.

Bringing consistency and clarity in the annotations will help in achieving generalized systems for page object detection. A total of 273 annotation files were updated. The update included the removal of false labels, updating missing labels, and bringing uniform labeling conventions. An overview of the ICDAR-2017 POD (corrected) dataset along with the ICDAR-2017 POD dataset is presented in Table 1. The updated annotations have been publicly released as the ICDAR-2017 POD (corrected) dataset to aid future research in this direction (https://bit.ly/2AUSlzI).

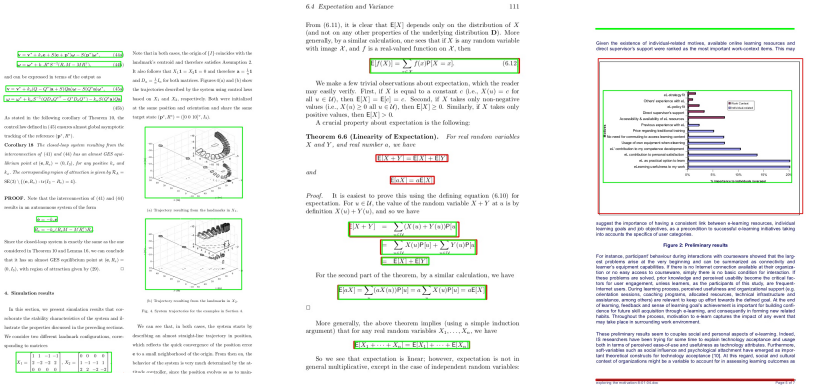


(**a**) Missing labels for formulas (**b**) Formulas labeled as table (**c**) Text labeled as formula (**d**) Problems with figure labels



(**e**) Corrected labels.

Figure 4. Problems with ICDAR-2017 POD dataset (1st row) in comparison to ICDAR-2017 POD (corrected) dataset (2nd row) (green, blue, and brown color represent figures, formulas, and tables, respectively, thanks to Fi-Fo Detector).



(**a**) Correct detection (**b**) Formula labels missing (**c**) Inconsistent labeling (**d**) Figure label missing

Figure 5. Fi-Fo evaluation results on ICDAR-2017 POD dataset with original annotations (green color represents annotated objects; while red color highlights the detection by Fi-Fo Detector).

Table 1. Overview of ICDAR-2017 POD (corrected) with class-wise comparison and modifications in the ICDAR-2017 POD dataset.

Class	ICDAR-2017 POD	ICDAR-2017 POD (Corrected)	# of Files Modified
Figure	2939	2912	135
Formula	5427	5463	156
Table	1016	1053	30

4.2. Model Configuration

We used deformable ResNet-101 as the backbone of our deformable detection models, along with model weights trained on the ImageNet dataset as described previously. We keep the rest of the object detection pipeline, after deformable pooling, intact—including per-region classification and bounding-box regression. Use of pretrained weights enables our approach for domain adaptation from natural scene images to document images. We trained three different variants of deformable models, which include deformable Faster-RCNN, deformable FCN, and deformable FPN. We used three different anchor ratios for all our models and they were set to $[0.5, 1, 2]$. We used five different anchor scales for R-FCN and Faster-RCNN set to $[2, 4, 8, 16, 32]$. FPNs have a built-in feature for multi-scale detection because of their top-down architecture, so only a single anchor scale of $[8]$ is used. We trained our models for 50 epochs with a learning rate of 0.000125 (with a learning rate schedule). We used aspect-aware image resizing with a max image size of 1280×800. All models were trained on a single NVIDIA V-100 GPU.

4.3. Evaluation Protocol

We follow the evaluation protocol defined for the ICDAR-2017 POD competition. We compute true positives (*TP*), false positives (*FP*), and false negatives (*FN*) during the testing phase. These results are computed by evaluating the test set on intersection over union (IoU) threshold of 0.6, and 0.8 for calculation of the given metrics. Results are reported using the metrics of precision, recall, f1-score, and mean average precision (mAP).

The precision metric evaluates how accurate the system's predictions are. It is calculated as follows:

$$Precision = \frac{\text{correct detection}}{\text{total detection}} \tag{4}$$

The recall metric is the measure of how good a system performs in finding all positive examples given in the test set. It is given by,

$$Recall = \frac{\text{correct detection}}{\text{total ground-truth annotations}} \tag{5}$$

Mean average precision (mAP) is computed as an average of maximum precision at different recall levels. Mathematical formulation of mean average precision is given as:

$$mAP = \frac{1}{Q} \sum_{r=1}^{Q} AP_r \tag{6}$$

We report results for every individual class using the official evaluation code provided by the ICDAR-2017 POD competition organizers.

4.4. Results and Discussions

We present an ablation study, which covers the potential of Fi-Fo image representation in comparison to raw image representation along with analysis of deformable networks with regards to non-deformable counterparts. Moreover, we present a comparative analysis to establish the utility and

effectiveness of the ICDAR-2017 POD (corrected) dataset with the existing state-of-the-art systems. We also evaluate the deformable variants of Faster-RCNN, R-FCN, and FPN. FPN forms the basis of the Fi-Fo detector as it comprehensively outperformed other deformable variants, owing to its multi-scale detection capabilities.

Using the ICDAR-2017 POD dataset, visual results looked convincing, as both formulas and figures were detected properly, but numbers were surprisingly not up to the expectations keeping in mind the potential of FPNs implemented with deformable convolution, as shown in Table 2. On an IoU threshold of 0.6, formulas were detected with the precision and recall of 0.909 and 0.927, along with an f1-score and average precision (AP) of 0.918 and 0.911, respectively. Once the IoU threshold was increased to 0.8, precision and recall dropped to 0.856 and 0.878 with an f1-score and AP of 0.867 and 0.854, respectively. Considering figures, the obtained numbers were 0.918, 0.883, 0.90, and 0.894 in terms of precision, recall, f1-score, and average precision, respectively. For the IoU threshold of 0.8, precision, recall, and f1-score went down to 0.871, 0.838, and 0.854, respectively.

Table 2. Comparison of Fi-Fo with existing state-of-the-art methods using ICDAR-2017 POD annotations both for training and testing.

ICDAR-2017 POD									
Method	**Class**	$IoU = 0.6$				$IoU = 0.8$			
		Precison	**Recall**	**F1-Score**	**AP**	**Precision**	**Recall**	**F1-Score**	**AP**
NLPR-PAL [19]	Formula	0.901	0.929	0.915	0.839	0.888	0.916	0.902	0.816
	Figure	0.920	0.933	0.927	0.849	0.892	0.904	0.898	0.805
Li et al. [9] ICDAR-2017 POD	Formula	0.93	0.953	0.942	0.878	0.921	0.944	0.932	0.863
	Figure	0.948	0.940	0.944	0.896	0.921	0.913	0.917	0.85
Deformable Faster R-CNN	Formula	0.882	0.738	0.803	0.660	0.638	0.534	0.582	0.337
	Figure	0.929	0.872	0.899	0.660	0.855	0.802	0.828	0.720
Deformable R-FCN	Formula	0.914	0.918	0.916	0.915	0.832	0.836	0.834	0.826
	Figure	0.904	0.920	0.912	0.903	0.86	0.875	0.867	0.864
Fi-Fo Detector	Formula	0.909	0.927	0.918	0.911	0.856	0.878	0.867	0.854
	Figure	0.918	0.883	0.90	0.894	0.871	0.838	0.854	0.861

As exhibited in Figure 5, the Fi-Fo detector is working fine, correctly detecting the page objects, i.e., figures and formulas in particular, but the performance is not reflected in numbers. Upon investigating the results, we discovered the irregularities and inconsistencies in the original annotations available for the ICDAR-2017 POD dataset. There were clear examples of missing annotations for page objects, as the case in Figure 5b where annotations for formulas were missing. Confusions between figure and table annotations are briefly shown in Figure 5d. Figure 5c establishes the case of inconsistent labelling for figure annotations. There were examples of over-segmented ground-truth where captions or text-lines were annotated along with figure or table, as shown in Figure 4c. These inconsistencies have been already discussed in detail in Section 4.1.2.

Problems in the original annotations of the ICDAR-2017 POD dataset led us to update the annotations, which included removal of discrepancies, confusions, along with the addition of missing labels. After addressing the problems found in the ICDAR-2017 POD dataset, a clean dataset is presented as ICDAR-2017 POD (corrected). To establish a fair comparison with existing state-of-the-art methods, it is necessary to present their results on ICDAR-2017 POD (corrected). Li et al.'s [9] method is not an end-to-end system, which means it is a combination of trainable and heuristic-based parts. Therefore, it is almost impossible to replicate the system performance by building it up from scratch. Upon request, the authors [9] excused to provide us their system because of the system's complexities but provided the results on the ICDAR-2017 POD dataset for comparison. It limits the scope of comparison on the corrected dataset, as both the system and results on the corrected dataset from the Li et al. [9] were not accessible. So, in the given circumstances, we opted for the best possible way

to establish a fair comparison with the existing state-of-the-art system. We trained the Fi-Fo detector using the ICDAR-2017 POD dataset, while evaluation was performed on ICDAR-2017 POD (corrected) for both the Fi-Fo detector and Li et al. [9], and the results are furnished in Table 3.

Table 3. Comparison of Fi-Fo with existing state-of-the-art methods using ICDAR-2017 POD in training and ICDAR-2017 POD (corrected) for testing.

Trained: ICDAR-2017 POD, Tested: ICDAR-2017 POD (Corrected)									
Method	**Class**	$IoU = 0.6$				$IoU = 0.8$			
		Precison	**Recall**	**F1-score**	**AP**	**Precision**	**Recall**	**F1-Score**	**AP**
Li et al. [9]	Formula	0.935	0.331	0.489	0.312	0.877	0.310	0.459	0.274
	Figure	0.918	0.292	0.443	0.271	0.888	0.283	0.429	0.253
Fi-Fo Detector	Formula	0.949	0.945	0.947	0.967	0.897	0.893	0.895	0.941
	Figure	0.930	0.932	0.931	0.97	0.899	0.900	0.899	0.952

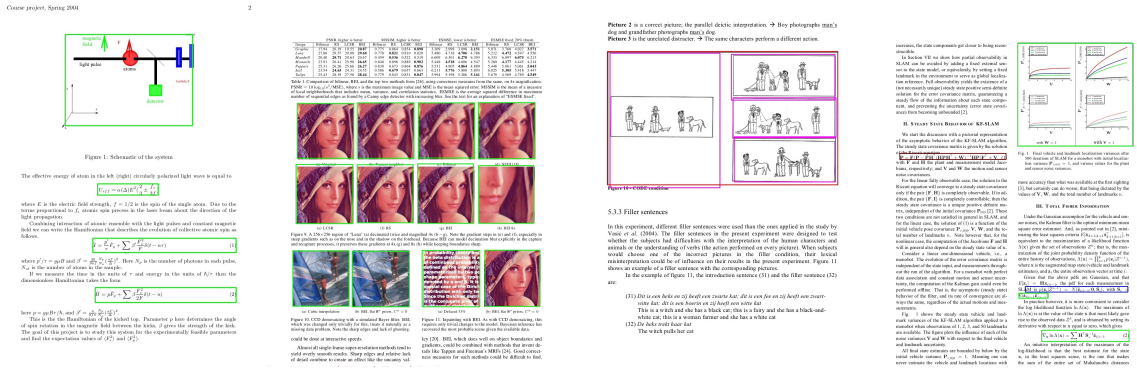
Results produced by the Fi-Fo detector on ICDAR-2017 POD (corrected) outperformed existing state-of-the-art system by large margins. On an IoU threshold of 0.6, Fi-Fo achieved the F1-score of 0.947 with the average precision (AP) of 0.967 in comparison to Li et al.'s [9] system with an F1-score of 0.489 and AP of 0.312, for formula detection. Similarly, there is an enormous difference in the results of figure detection produced by the Fi-Fo detector and Li et al. [9] on ICDAR-2017 POD (corrected). Fi-Fo detected the figures with an F1-score of 0.931, whereas Li et al.'s F1-score is 0.271, as shown in Table 3. On IoU threshold of 0.8, AP for Li et al. was computed to be 0.253 and 0.274 in comparison to Fi-Fo's AP score of 0.941 and 0.952 for figure and formula detection, respectively.

One of the potential reasons for dramatic decline in performance of Li et al.'s method on updated annotations might be its inability as an end-to-end system. Since the dataset itself had inconsistencies, state-of-the-art methods had to leverage hand-defined heuristics, which catered for these inconsistencies. Upon removal of these inconsistencies from the dataset, the heuristics themselves had to be adapted, which is a major short-coming of heuristics-based systems. The given results establish the weakness of the existing state-of-the-art system both on the ICDAR-2017 POD as well as the ICDAR-2017 POD (corrected) dataset. Using ICDAR-2017 POD, their system failed to detect and highlight the missing or wrong labels, and it failed to capitalize on the ICDAR-2017 POD (corrected). It is worth emphasizing again that all the reported results on the original and updated annotations in this paper are generated using the official evaluation code released by organizers of the ICDAR-2017 competition on page object detection [18].

Despite reporting significantly high metric-scores, Li et al.'s [9] method was unable to find any inconsistencies in the ICDAR-2017 POD dataset. As their approach relies heavily on heuristics along with pre/post-processing, which were specifically tuned to the inconsistencies present in the dataset. This is one of the primary reasons why purely data-driven techniques were not found to be very effective for this dataset in comparison to using hand-defined heuristics. This provides a clear edge to the Fi-Fo detector over Li et al.'s [9] method, in terms of generalization. Fi-Fo detector demonstrated to be a generic network by pointing out discrepancies in the ICDAR-2017 POD dataset, highlighted in Figure 5. Secondly, the Fi-Fo detector does not rely on any pre/post-processing, rather simple image transforms providing a clear edge not only in terms of efficiency and computation costs, but also the state-of-the-art results, as shown in Figure 6a,b. We also report results from the ablation study where we removed components from the Fi-Fo detector to identify the contribution of the individual components to the system.

To further validate the performance and potential of Fi-Fo detector, we present an ablation study that covers the potential of Fi-Fo image representation and deformable neural networks in comparison to raw image representation and non deformable neural networks. We observed a clear performance boost from raw image representation to Fi-Fo image representation and non-deformable neural networks to deformable neural networks, results are furnished in Table 4. The state-of-the-art

performance is achieved by using the combination of Fi-Fo image representation with deformable neural networks. Using an IoU threshold of 0.6 as per ICDAR-2017 POD competition standards, we achieved state-of-the-art results for formula detection with a precision of 0.957 and recall of 0.952 which translates to an F1-score of 0.954. Results with an IoU of 0.8 are 0.913, 0.908, and 0.91 in terms of precision, recall, and F1-score for formula detection, respectively. In figure detection, at an IoU threshold of 0.6, Fi-Fo detector achieved a precision of 0.931, recall of 0.913 along with an F1-score of 0.905. Setting the IoU threshold to 0.8 translated into a precision of 0.901, recall of 0.885 along with F1-score of 0.893. In terms of average precision, Fi-Fo detector outperformed other methods with a significant margin. At IoU=0.6, average precision of figure and formula detection was found to be 0.949 and 0.905, while at IoU=0.8, the average precision of 0.898 and 0.870 for formulas and figures was achieved, respectively.



(**a**) TPs of figures formulas (**b**) TPs of figure grid (**c**) FPs & FNs of figures (**d**) FPs & FNs for formulas

Figure 6. Analysis of results generated by Fi-Fo detector using the ICDAR-2017 POD (corrected) dataset, green color highlights true positives (TPs), blue color signifies false positives (FPs) for formulas and magenta color flags FPs for figures, and red color annotates FNs for both classes.

Table 4. An ablation study on the performance of Fi-Fo detector using raw image representation, Fi-Fo image representation, using non deformable FPN, and deformable FPN, both for ICDAR-2017 POD and ICDAR-2017 POD (corrected).

ICDAR-2017 POD										
Method	**Image Representation**	**Class**	$IoU = 0.6$				$IoU = 0.8$			
			Precison	**Recall**	**F1-Score**	**AP**	**Precision**	**Recall**	**F1-Score**	**AP**
Fi-Fo Detector Deformable	Raw	Formula	0.867	0.918	0.892	0.893	0.780	0.826	0.802	0.780
		Figure	0.860	0.869	0.864	0.847	0.818	0.827	0.822	0.799
Fi-Fo Detector Non Deformable	Fi-Fo	Formula	0.867	0.874	0.871	0.917	0.712	0.694	0.703	0.837
		Figure	0.856	0.821	0.838	0.929	0.801	0.739	0.769	0.889
Fi-Fo Detector Deformable	Fi-Fo	Formula	0.909	0.927	0.918	0.911	0.856	0.878	0.867	0.854
		Figure	0.918	0.883	0.90	0.894	0.871	0.838	0.854	0.861
ICDAR-2017 POD (Corrected)										
Fi-Fo Detector Deformable	Raw	Formula	0.949	0.945	0.947	0.973	0.897	0.893	0.895	0.967
		Figure	0.930	0.932	0.931	0.971	0.897	0.90	0.899	0.959
Fi-Fo Detector Non Deformable	Fi-Fo	Formula	0.910	0.927	0.918	0.953	0.860	0.877	0.868	0.928
		Figure	0.879	0.822	0.850	0.948	0.847	0.792	0.819	0.958
Fi-Fo Detector Deformable	Fi-Fo	Formula	0.957	0.952	0.954	0.949	0.913	0.908	0.910	0.898
		Figure	0.931	0.913	0.922	0.905	0.901	0.885	0.893	0.870

The obtained results highlight the superiority of the deformable model family for this task where the models either outperformed or achieved performance on par with heuristic-based methods.

Moreover, it is also demonstrated that the Fi-Fo detector shows progressive performance moving from ICDAR-2017 POD to ICDAR-2017 POD (corrected) at every subsequent step, whereas existing state-of-the-art failed to do so, as shown in Table 4. Since Fi-Fo is a data-driven approach, significant improvements in performance could be achieved by increasing the amount of training data. Results can further be boosted by post-processing particularly for the detected figure regions using computer-vision approaches, considering the case of Figure 6c into account.

We did fine-tune the annotations of ICDAR-2017 POD and make them publicly available as ICDAR-2017 POD (corrected), but the problem requires much more effort in terms of relabeling the document page objects, potentially requiring an introduction of a new and cleaned POD dataset based on hard defined conventions, which is out of the scope of current work. Here, we aim to demonstrate the potential of a hybrid approach, a combination of Fi-Fo image representation and deformable neural networks to detect page objects from document images.

5. Conclusions

We propose a combination of Fi-Fo image representation and deep neural networks to detect figures and formulas from document images. Our presented approach is novel, as it utilized traditional computer vision techniques to complement the performance of deep-learning models. The presented approach is generic, as it is capable of detecting page objects, i.e., figures and formulas, despite varying page formats and layouts. This paper establishes the state-of-the-art results with the Fi-Fo detector, an end-to-end system for figure and formula detection from document images. An ablation study is also presented to signify the importance of the parts of the Fi-Fo detector, and the impact of their combination for document image processing. There are many confusions and inconsistencies in the ICDAR-2017 POD dataset. We fine-tuned the dataset to remove inconsistencies and confusions, and presented as the ICDAR-2017 POD (corrected) dataset. This will contribute to the development of more generalized methods and systems for page-object detection in the future.

In the future, the existing region proposal network can be replaced with attention-based region proposal networks. This can be beneficial in generating improved and better region proposals. We strongly believe that curating a new dataset, which includes more page objects like title, headings, captions, etc., with existing classes, is an important avenue for pushing the state-of-the-art in this direction.

Author Contributions: Methodology, experimentation, writing, J.Y., S.A.S.; hypothesis, evaluation, review/updation, J.Y., M.M., M.I.M., S.A., F.S.; Project administration, P.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research work is partially funded by the Higher Education Commission of Pakistan and Forschungsinitiative under the project "AI Enhanced Cognition and Learning".

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kieninger, T.; Dengel, A. The T-Recs Table Recognition and Analysis System. In *Document Analysis Systems: Theory and Practice*; Springer: Berlin/Heidelberg, Germany, 1999, pp. 255–270.
2. Younas, J.; Afzal, M.Z.; Malik, M.I.; Shafait, F.; Lukowicz, P.; Ahmed, S. D-StaR: A Generic Method for Stamp Segmentation from Document Images. In Proceedings of the 14th International Conference on Document Analysis and Recognition, Kyoto, Japan, 10–15 November 2017; Volume 1, pp. 248–253.
3. Siddiqui, S.A.; Malik, M.I.; Agne, S.; Dengel, A.; Ahmed, S. DeCNT: Deep Deformable CNN for Table Detection. *IEEE Access* **2018**, *6*, 74151–74161. [CrossRef]
4. Ahmed, S.; Shafait, F.; Liwicki, M.; Dengel, A. A Generic Method for Stamp Segmentation Using Part-Based Features. In Proceedings of the 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 708–712.
5. Smith, R. An Overview of the Tesseract OCR Engine. In Proceedings of the 9th International Conference on Document Analysis and Recognition, Parana, Brazil, 23–26 September 2007; Volume 2, pp. 629–633.

6. Breuel, T. *The OCRopus Open Source OCR System*; International Society for Optics and Photonics: Bellingham, WA, USA, 2008; Volume 6815, p. 68150.
7. Haralick, R.M.; Phillips, I.T. Recursive X-Y cut using bounding boxes of connected components. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, USA, 14–16 August 1995; Volume 2, pp. 952–955. [CrossRef]
8. Das, A.K.; Chowdhury, S.P.; Mandal, S.; Chanda, B. Automated Segmentation of Math-Zones from Document Images. In Proceedings of the 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; Volume 3, p. 755. [CrossRef]
9. Li, X.H.; Yin, F.; Liu, C.L. Page Object Detection from PDF Document Images by Deep Structured Prediction and Supervised Clustering. In Proceedings of the 24th International Conference on Pattern Recognition, Beijing, China, 20–24 August 2018; pp. 3627–3632.
10. Yi, X.; Gao, L.; Liao, Y.; Zhang, X.; Liu, R.; Jiang, Z. CNN Based Page Object Detection in Document Images. In Proceedings of the 14th Internation Conference on Document Analysis and Recognition, Kyoto, Japan, 10–15 November 2017; Volume 1, pp. 230–235. [CrossRef]
11. Gilani, A.; Qasim, S.R.; Malik, M.I.; Shafait, F. Table Detection Using Deep Learning. In Proceedings of the 14th Internation Conference on Document Analysis and Recognition, Kyoto, Japan, 10–15 November 2017; pp. 771–776. [CrossRef]
12. Schreiber, S.; Agne, S.; Wolf, I.; Dengel, A.; Ahmed, S. DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images. In Proceedings of the 14th Internation Conference on Document Analysis and Recognition, Kyoto, Japan, 10–15 November 2017; pp. 1162–1167.
13. Lin, T.Y.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S.J. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
14. Iwatsuki, K.; Sagara, T.; Hara, T.; Aizawa, A. Detecting In-line Mathematical Expressions in Scientific Documents. In Proceedings of the 2017 ACM Symposium on Document Engineering, Valletta, Malta, 4–7 September 2017; pp. 141–144. [CrossRef]
15. Phong, B.H.; Hoang, T.M.; Le, T. A new method for displayed mathematical expression detection based on FFT and SVM. In Proceedings of the 4th NAFOSTED Conference on Information and Computer Science, Hanoi, Vietnam, 24–25 November 2017; pp. 90–95. [CrossRef]
16. Kavasidis, I.; Pino, C.; Palazzo, S.; Rundo, F.; Giordano, D.; Messina, P.; Spampinato, C. *A Saliency-Based Convolutional Neural Network for Table and Chart Detection in Digitized Documents*; Springer: Cham, Switzerland, 2019; pp. 292–302.
17. Chiu, P.; Chen, F.; Denoue, L. Picture Detection in Document Page Images. In Proceedings of the 10th ACM Symposium on Document Engineering, Manchester, UK, 21–24 September 2010; pp. 211–214. [CrossRef]
18. Gao, L.; Yi, X.; Liao, Y.; Jiang, Z.; Yan, Z.; Tang, Z. A Deep Learning-Based Formula Detection Method for PDF Documents. In Proceedings of the 14th Internation Conference on Document Analysis and Recognition, Kyoto, Japan, 10–15 November 2017; Volume 1, pp. 553–558.
19. Gao, L.; Yi, X.; Jiang, Z.; Hao, L.; Tang, Z. ICDAR2017 Competition on Page Object Detection. In Proceedings of the 14th International Conference on Document Analysis and Recognition,Kyoto, Japan, 10–15 November 2017; Volume 1, pp. 1417–1422.
20. Saha, R.; Mondal, A.; Jawahar, C.V. Graphical Object Detection in Document Images. In Proceedings of the 15th Internation Conference on Document Analysis and Recognition, Sydney, Australia, 22–27 September 2019.
21. Siegel, N.; Horvitz, Z.; Levin, R.; Divvala, S.; Farhadi, A. *FigureSeer: Parsing Result-Figures in Research Papers*; Computer Vision—ECCV 2016; Springer International Publishing: Cham, Switzerland, 2016; pp. 664–680.
22. Clark, C.; Divvala, S. PDFFigures 2.0: Mining Figures from Research Papers. In Proceedings of the JCDL '16 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, Newark, NJ, USA, 19–23 June 2016; pp. 143–152. [CrossRef]
23. Siegel, N.; Lourie, N.; Power, R.; Ammar, W. Extracting Scientific Figures with Distantly Supervised Neural Networks. In Proceedings of the JCDL '18 18th ACM/IEEE on Joint Conference on Digital Libraries, Fort Worth, TX, USA, 3–7 June 2018; pp. 223–232. [CrossRef]

24. Bukhari, S.; Al Azawi, M.; Shafait, F.; Breuel, T. Document Image Segmentation Using Discriminative Learning over Connected Components. In Proceedings of the 9th International Workshop on Document Analysis Systems, Boston, MA, USA, 9–11 June 2010; pp. 183–190. [CrossRef]
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
26. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 764–773.
27. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable ConvNets v2: More Deformable, Better Results. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
28. Yosinski, J.; Clune, J.; Nguyen, A.M.; Fuchs, T.J.; Lipson, H. Understanding Neural Networks Through Deep Visualization. *arXiv* **2015**, arXiv:1506.06579.
29. Bau, D.; Zhou, B.; Khosla, A.; Oliva, A.; Torralba, A. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3319–3327. [CrossRef]
30. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems. *arXiv* **2015**, arXiv:1506.01497.
31. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object Detection via Region-based Fully Convolutional Networks. Advances in Neural Information Processing Systems 29. *arXiv* **2016**, arXiv:1605.06409.
32. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 843–852. [CrossRef]