



Article

# Toward Semi-Supervised Graphical Object Detection in Document Images

Goutham Kallempudi <sup>1</sup>, Khurram Azeem Hashmi <sup>1,2,3,\*</sup> , Alain Pagani <sup>3</sup>, Marcus Liwicki <sup>4</sup>, Didier Stricker <sup>1,3</sup> and Muhammad Zeshan Afzal <sup>1,2,3</sup> 

<sup>1</sup> Department of Computer Science, Technical University of Kaiserslautern, 67663 Kaiserslautern, Germany; kallempu@rhrk.uni-kl.de (G.K.); didier.stricker@dfki.de (D.S.); muhammad\_zeshan.afzal@dfki.de (M.Z.A.)

<sup>2</sup> Mindgarage, Technical University of Kaiserslautern, 67663 Kaiserslautern, Germany

<sup>3</sup> German Research Institute for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany; alain.pagani@dfki.de

<sup>4</sup> Department of Computer Science, Luleå University of Technology, 97187 Lulea, Sweden; marcus.liwicki@ltu.se

\* Correspondence: khurram\_azeem.hashmi@dfki.de

**Abstract:** The graphical page object detection classifies and localizes objects such as Tables and Figures in a document. As deep learning techniques for object detection become increasingly successful, many supervised deep neural network-based methods have been introduced to recognize graphical objects in documents. However, these models necessitate a substantial amount of labeled data for the training process. This paper presents an end-to-end semi-supervised framework for graphical object detection in scanned document images to address this limitation. Our method is based on a recently proposed Soft Teacher mechanism that examines the effects of small percentage-labeled data on the classification and localization of graphical objects. On both the PubLayNet and the IIIT-AR-13K datasets, the proposed approach outperforms the supervised models by a significant margin in all labeling ratios (1%, 5%, and 10%). Furthermore, the 10% PubLayNet Soft Teacher model improves the average precision of Table, Figure, and List by +5.4, +1.2, and +3.2 points, respectively, with a similar total mAP as the Faster-RCNN baseline. Moreover, our model trained on 10% of IIIT-AR-13K labeled data beats the previous fully supervised method +4.5 points.

**Keywords:** graphical page objects; object detection; document image analysis; semi-supervised; soft teacher



**Citation:** Kallempudi, G.; Hashmi, K.A.; Pagani, A.; Liwicki, M.; Stricker, D.; Afzal, M.Z. Toward Semi-Supervised Graphical Object Detection in Document Images. *Future Internet* **2022**, *14*, 176. <https://doi.org/10.3390/fi14060176>

Academic Editor: Salvatore Carta

Received: 29 April 2022

Accepted: 5 June 2022

Published: 8 June 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Visual summary is an important aspect in a variety of applications, including summarizing the contents of a document and detecting graphical elements in the visualization pipeline. As a result, identifying and localizing graphical items will be an important step in document summary and analysis. With the increase in documents, it has become impractical to manually extract visual objects. Automated methods provide reliable, effective and efficient solutions for manual tasks. For instance, ref. [1] verified that the machine learning method performs better than humans in domain knowledge and attention-demanding tasks. Similarly, several automated methods [2–4] have been proposed to identify graphical objects, but these automated methods are typically rule-based, because the documents lack established dimension or structure [5].

The graphical page object detection aims at localizing and classifying multiple objects such as tables, images, and figures in a document. For instance, Figures 1 and 2 detect and localize objects in PubLayNet and IIIT-AR-13K datasets, respectively. As opposed to natural images and scenes, graphical objects have very little difference in their appearance; for example, Figure 1 (right) consists of the text block and a list of items block which appear to be similar, but it is necessary to classify them separately, making graphical page object detection more challenging.

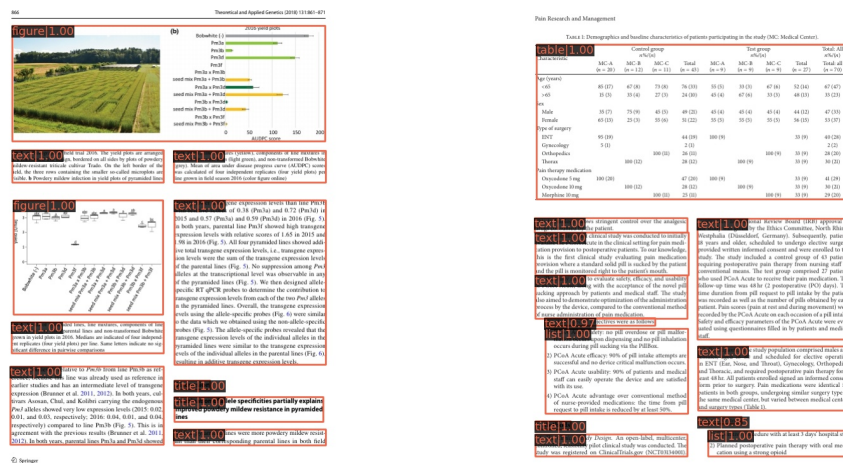


Figure 1. Left: graphical objects Figure, Text and Title. Right: graphical objects Table, List Text and Title. Both the images are generated by the Soft Teacher model, which is trained on 10% of the labeled PubLayNet data.



Figure 2. Left: graphical objects Table and Figure. Right: graphical objects Natural image and Signature. Both the images are generated by the the Soft Teacher model, which is trained on 5% of the labeled IIIT-AR-13K data.

Although defining rules for all graphical objects and running an optical character recognition tool helps localize objects, they cannot generalize to the new objects. This problem has resulted in the usage of deep learning models to detect graphical objects [6–9]. The deep learning models are rule independent and overcome the generalization problem with high precision. To train a deep learning model, a large amount of annotated data must be available, and as a result, either a manual labeling process or another pre-processing technique must be involved in generating high-quality data, which is both time-consuming and error-prone [10]. Owing to the concerns with training data, the problem changes its state from supervised to semi-supervised. This paper’s primary goal is to leverage the unlabeled data in identifying graphical objects without compromising the mean average precision and performance. We leverage the recently proposed Soft Teacher mechanism [10] to design a semi-supervised pipeline for graphical object detection on scanned document images.

The semi-supervised deep learning techniques in object detection [11–14] follow a predefined pipeline where an initial detector is trained using a small amount of available annotated data, and this detector is used to generate pseudo labels for unlabeled data. Finally, the labeled data along with unlabeled data with pseudo labels are used to retrain the model. The quality of these pseudo-labeling approaches depends on the initial model, and a weak initial model may degrade the pseudo-labeling process.

To alleviate the weak initial model problem, the end-to-end semi-supervised framework simultaneously performs pseudo labeling for unlabeled data and trains a detector using these pseudo labels with few annotated data in each iteration. The approach trains two models, one for detection known as Student and the other for pseudo labeling known as Teacher. In addition to this, the teacher model is simply an Exponential Moving Average (EMA) of the student model; this ensures that the pseudo-labeling process is constantly updated by the detection process and vice versa. Therefore, both the models reinforce each other.

The problem of GOD can be formulated as a generic object detection problem where we treat document images and graphical page objects, such as figures and tables, as natural objects. Therefore, inspired by the idea of [10], we leverage the use of the end-to-end semi-supervised framework on the graphical page object detection (GOD). Unlike the traditional pseudo-labeling approaches which are multi-staged but not end-to-end, the Soft Teacher ensures that a single round of iteration is complete from pseudo labeling to generating the loss by combining both labeled and pseudo-labeled data. Additionally, the framework provides a reinforcement effect, ensuring that the student model is always monitored by the teacher. Furthermore, this paper empirically shows that semi-supervised GOD methods can produce comparable results to fully labeled supervised GOD approaches.

The rest of the paper is structured as follows. Section 2 introduces multiple ways for object detection and categorizes them into rule-based, learning-based, deep learning and semi-supervised methods. The novel Soft Teacher framework is described in Section 3, which is followed by a dataset introduction in Section 4.1, evaluation protocol in Section 4.2, implementation details in Section 4.3 and results in Section 4.4. Finally, Section 5 concludes the paper with some future directions.

## 2. Related Work

There are numerous techniques for localizing objects in an image, ranging from conventional OCR rule-based systems to more current and accurate deep learning methods. Although each method aims to tackle the same problem, they confront a few issues with data, procedure, and performance. This section discusses the object detection methods, a few disadvantages, and a semi-supervised deep learning-based solution to overcome the problems.

### 2.1. Rule-Based Methods

Before the success of deep learning techniques, rule-based solutions for graphical object recognition began with the requirement to recognize tables and figures from the scanned text. There were various rule-based algorithms developed to localize the tables. They identify graphical objects in a document by using predefined rules; Ref. [2] uses the spectator mining technique to detect potential objects in a PDF document. This method assumes that white spaces or lines can differentiate a table cell.

To tackle the problems with varying table layouts, ref. [15] proposes a practical algorithm that can detect tables from various sources, including newspapers and company articles. Unlike most of the rule-based algorithms, which focus on detecting objects from structured documents, ref. [3] employs a correlation-based approach along with dynamic programming to identify tables in noisy handwritten documents. The fundamental problem of rule-based techniques is that they rely on predefined structure and content arrangement in documents.

## 2.2. Learning-Based Methods

The supervised learning models are the initial and most successful approaches to solving the graphical page object detection. With the popularity of machine learning in object detection, Ref. [16] used an SVM model to identify table cells in a document by creating 26 low-level features for each group of intersecting horizontal and vertical lines. The classifier then determines whether a cell belongs to a table or not.

In [17], the authors verify that the documents can be described by the MXY tree, which is a hierarchical representation. The method identifies tables by locating perpendicular lines and white spaces. Ref. [18] describes a probabilistic graphical model for document analysis. This model incorporates different document structures into Hidden Markov Models. Unsupervised learning algorithms are also proposed to cluster tables and text separately. The domain-independent technique proposed by [19] applies bottom-up clustering to word segments in recognizing tables. Although the learning-based models are accurate and independent compared to rule-based models, they do not generalize to new structures and need a large amount of training data.

## 2.3. Deep Learning Methods

With the success of Convolution Neural Networks (CNN) in the field of object detection and computer vision, the CNNs are also applied to predict and localize the documents; Ref. [20] proposes a method that first identifies table-like areas using a rule-based approach and applies a CNN on the output to detect potential tables. This method also considers non-visual features such as characters for better localization. To eliminate the dependency of rules and other metadata in training, Ref. [21] introduced DeepDeSRT, a two-stage approach for detection and structure detection, and later proved that this method generalizes to new structures. A Deep Convolution Neural Network (DCNN), which adds fully connected conditional random fields to convolution layers, performs multi-scale reasoning on visual cues for localizing objects. Hashmi et al. [22] made a few modifications to Cascade Mask R-CNN to identify mathematical formulas in document images.

The state-of-the-art object detection networks that depend on the regional proposals are introduced to the field of graphical object detection by [7]. They applied a Region Proposal Network (RPN), which shares full-image convolution features, enabling cost-free region proposals. There are many [23–25] region, pixel and connected component-based models proposed to identify textual and non-textual components in a document. The Faster-RCNN and Mask-RCNN detectors are used by [26] to localize and segment objects.

Based on a dynamic programming approach, [27] identifies region proposals in page object detection and [28] outlines and summarizes multiple deep learning approaches for graphical page object detection. The authors from [29] conduct performance analysis on neural networks which recognize tables in document images. The deep learning-based approaches are highly effective with some detectors trained on a particular dataset capable of detecting and localizing graphical structures on unseen new datasets [30]. The major disadvantage of the deep learning-based methods is the hunger for annotated data. These approaches require a lot of characterized data that is hard to obtain, and the annotation labeling process is manual, error-prone, and time-consuming.

## 2.4. Semi-Supervised Approaches

Due to the bottleneck in the labeling annotation process, the object detection techniques are leveraging unlabeled data via semi-supervised methods. Every day, numerous new documents are created that may or may not be related to current datasets and may represent new datasets. One of the early studies [31] employs semi-supervised learning on fully convolution networks to accomplish MS Lesion Segmentation. Attention-based semi-supervised deep networks proposed by [32] use region-attention to leverage unlabeled training data to segment medical images in an end-to-end fashion. Similarly, Ref. [33] exploits unlabeled endoscopic videos to learn representations of the target domain. The

role of semi-supervised deep learning is employed in various fields such as label propagation [34,35], anomaly detection [36], and segmentation [37,38].

In image classification, consistency-based semi-supervised methods improve classification performance by using all the unlabeled input. In these methods, a consistency requirement is used, with multiple modifications of the same image producing comparable results. Ref. [39] modifies the model, Ref. [40] proposes a new regularization method based on virtual adversarial loss, and [41] uses randomized image augmentations. Self-training approaches also known as pseudo-labeling approaches such as [42,43] train an initial classification model on unlabeled data to generate pseudo labels that help refine the initial classifier.

Semi-supervised object detection algorithms, such as image classification, can be consistency [39,44] or pseudo label based [11,45,46]. Some of the semi-supervised approaches are proved to perform better object detection than supervised approaches on the MS COCO image dataset [11,12,14,47,48]. These approaches are multi-staged, where an initial detector is applied to pre-labeled data before generating pseudo labels for the unlabeled data. Finally, the combination of annotated and non-annotated data is trained for accuracy. To improve the performance of these multi-stage detectors, an end-to-end pseudo label approach is proposed by [10], which simultaneously proposes pseudo boxes and performs detection training. The application of this model on the MS COCO dataset to detect and localize images showed significant improvement from the previous supervised and semi-supervised approaches. Hence, the primary objective of this paper is to leverage this novel end-to-end semi-supervised framework on graphical page object detection.

### 3. Proposed Framework

Semi-supervised object detection algorithms are broadly divided into consistency and pseudo labeled. The pseudo-labeled algorithms are multi-staged and run in two stages where an initial classifier is generated using the labeled data in the first stage, In the next stage, the initial classifier is used to create pseudo labels before updating the initial classifier with the combination of labeled and pseudo-labeled data. The Soft Teacher [10] framework, which is proven to produce supervised-like results in object detection is being used to prove the same for graphical page object detection in a pseudo-labeling framework with few modifications to the training process. Firstly, the framework is made end-to-end, without any initial classifier, and then, weak augmentation is used to generate pseudo labels. To address the issue of low training data, the strongly augmented data along with labeled data are used for detection training.

The framework comprises two models: the “student” model, which is in charge of detection training, and the “teacher” model creates pseudo boxes for unlabeled input. The teacher model is the student model’s exponential moving average (EMA) and learns to generate pseudo labels on weakly augmented unlabeled data. In contrast, the student model is trained on both labeled and strongly augmented unlabeled data to minimize the loss. The teacher model generates two sets of pseudo labels: one for the classification task and the other for detecting bounding boxes. We employ the FixMatch technique [49] of using data augmentation for training two multiple branches.

During the training process shown in Figure 3, the training data are split into different batches, each of which contains a random selection of labeled and unlabeled data. The teacher model creates pseudo labels from weakly augmented unlabeled data, while the student model uses both labeled and strongly augmented data. The pseudo labels generated by the teacher model are exploited as ground truth for strongly augmented unlabeled data to accomplish detection training. In contrast to the traditional classification problem, which determines whether a specific image corresponds to a specific label, the localization problem should identify and localize several entities in the same image. Identifying a particular object can be treated as a classification task, and the localization branch can be handled

as a regression task. The framework’s total loss is the weighted sum of supervised and unsupervised losses

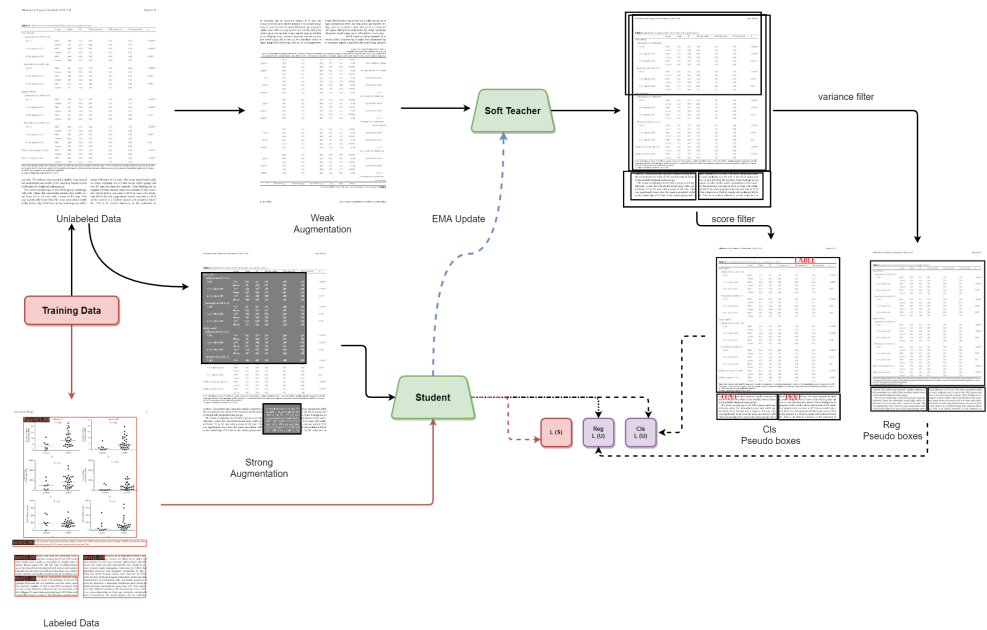
$$L_{total} = L_{sup} + \alpha L_{unsup} \tag{1}$$

Here,  $L_{sup}$  and  $L_{unsup}$  represent the supervised and unsupervised loss, respectively, and if  $N_{la}$  and  $N_{un}$  represent the number of labeled and unlabeled images in a batch with  $I_{la}^i$  and  $I_{un}^i$  representing a particular labeled and unlabeled images, respectively, the supervised and unsupervised loss can be defined as below

$$L_{sup} = \frac{1}{N_{la}} \sum_{i=1}^{N_{la}} (L_{cls}(I_{la}^i) + L_{reg}(I_{la}^i)) \tag{2}$$

$$L_{unsup} = \frac{1}{N_{un}} \sum_{i=1}^{N_{un}} (L_{unsup}^{cls}(I_{un}^i) + L_{unsup}^{reg}(I_{un}^i)) \tag{3}$$

Both the teacher and student model are randomly initialized initially, and during the training, the teacher model, in addition to providing pseudo labels, guides the student model. This effect occurs because the teacher model is a simple EMA of the student. During the training of the teacher model, multiple pseudo boxes are generated, and the Non-Maxima Suppression technique is leveraged to discard a few pseudo boxes. However, there can still be multiple pseudo boxes after the suppression technique. To segregate the available boxes into foreground and background, a threshold is employed, and the predictions with box scores higher than the threshold are used as foreground boxes in training.



**Figure 3.** The image represents the Soft Teacher process: (1) The training data are divided into labeled and unlabeled data. (2) Weak and strong augmentation are applied to unlabeled data. (3) The Teacher model uses the weakly augmented data to generate two sets of pseudo boxes: one for classification and the other for bounding box regression. (4) The Student model takes both the labeled and strongly augmented data with pseudo labels to perform detection training. (5) The Teacher model is an EMA of the Student model.

### 3.1. Limitations

Although the above base framework is capable of handling unlabeled data for training, there are two different problems. The first problem is due to the high threshold value for identifying the foreground boxes. Setting a high threshold value will result in higher precision, but the recall drops. Hence, if an IoU is used between the student and teacher-

generated boxes for label assignment, most of the foreground boxes will be assigned as background. Another problem is that the localization accuracy and foreground scores are not strongly co-related. Therefore, the localization is inaccurate. These problems are solved in [10] by introducing two techniques, namely the Soft Teacher and Box Jittering, which handle the classification and regression losses of the unsupervised branch.

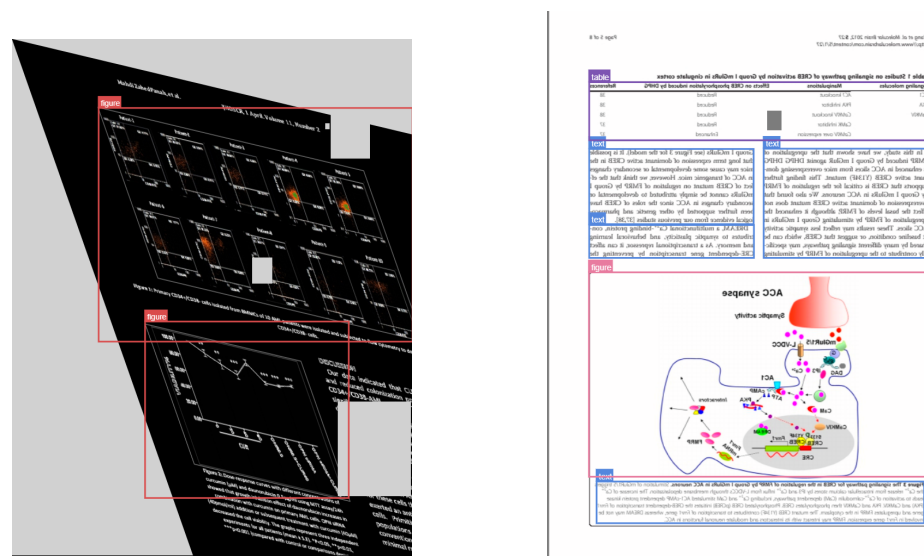
### 3.2. Soft Teacher

The Teacher model and pseudo boxes are leveraged to reduce the wrong assignment of foreground boxes to background boxes. The main idea of the Soft Teacher is to assess the reliability of each Student-generated box to be a real background, which is then used to calculate the background loss. Let  $b_i^{fg}$  and  $b_i^{bg}$  denote the set of foreground and background boxes; the unsupervised classification loss can be calculated as the following

$$L_{unsup}^{cls} = \frac{1}{N_{fg}} \sum_{i=1}^{N_{fg}} L_{cls}(b_i^{fg}, G_{cls}) + \sum_{j=1}^{N_{bg}} w_j L_{cls}(b_j^{bg}, G_{cls}) \tag{4}$$

$$w_j = \frac{r_j}{\sum_{k=1}^{N_{bg}} r_k} \tag{5}$$

The  $N_{fg}$  and  $N_{bg}$  denote the number of images in  $b_i^{fg}$  and  $b_i^{bg}$ ,  $G_{cls}$  denotes all the pseudo labels generated by the teacher model. The  $L_{cls}$  is the classification loss and  $r_j$  denotes the reliability for  $j$ -th background box. Ref. [10] proved that the simple background score produced by the teacher model serves as the reliability score, and Figure 4 shows the classification bounding boxes that are generated during the training process.



**Figure 4.** Left: Augmentation with scale, shift, and color change. Right: Augmentation with rotation. The images represent the classification pseudo boxes on two different augmented images. They are generated by the teacher model during the training process of the Soft Teacher model on 10% labeled data.

### 3.3. Box Jittering

Due to a high threshold in identifying foreground boxes, the boxes generated are not accurate. In order to alleviate the issue, ref. [10] proposes a box-jittering approach in which

a random box  $b_i$  is identified across the real foreground box and is refined with the teacher model multiple times  $N_{jitter}$  to generate more accurate box coordinates  $b_i^*$

$$b_i^* = refine_{jitter}(b_i) \tag{6}$$

Hence,  $N_{jitter}$  number of new box coordinate sets  $\{b_{i,j}^*\}$  are generated. For each of the boxes generated, a regression variance is calculated to identify the regression box with high localization accuracy. The box regression variance is defined by

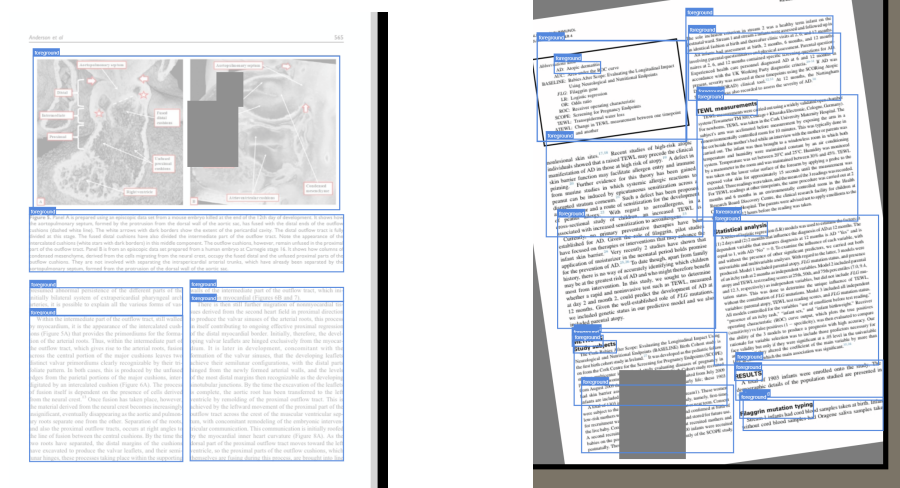
$$\bar{\sigma}_i = \frac{1}{4} \sum_{k=1}^k \sigma_k^* \tag{7}$$

$$\sigma_k^* = \frac{\sigma_k}{0.5 * (h(b_t) + w(b_t))} \tag{8}$$

Here,  $\sigma_k$  denotes the standard deviation of the  $k$ -th coordinate of the refined box coordinate set  $\{b_{i,j}^*\}$ ,  $\sigma_k^*$  is the normalized standard deviation and  $h$  and  $w$  are the height and width of the  $t$ -th refined box  $b_t$ . A smaller box regression variance  $\bar{\sigma}_i$  indicates better localization, but calculating the box regression variance for all the foreground boxes is highly time and resource-consuming. Hence, it is proposed to calculate it for foreground boxes with a threshold greater than 0.5. Finally, if  $G_{reg}$  are the coordinates of all the regression pseudo labels generated by the teacher model, the unsupervised regression loss can be calculated using

$$L_{unsup}^{reg} = \frac{1}{N_{fg}} \sum_{i=1}^{N_{fg}} L_{reg}(b_i^{fg}, G_{reg}) \tag{9}$$

Figure 5 shows the regression bounding boxes that are generated after the box jittering approach in the training process.



**Figure 5.** Left: Augmentation with rotation and blur. Right: Augmentation with scale and shift. The images represent the regression pseudo boxes on two different augmented images. They are generated by the Teacher model during the training process of the Soft Teacher model on 10% labeled data.

## 4. Experiments

### 4.1. Datasets

The experiments are performed on three datasets: namely, PubLayNet [26], IIT-AR-13K [50] and DocBank [51], and the following subsections describe the datasets in detail.



#### 4.1.1. Publaynet

The PubLayNet dataset is created by matching the XML representations and content of over 1 million publicly available documents and is primarily used for training the model. This dataset consists of 335,703 training images and 11,245 validation images for training and evaluation. Table 1 shows the distribution of labels in the training dataset. In addition to this, a new subset dataset termed sub-PubLayNet is created, which consists of two labels: Figure and Table. These data are used to cross-validate a PubLayNet trained model on the sub-DocBank and sub-IIIT-AR-13K datasets.

**Table 1.** Label-wise summary of PubLayNet training dataset.

PubLayNet	Number of Images	Number of Annotations
Table	86,460	102,514
Figure	91,968	109,292
Text	334,548	2,343,356
Title	255,731	627,125
List	53,049	80,759
Total	335,703	3,263,046

#### 4.1.2. IIIT-AR-13K

The IIIT-AR-13K consists of a set of business documents, mostly annual reports. The dataset contains 9333 training images and 1955 validation images. Table 2 represents the class-wise training dataset distribution. Similar to the PubLayNet, a subset (sub-IIIT-AR-13K) dataset is created with two labels: Figure and Table. This subset dataset is used to train a Soft Teacher model and to validate the model on sub-PubLayNet and sub-DocBank datasets.

**Table 2.** Label-wise summary of IIIT-AR-13K training dataset.

IIIT-AR-13K	Number of Images	Number of Annotations
Table	6903	11,163
Figure	1293	2004
Natural Image	1258	1987
Logo	165	379
Signature	208	420
Total	9333	15,953

#### 4.1.3. Docbank

The DocBank dataset consists of 16 classes. To understand the cross-validated performance of the DocBank trained model on other datasets, a new subset (sub-DocBank) is created, which contains the labels Table and Figure. Tables 3 and 4 show the total image and annotation count in the subset datasets sub-PubLayNet, sub-IIIT-AR-13K, and sub-DocBank, respectively.

**Table 3.** Actual datasets are used to create subset datasets that only consist of the Table and Figure annotations. The table represents the distribution of the Table and Figure images in the training datasets of sub-PubLayNet, sub-IIIT-AR-13K and sub-DOCBANK.

Dataset	Table	Figure	Total
sub-PubLayNet	86,460	91,968	102,514
sub-IIIT-AR-13K	6903	1293	7837
sub-DocBank	19,528	89,612	103,285

**Table 4.** The table represents the distribution of the Table and Figure annotations in the three subset training datasets.

Dataset	Tables	Figures	Total
sub-PubLayNet	102,514	109,292	211,806
sub-IIIT-AR-13K	2222	481	2703
sub-DocBank	25,991	128,312	154,303

#### 4.2. Evaluation Protocol

The partially labeled data creation follows the STAC [52] setting, where the training data are divided into 1%, 5% and 10%, respectively, and these samples are considered annotated data in these three training settings, respectively. The rest of the data in each of the setting are used as unlabeled data in the training process. For each protocol, STAC provides five different folds, and the final performance is the average of all the five folds. To compare the supervised training models with this semi-supervised framework, the unlabeled data are treated as not useful for the training of supervised models. The mean Average Precision (mAP) on the validation data is evaluated for the comparison of various semi-supervised models.

##### 4.2.1. Precision

The Precision [53] is the fraction of relevant instances (True Positives) among the retrieved instances (True Positives + False Positives).

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (10)$$

##### 4.2.2. Recall

The Recall [53] is the fraction of relevant instances (True Positives) that were retrieved (True Positives + False Negatives).

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (11)$$

##### 4.2.3. F1-Score

The F1-score [53] is the harmonic mean between the Precision and Recall. It is mathematically defined as follows

$$F1\text{-Score} = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \quad (12)$$

##### 4.2.4. Intersection over Union (IoU)

The Intersection over Union (IoU) [22], which is also known as the Jaccard index, measures the similarity between finite sample sets. It is defined as the ratio between the

size of the intersection and the size of the union of the two sample sets. In the machine learning setting, it predicts the region between the predicted and ground truth region.

$$\text{IoU}(A,B) = \frac{\text{Area of Overlap region}}{\text{Area of Union region}} = \frac{|A \cap B|}{|A \cup B|} \quad (13)$$

#### 4.2.5. Average Precision (Ap)

Average Precision is defined as the area under the Precision–Recall curve. In the context of object localization, the average precision is defined for multiple IoUs. For instance, the AP@[0.5] indicates the area under the Precision–Recall curve when the IoU threshold is set to 0.5. In this case, an element is correctly classified if it overlaps with 50% of the ground truth. In the context of MS COCO [48] evaluation, the Average Precision of the classes is averaged to get the final Average Precision.

#### 4.2.6. Mean Average Precision (Map)

The mean Average Precision (mAP) is calculated by taking the mean of all classes average precision over all the IoU thresholds defined. For MS COCO [48] evaluation, the mAP is averaged over 10 IoU thresholds from 0.50 to 0.95 with a step size of 0.5.

### 4.3. Implementation Details

For the PubLayNet Partial data (i.e., based on the percentage of annotated data) and the PubLayNet intersection data (i.e., based on the percentage of annotated data for Table and Figure), the implementation considers Faster R-CNN [7] equipped with FPN [54] as the default detection framework for the evaluation. Along with the ImageNet pre-trained ResNet-50 [55] and ResNet-101 [55] backbones on 1%, 5% and 10%, a Swin-T [56] model on 5% and 10% of annotated data is also trained for the comparison. Anchors with five scales and three aspect ratios are used, and 2000 and 1000 region proposals are generated with a non-maximum suppression threshold of 0.7 on both training and inference. Finally, in each training step, 512 proposals are samples from 2000 proposals as box candidates to train RCNN.

All the models are trained for 150,000 iterations on two GPUs with eight images per GPU batch size. The foreground threshold is set to 0.9, and the data sampling ratio, which is the ratio of annotated to non-annotated images in each batch, is set to 0.2 and gradually decreases to 0 in the last 10,000 iterations. For the stochastic gradient descent, the learning rate is set to 0.01 and is divided by ten at 110,000 iterations. To identify the pseudo labels for bounding boxes with high localization reliability, the  $N_{jiter}$  is set to 10 with a threshold of 0.02. In addition, the same augmentation techniques as in [52] are considered for training.

The hyper-parameters foreground threshold, suppression threshold, and  $N_{jiter}$  are proved to be optimal by [10] while training the Soft Teacher model on the MS-COCO object detection dataset [48]. Finally, the IIIT-AR-13K is trained only for 50,000 iterations with Faster-RCNN equipped with FPN using ImageNet pre-trained ResNet-50 as the backbone. The learning rate is set to 0.01, and other parameters are similar to the PublayNet trained models, since the IIIT-AR-13K data in all the splits are small (less than 1100 images).

### 4.4. Results and Discussion

#### 4.4.1. Publaynet

In this section, the results of multiple models trained on the PubLayNet dataset with the proposed framework are compared with the models trained using supervised techniques. The partially annotated PubLayNet models are first compared with the corresponding supervised models. Table 5 shows the comparison of the Faster-RCNN supervised models on the ResNet-50 backbone with different versions of Soft Teacher that are trained using Faster-RCNN and Swin-T transformers. For the detailed comparison, the Faster-RCNN is further trained using two backbones, i.e., ResNet50 and ResNet101. The semi-supervised models showed around 4.9 points, 5.8 points and 5.5 points improvement

in the bounding box mean Average Precision when compared under 1%, 5% and 10% of annotated data, respectively.

Further, a Soft Teacher (Faster-RCNN + ResNet101) model which is trained on 10% of labeled data for 180,000 iterations with a batch size of 8 is compared with the existing baseline model by [26]. The baseline models are trained using fully labeled PubLayNet data on Faster-RCNN for 180,000 iterations. The results in Table 6 show that the Soft Teacher model outperforms the supervised model's average precision at (IoU = 0.5) by 2.7% and exhibits a similar mean Average Precision of around 90.0%. Finally, the inference times in terms of FPS of our models are shown in Table 5. Although there is no reference FPS from the earlier methods, we tried to compare inference time with our models and found that our model equipped with Faster R-CNN and ResNet50 outperformed ResNet101 and Swin-T.

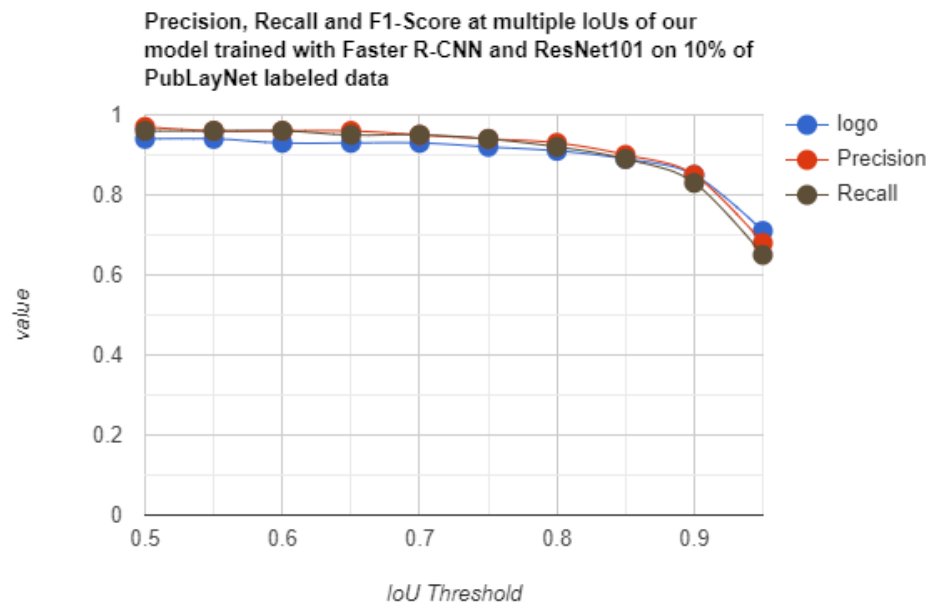
**Table 5.** Performance mAP (0.5:0.95) comparison of supervised and ours models trained on 1%, 5% and 10% of PubLayNet dataset.

Technique	Detector + Backbone	1%	5%	10%	FPS
Supervised	Faster R-CNN + ResNet50	82.5	83.3	83.4	12.3
Ours	Faster R-CNN + ResNet50	84.9	87.2	87.3	16.7
	Faster R-CNN + ResNet101	87.4	88.2	88.9	15.8
	Swin-T	88.3	89.1	88.6	14.9

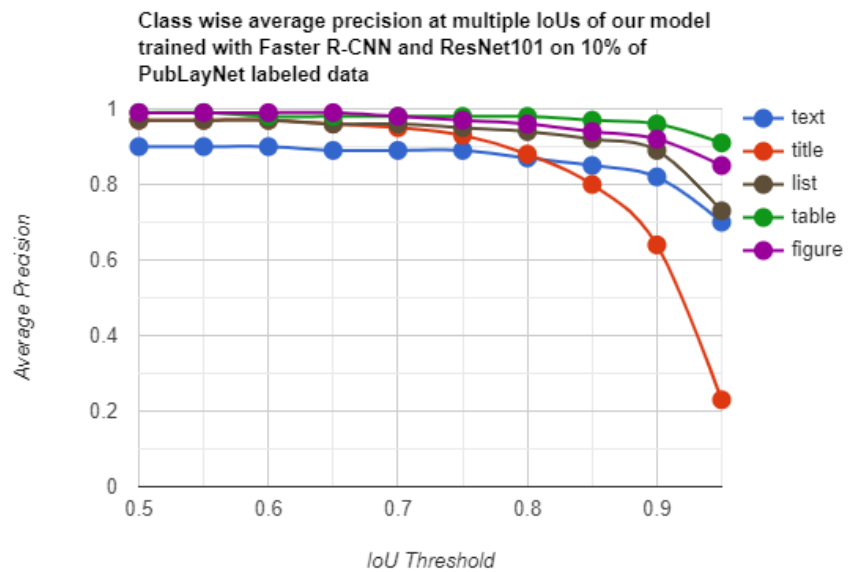
**Table 6.** Performance comparison of Faster R-CNN baseline model trained on fully labeled data with our semi-supervised model which is trained on 10% labeled PubLayNet dataset. The model uses Faster R-CNN and ResNet101.

Technique	Detector	AP@0.50	AP@0.75	mAP	FPS
Zhong et al. [26]	Supervised + Faster R-CNN	93.7	91.1	90.2	-
Ours	Faster R-CNN (10% labeled data)	96.4	93.8	90.0	15.8

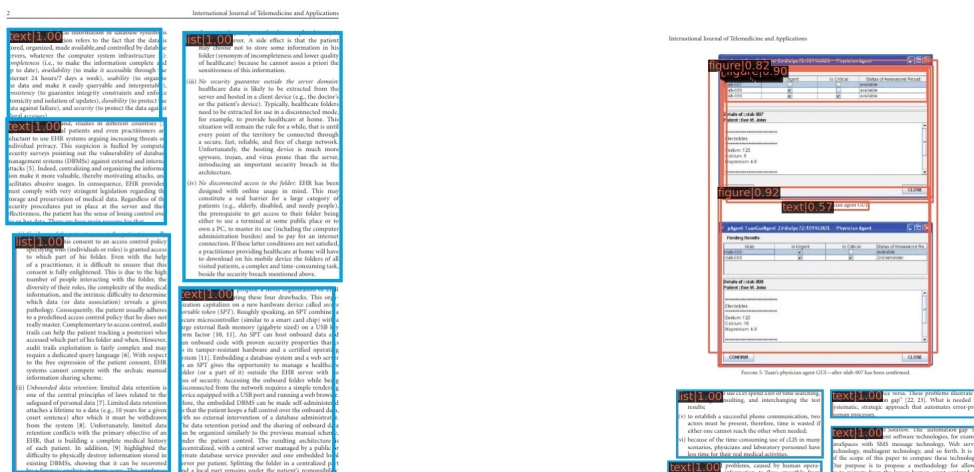
In addition to the total mAP, the label-wise AP comparison is shown in Table 7. The graphical objects Table, Figure and List outperform the baseline model by 5.4%, 1.2% and 3.2%, respectively. Moreover, Table 8 as well as Figures 6 and 7 explain and show the detailed metrics of Precision, Recall and F-1 score at various IoU thresholds. The results for the Soft Teacher model trained on 10% labeled PubLayNet data are shown in Figure 1. Additionally, Figure 8 portrays the qualitative results, and Figure 8 (right) shows the model figure and table resulting in False Positives.



**Figure 6.** A visual representation of Precision, Recall and F1-Score at multiple IoUs. Our model is trained on 10% PubLayNet labeled data with Faster R-CNN and ResNet101.



**Figure 7.** A label-wise visual representation of Precision at multiple IoUs. Our model is trained on 10% PubLayNet labeled data with Faster R-CNN and ResNet101.



**Figure 8.** The results of our model with Faster R-CNN and ResNet101 that is trained on 10% of PubLayNet labeled data. Blue represents True Positives and Red represents False Positives. In this figure, (left) represents a couple of samples with True Positives, (right) depicts True Positives and False Positives.

**Table 7.** Class-wise performance AP (0.5:0.95) comparison of Faster R-CNN baseline model trained on fully labeled data with our semi-supervised model which is trained on 10% labeled PubLayNet dataset. The model uses Faster R-CNN and ResNet101.

Technique	Detector	Table	Figure	Text	Title	List
Zhong et al. [26]	Supervised + Faster R-CNN	90.2	93.7	91.0	82.6	88.3
Ours	Faster R-CNN (10% labeled data)	96.6	94.9	85.5	81.4	91.5

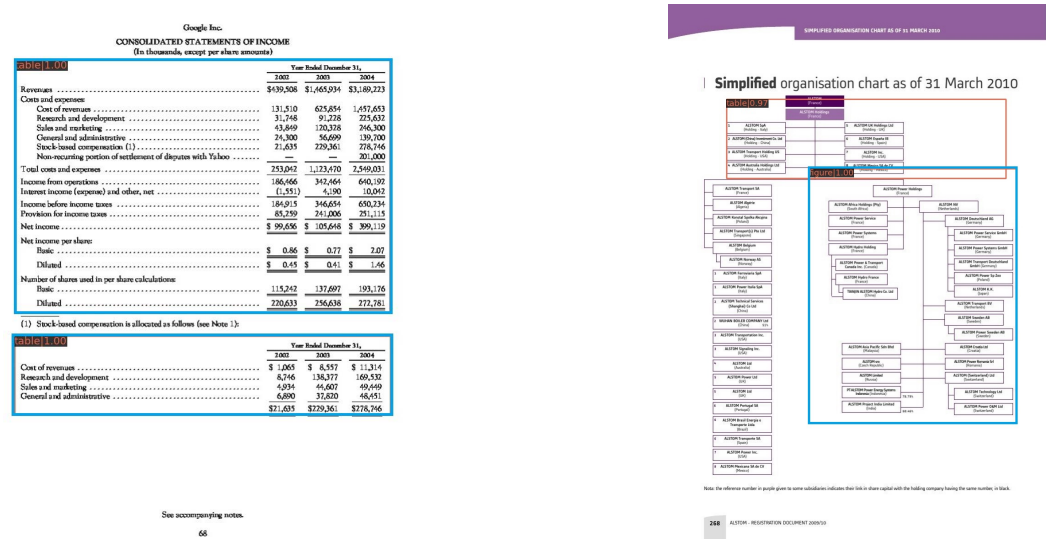
**Table 8.** A detailed representation of Precision, Recall and F1-Score at multiple IoUs. Our model is trained on 10% PubLayNet labeled data with Faster R-CNN and ResNet101.

Method	IoU	Precision	Recall	F1-Score
Semi-Supervised Soft Teacher Faster R-CNN + ResNet101 10% labeled	0.50	0.97	0.96	0.94
	0.60	0.96	0.96	0.93
	0.70	0.95	0.95	0.93
	0.80	0.93	0.92	0.91
	0.90	0.85	0.85	0.83

4.4.2. IIIT-AR-13K

The Supervised Faster RCNN and Soft Teacher models trained on IIIT-AR-13K with ResNet50 backbone are compared under three different scenarios of labeled data: 1%, 5%, and 10%. Table 9 shows that the proposed framework performs better than the Faster RCNN model by 6.4%, 1% and 5.1%, respectively. Additionally, a YOLO F [57] model trained on the complete data is compared with the Soft Teacher model that is trained on 10% labeled data. The Soft Teacher model showed a 3.7% mAP improvement, and it outperforms the YOLO F model in detecting natural images, logos, and signatures. Tables 10 and 11 shows the detailed comparison of YOLO F and Soft Teacher models along with the inference times. Figure 2 show the qualitative results of two IIIT-AR-13K, and Figure 9 displays a few samples of True Positives and False Positives documents trained using the Soft Teacher

model. Finally, Table 12 and Figures 10 and 11 visually explain the Precision, Recall and F1 scores of a soft teacher model trained on 10% of IIIT-AR-13K labeled dataset.



**Figure 9.** The results of our model with Faster R-CNN and ResNet50 that is trained on 10% of IIIT-AR-13K labeled data. Blue represents True Positives and Red represents False Positives. In this figure, (left) represents a couple of samples with True Positives, (right) depicts True Positive and False Positive.

**Table 9.** Performance comparison in terms of mAP (0.5:0.95) between supervised and our semi-supervised model trained on 1%, 5% and 10% of the IIIT-AR-13K dataset. Both the models use Faster-RCNN with ResNet 50 backbone.

Technique	Detector + Baseline	1%	5%	10%	FPS
Supervised	Faster R-CNN + ResNet50	35.8	49.7	57.4	12.3
Ours	Faster R-CNN + ResNet50	42.2	51.8	63.3	16.8

**Table 10.** Performance comparison of YOLO F baseline model trained on fully labeled data with our model which is equipped with Faster R-CNN and ResNet50 and is trained on 10% labeled IIIT-AR-13K dataset.

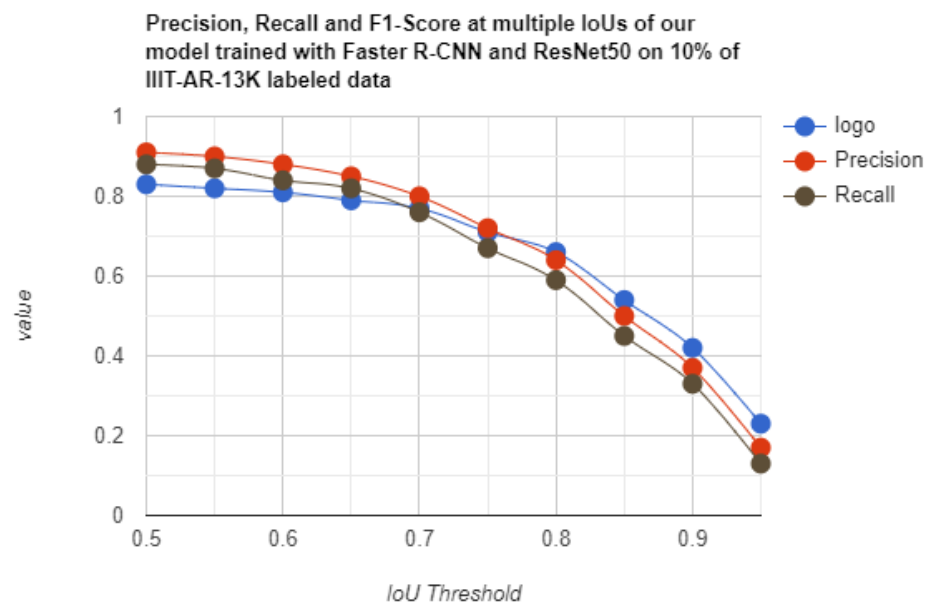
Technique	Detector	AP@0.50	AP@0.75	mAP	FPS
Nguyen et al. [57]	Supervised + YOLO F	81.2	64.9	58.8	-
Ours	Faster R-CNN (10% labeled data)	87.8	67.0	63.3	16.8

**Table 11.** Class-wise performance AP (0.5:0.95) comparison of YOLO F baseline model trained using fully labeled data with our model, which is equipped with Faster-RCNN and ResNet50 and is trained on 10% labeled data.

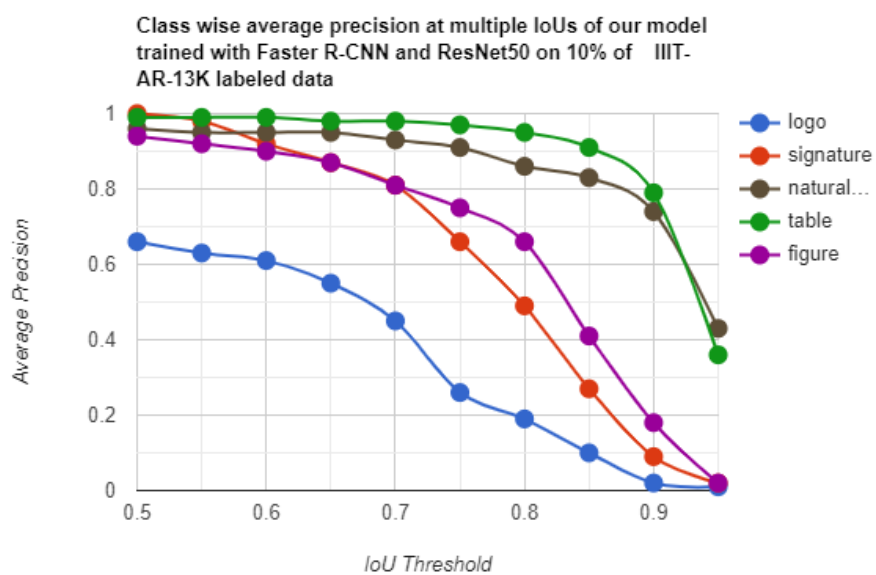
Technique	Detector	Table	Figure	Natural Image	Logo	Signature
Nguyen et al. [57]	YOLO F	88.3	63.7	73.0	18.3	50.6
Ours	Faster R-CNN (10% labeled)	87.1	60.7	82.5	28.2	58.0

**Table 12.** A detailed representation of Precision, Recall and F1-Score at multiple IoUs. Our model is trained on 10% IIIT-AR-13K labeled data with Faster R-CNN and ResNet50 backbone.

Method	IoU	Precision	Recall	F1-Score
Semi-Supervised Soft Teacher Faster R-CNN + ResNet50 10% labeled	0.50	0.91	0.88	0.83
	0.60	0.88	0.84	0.81
	0.70	0.80	0.76	0.77
	0.80	0.64	0.59	0.66
	0.90	0.37	0.33	0.42



**Figure 10.** A visual representation of Precision, Recall and F1-Score at multiple IoUs. The model is trained on 10% IIIT-AR-13K labeled data and is equipped with Faster R-CNN and ResNet50.



**Figure 11.** A class-wise visual representation of precision at multiple IoUs. The model is trained on 10% IIIT-AR-13K labeled data and is equipped with Faster R-CNN and ResNet50.



#### 4.4.3. Cross-Validation

The datasets PubLayNet, IIIT-AR-13K, and DocBank are the most common datasets used for graphical object detection, and the three datasets possess objects which are different from each other. To validate a model trained on one dataset with another dataset, there is a need to train models with the objects common to all the datasets. Table and Figure are the common classes among datasets, and they are different from each other; for instance, the IIIT-AR-13K dataset is based on business documents, the PubLayNet is created from publicly available docs, and DocBank is created in a way to ease layout analysis. The primary idea is to identify how similar the datasets are in terms of the graphical objects. The new datasets sub-PubLayNet, sub-IIIT-AR-13K and sub-DocBank that contain Table and Figure annotations are created, and multiple Soft Teacher models are trained with 5 and 10% of sub-PubLayNet, sub-IIIT-AR-13K, and sub-DocBank.

The results of models trained on sub-PubLayNet and cross-validated on sub-IIIT-AR-13K and sub-DocBank are shown in Table 13. The Soft Teacher model trained on 10% of labeled trained data on Faster-RCNN with ResNet101 showed the highest mAP of 95.1%. The validation performed on other datasets proves that sub-DocBank is more similar to sub-PubLayNet than the sub-IIIT-AR-13K. Table 14 also proves that sub-DocBank is similar to sub-PubLayNet, and interestingly, the mAP on sub-PubLayNet is higher than that of sub-DocBank, even though the model is trained completely trained on sub-Docbank. From Table 15, it is shown the models trained on IIIT-AR-13K perform badly on the other two datasets. This is because the dataset has mostly business reports, unlike the other two datasets. Figure 12 depicts the qualitative evaluation of the table and figure on the sub-PubLayNet dataset by displaying samples of True Positives and False Positives.

**Table 13.** The table represents the mAP@(0.5:0.95) of sub-PubLayNet, sub-IIIT-AR-13K and sub-DocBank validation datasets on a sub-PubLayNet trained model.

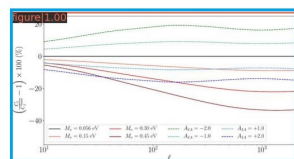
Methods	Labeled Data	Sub-PubLayNet			Sub-IIIT-AR-13K			Sub-DocBank		
		Table	Figure	mAP	Table	Figure	mAP	Table	Figure	mAP
ST_FRCNN_R50	5%	95.9	92.1	94.0	59.7	23.0	41.4	72.7	49.0	60.8
	10%	95.9	92.2	94.0	60.2	24.2	42.2	72.3	49.4	60.8
ST_FRCNN_R101	5%	96.0	94.2	94.5	61.1	22.6	41.9	71.1	48.0	59.6
	10%	96.1	92.9	95.1	57.3	16.6	37.0	66.2	48.0	57.1

**Table 14.** The table represents the mAP@(0.5:0.95) of sub-DocBank, sub-IIIT-AR-13K and sub-PubLayNet validation datasets on a sub-DocBank trained model.

Methods	Labeled Data	Sub-DocBank			Sub-PubLayNet			Sub-IIIT-AR-13K		
		Table	Figure	mAP	Table	Figure	mAP	Table	Figure	mAP
ST_FRCNN_R50	5%	82.4	64.5	73.4	88.5	71.2	79.8	52.3	14.3	33.3
	10%	82.5	65.3	73.9	89.0	72.0	80.5	60.6	28.0	40.3
ST_FRCNN_R101	5%	80.7	65.3	73.0	84.9	65.1	75.0	55.5	18.2	36.9
	10%	80.4	68.5	74.4	87.1	65.0	76.0	55.4	17.9	36.7

**Table 15.** The table represents the mAP@(0.5:0.95) of sub-IIIT-AR-13K, sub-PubLayNet and sub-DocBank validation datasets on a sub-IIIT-AR-13K trained model.

Methods	Labeled Data	Sub-IIIT-AR-13K			Sub-PubLayNet			Sub-DocBank		
		Table	Figure	mAP	Table	Figure	mAP	Table	Figure	mAP
ST_FRCNN_R50	5%	86.3	51.5	68.9	73.1	25.8	49.4	33.6	20.5	27.0
	10%	88.0	54.8	71.4	75.7	31.0	53.3	54.3	29.0	41.7
ST_FRCNN_R101	5%	87.3	45.8	66.5	75.4	31.3	53.4	60.4	31.3	45.8
	10%	85.5	48.3	67.4	78.3	21.2	49.7	56.0	23.0	39.5



**Figure 6.** Percentage difference on the shear power spectrum  $C_s$  due to an increasing intrinsic mass (from left to right) or an increasing intrinsic alignment (from green to yellow) with respect to a model with minimal intrinsic mass and no intrinsic alignment. We assume here a single tomographic bin.

with

$$N_s^{(2)}(z) = n_s^{(2)}(z) \frac{d\ln}{dz} = n_s^{(2)}(z) \frac{H(z)}{z} \quad (4.5)$$

$$F_{11}(k, z) = -A_1 C_1 \rho_c \frac{D_m(z)}{D_m(z_0)} \quad (4.6)$$

where  $\mu_c$  and  $D_m$  are the critical density and the matter density parameter today,  $D_m(k, z)$  is the linear growth factor, scale-dependent for massive neutrino cosmologies, while  $C_1 = 5 \times 10^{-10} h^2 M_{\odot}^2 \text{Mpc}^3$  is a normalisation constant chosen such that the intrinsic alignment free parameter  $A_1$  takes values around unity. For instance, Ref. [13] found  $A_1 = -1.8^{+1.2}_{-1.5}$  and  $A_1 = -1.7^{+1.2}_{-1.8}$  for the analyses using 3- $z$  and 2- $z$  bins respectively, while Ref. [2], although using another model, obtained  $A_1 = 1.3^{+0.4}_{-0.5}$ .

In Figure 6 we plot the relative difference on the shear power spectrum (we consider a single tomographic bin for simplicity) of models with different neutrino masses and models with different intrinsic alignment parameter with respect to a model with minimal neutrino mass and  $A_1 = 0$ . Intrinsic alignment can either enhance (if  $A_1 < 0$ ) or damp (if  $A_1 > 0$ ) the signal at all multipoles, and this effect may in principle mimic the neutrinos and introduce a possible degeneracy with  $M_\nu$ .

We perform the MCMC with the usual method but this time setting the  $z_1$  parameter to a fixed value of 2. We choose to do so because some of the runs of the previous analysis (Section 3) was able to constrain such parameter, due to the weak dependence of the shear spectra on it (see the bottom right point of Figure 2). Therefore, we will have again 4 free



**Figure 12.** The results of our model equipped with Faster R-CNN and ResNet101 and trained on 5% of sub-PubLayNet labeled data. Blue represents True Positives and Red represents False Positives. In this figure, (left) represents a sub-DocBank sample with a True Positive and False Positives, (right) shows True Positives and False Positives on a sub-IIIT-AR-13K sample.

### 5. Conclusions and Future Work

This paper presents the semi-supervised framework for graphical page object detection by employing a multi-stage semi-supervised technique known as Soft Teacher for the detection of graphical objects. In addition to operating on minimal data, this approach combines the complex process of pseudo labeling into a single pipeline. As the framework generates pseudo labels simultaneously with detection training, the flywheel effect occurs, which means one model constantly refines the pseudo boxes generated by the other model during the training process. This framework refines the classification and regression pseudo boxes using two different techniques, Soft Teacher and Box Jittering. These two processes work independently to render accurate classification and bounding box predictions.

This approach outperforms the supervised models in the labeling ratios (1%, 5%, and 10%) of PubLayNet and IIIT-AR-13K training data. Additionally, the Soft Teacher models trained on 10% PubLayNet labeled data performed similarly to the existing supervised baseline, whereas the Soft Teacher model on IIIT-AR-13K outperformed the YOLO F model. Finally, new datasets are created to perform inter-dataset cross-validation and prove that PubLayNet and DocBank datasets are more related compared to IIIT-AR-13K.

In future work, we plan to use powerful two-stage detectors such as Cascade-RCNN for better detection results compared to the current Faster-RCNN. Furthermore, we will examine the effect of the percentage of labeled data on the final performance and try to design robust models relying on minimal annotated data. Finally, the aforementioned

results show that the proposed framework with Faster R-CNN performs similarly to Faster R-CNN; nevertheless, we must show that this statement holds for multiple backbones. Finally, to establish a baseline for semi-supervised graphical page object detection, we want to assess various semi-supervised frameworks, including consistency and pseudo labeling.

**Author Contributions:** Writing—original draft preparation, G.K., K.A.H., M.Z.A.; writing—review and editing, K.A.H., M.Z.A., M.L.; supervision and project administration, A.P., D.S. All authors have read and agreed to the submitted version of the manuscript.

**Funding:** The work leading to this publication has been partially funded by the European project INFINITY under Grant Agreement ID 883293.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

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

## References

1. Orosz, T.; Vági, R.; Csányi, G.M.; Nagy, D.; Üveges, I.; Vadász, J.P.; Megyeri, A. Evaluating Human versus Machine Learning Performance in a LegalTech Problem. *Appl. Sci.* **2021**, *12*, 297. [[CrossRef](#)]
2. Fang, J.; Gao, L.; Bai, K.; Qiu, R.; Tao, X.; Tang, Z. A Table Detection Method for Multipage PDF Documents via Visual Separators and Tabular Structures. In Proceedings of the 2011 International Conference on Document Analysis and Recognition, ICDAR 2011, Beijing, China, 18–21 September 2011; IEEE Computer Society: Piscataway, NJ, USA, 2011; pp. 779–783. [[CrossRef](#)]
3. Chen, J.; Lopresti, D.P. Table Detection in Noisy Off-line Handwritten Documents. In Proceedings of the 2011 International Conference on Document Analysis and Recognition, ICDAR 2011, Beijing, China, 18–21 September 2011; IEEE Computer Society: Piscataway, NJ, USA, 2011; pp. 399–403. [[CrossRef](#)]
4. Hashmi, K.A.; Ponnappa, R.B.; Bukhari, S.S.; Jenckel, M.; Dengel, A. Feedback learning: Automating the process of correcting and completing the extracted information. In Proceedings of the 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), Sydney, Australia, 22–25 September 2019; IEEE: Piscataway, NJ, USA, 2019; Volume 5, pp. 116–121.
5. Saha, R.; Mondal, A.; Jawahar, C.V. Graphical Object Detection in Document Images. In Proceedings of the 2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, 20–25 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 51–58. [[CrossRef](#)]
6. Girshick, R.B. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 7–13 December 2015; IEEE Computer Society: Piscataway, NJ, USA, 2015; pp. 1440–1448.
7. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)] [[PubMed](#)]
8. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. *arXiv* **2016**, arXiv:1612.08242.
9. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017.
10. Xu, M.; Zhang, Z.; Hu, H.; Wang, J.; Wang, L.; Wei, F.; Bai, X.; Liu, Z. End-to-End Semi-Supervised Object Detection with Soft Teacher. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021.
11. Wang, K.; Yan, X.; Zhang, D.; Zhang, L.; Lin, L. Towards Human-Machine Cooperation: Self-supervised Sample Mining for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
12. Tang, P.; Ramaiah, C.; Xu, R.; Xiong, C. Proposal Learning for Semi-Supervised Object Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Montreal, QC, Canada, 11–17 October 2021.
13. Rhee, P.; Erdenee, E.; Kyun, S.D.; Ahmed, M.U.; Jin, S. Active and semi-supervised learning for object detection with imperfect data. *Cogn. Syst. Res.* **2017**, *45*, 109–123. [[CrossRef](#)]
14. Xie, Q.; Dai, Z.; Hovy, E.H.; Luong, T.; Le, Q. Unsupervised Data Augmentation for Consistency Training. In Proceedings of the Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Virtual, 6–12 December 2020. Available online: <https://proceedings.neurips.cc/paper/2020/file/44feb0096faa8326192570788b38c1d1-Paper.pdf> (accessed on 28 April 2022).

15. Shafait, F.; Smith, R. Table detection in heterogeneous documents. In Proceedings of the The Ninth IAPR International Workshop on Document Analysis Systems, DAS 2010, Boston, MA, USA, 9–11 June 2010; Doermann, D.S., Govindaraju, V., Lopresti, D.P., Natarajan, P., Eds.; ACM: New York, NY, USA, 2010; pp. 65–72. [[CrossRef](#)]
16. Kasar, T.; Barlas, P.; Adam, S.; Chatelain, C.; Paquet, T. Learning to Detect Tables in Scanned Document Images Using Line Information. In Proceedings of the 12th International Conference on Document Analysis and Recognition, ICDAR 2013, Washington, DC, USA, 25–28 August 2013; IEEE Computer Society: Piscataway, NJ, USA, 2013; pp. 1185–1189. [[CrossRef](#)]
17. Cesarini, F.; Marinai, S.; Sarti, L.; Soda, G. Trainable Table Location in Document Images. In Proceedings of the 16th International Conference on Pattern Recognition, ICPR 2002, Quebec, QC, Canada, 11–15 August 2002; IEEE Computer Society: Piscataway, NJ, USA, 2002; pp. 236–240. [[CrossRef](#)]
18. e Silva, A.C. Learning Rich Hidden Markov Models in Document Analysis: Table Location. In Proceedings of the 10th International Conference on Document Analysis and Recognition, ICDAR 2009, Barcelona, Spain, 26–29 July 2009; IEEE Computer Society: Piscataway, NJ, USA, 2009; pp. 843–847. [[CrossRef](#)]
19. Kieninger, T.; Dengel, A. The T-Recs Table Recognition and Analysis System. In Proceedings of the Document Analysis Systems: Theory and Practice, Third IAPR Workshop, DAS'98, Nagano, Japan, 4–6 November 1998; Lee, S., Nakano, Y., Eds.; Springer: Berlin/Heidelberg, Germany, 1998; Volume 1655, pp. 255–269. [[CrossRef](#)]
20. Hao, L.; Gao, L.; Yi, X.; Tang, Z. A Table Detection Method for PDF Documents Based on Convolutional Neural Networks. In Proceedings of the 12th IAPR Workshop on Document Analysis Systems, DAS 2016, Santorini, Greece, 11–14 April 2016; IEEE Computer Society: Piscataway, NJ, USA, 2016; pp. 287–292. [[CrossRef](#)]
21. 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 IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, 9–15 November 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1162–1167. [[CrossRef](#)]
22. Hashmi, K.A.; Pagani, A.; Liwicki, M.; Stricker, D.; Afzal, M.Z. CasTabDetectorRS: Cascade Network for Table Detection in Document Images with Recursive Feature Pyramid and Switchable Atrous Convolution. *J. Imaging* **2021**, *7*, 214. [[CrossRef](#)] [[PubMed](#)]
23. Moll, M.A.; Baird, H.S. Segmentation-based retrieval of document images from diverse collections. In Proceedings of the Document Recognition and Retrieval XV, part of the IS&T-SPIE Electronic Imaging Symposium, San Jose, CA, USA, 29–31 January 2008; Yanikoglu, B.A., Berkner, K., Eds.; SPIE: Bellingham, WA, USA, 2008; Volume 6815, p. 68150L. [[CrossRef](#)]
24. Nayef, N.; Ogier, J. Text zone classification using unsupervised feature learning. In Proceedings of the 13th International Conference on Document Analysis and Recognition, ICDAR 2015, Nancy, France, 23–26 August 2015; IEEE Computer Society: Piscataway, NJ, USA, 2015; pp. 776–780. [[CrossRef](#)]
25. Tombre, K.; Tabbone, S.; Pélissier, L.; Lamiroy, B.; Dosch, P. Text/Graphics Separation Revisited. In Proceedings of the Document Analysis Systems V, 5th International Workshop, DAS 2002, Princeton, NJ, USA, 19–21 August 2002; Lopresti, D.P., Hu, J., Kashi, R.S., Eds.; Springer: Berlin/Heidelberg, Germany, 2002; Volume 2423, pp. 200–211. [[CrossRef](#)]
26. Zhong, X.; Tang, J.; Jimeno-Yepes, A. PubLayNet: Largest dataset ever for document layout analysis. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, 20–25 September 2019.
27. Zach, C.; Sánchez, A.P.; Pham, M. A dynamic programming approach for fast and robust object pose recognition from range images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015; IEEE Computer Society: Piscataway, NJ, USA, 2015; pp. 196–203. [[CrossRef](#)]
28. Bhatt, J.; Hashmi, K.A.A.; Afzal, M.Z.; Stricker, D. A Survey of Graphical Page Object Detection with Deep Neural Networks. *Appl. Sci.* **2021**, *11*, 5344. [[CrossRef](#)]
29. Hashmi, K.A.; Liwicki, M.; Stricker, D.; Afzal, M.A.; Afzal, M.A.; Afzal, M.Z. Current Status and Performance Analysis of Table Recognition in Document Images with Deep Neural Networks. *IEEE Access* **2021**, *9*, 87663–87685. [[CrossRef](#)]
30. Nazir, D.; Hashmi, K.A.; Pagani, A.; Liwicki, M.; Stricker, D.; Afzal, M.Z. HybridTabNet: Towards better table detection in scanned document images. *Appl. Sci.* **2021**, *11*, 8396. [[CrossRef](#)]
31. Baur, C.; Albarqouni, S.; Navab, N. Semi-supervised Deep Learning for Fully Convolutional Networks. In Proceedings of the Medical Image Computing and Computer Assisted Intervention - MICCAI 2017-20th International Conference, Quebec, QC, Canada, 11–13 September 1 2017; Descoteaux, M., Maier-Hein, L., Franz, A.M., Jannin, P., Collins, D.L., Duchesne, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2017; Volume 10435, pp. 311–319. [[CrossRef](#)]
32. Nie, D.; Gao, Y.; Wang, L.; Shen, D. ASDNet: Attention Based Semi-supervised Deep Networks for Medical Image Segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention - MICCAI 2018-21st International Conference, Granada, Spain, 16–20 September 2018; Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2018; Volume 11073, pp. 370–378. [[CrossRef](#)]
33. Roß, T.; Zimmerer, D.; Vemuri, A.S.; Isensee, F.; Bodenstedt, S.; Both, F.; Kessler, P.; Wagner, M.; Müller, B.; Kenngott, H.; et al. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *Int. J. Comput. Assist. Radiol. Surg.* **2018**, *13*, 925–933. [[CrossRef](#)] [[PubMed](#)]
34. Iscen, A.; Toliás, G.; Avrithis, Y.; Chum, O. Label Propagation for Deep Semi-supervised Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
35. Gan, Y.; Zhu, H.; Guo, W.; Xu, G.; Zou, G. Deep semi-supervised learning with contrastive learning and partial label propagation for image data. *Knowl. Based Syst.* **2022**, *245*, 108602. [[CrossRef](#)]

36. Kiran, B.R.; Thomas, D.M.; Parakkal, R. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *J. Imaging* **2018**, *4*, 36. [[CrossRef](#)]
37. Papandreou, G.; Chen, L.; Murphy, K.P.; Yuille, A.L. Weakly-and Semi-Supervised Learning of a Deep Convolutional Network for Semantic Image Segmentation. In Proceedings of the 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 7–13 December 2015; IEEE Computer Society: Piscataway, NJ, USA, 2015; pp. 1742–1750. [[CrossRef](#)]
38. Olsson, V.; Tranheden, W.; Pinto, J.; Svensson, L. ClassMix: Segmentation-Based Data Augmentation for Semi-Supervised Learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Montreal, QC, Canada, 11–17 October 2021.
39. Jeong, J.; Lee, S.; Kim, J.; Kwak, N. Consistency-based Semi-supervised Learning for Object detection. In Proceedings of the Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019; Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R., Eds.; 2019; pp. 10758–10767. Available online: <https://papers.nips.cc/paper/2019/hash/d0f4dae80c3d0277922f8371d5827292-Abstract.html> (accessed on 28 April 2022).
40. Miyato, T.; Maeda, S.; Koyama, M.; Ishii, S. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1979–1993. [[CrossRef](#)] [[PubMed](#)]
41. Sajjadi, M.; Javanmardi, M.; Tasdizen, T. Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning. 2016. Available online: <https://proceedings.neurips.cc/paper/2016/file/30ef30b64204a3088a26bc2e6ecf7602-Paper.pdf> (accessed on 28 April 2022).
42. Grandvalet, Y.; Bengio, Y. Semi-supervised Learning by Entropy Minimization. In Proceedings of the Neural Information Processing Systems 17 Neural Information Processing Systems, NIPS 2004, Vancouver, BC, Canada, 13–18 December 2004; pp. 529–536.
43. Berthelot, D.; Carlini, N.; Goodfellow, I.J.; Papernot, N.; Oliver, A.; Raffel, C. MixMatch: A Holistic Approach to Semi-Supervised Learning. 2019. Available online: <https://proceedings.neurips.cc/paper/2019/file/1cd138d0499a68f4bb72bee04bbec2d7-Paper.pdf> (accessed on 28 April 2022).
44. Jeong, J.; Verma, V.; Hyun, M.; Kannala, J.; Kwak, N. Interpolation-based semi-supervised learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
45. Radosavovic, I.; Dollár, P.; Girshick, R.B.; Gkioxari, G.; He, K. Data Distillation: Towards Omni-Supervised Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, USA, 21–26 July 2017.
46. Yang, Q.; Wei, X.; Wang, B.; Hua, X.; Zhang, L. Interactive Self-Training With Mean Teachers for Semi-Supervised Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, 19–25 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 5941–5950.
47. Tang, Y.; Chen, W.; Luo, Y.; Zhang, Y. Humble Teachers Teach Better Students for Semi-Supervised Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
48. Lin, T.; Maire, M.; Belongie, S.J.; Bourdev, L.D.; Girshick, R.B.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
49. Sohn, K.; Berthelot, D.; Li, C.; Zhang, Z.; Carlini, N.; Cubuk, E.D.; Kurakin, A.; Zhang, H.; Raffel, C. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 596–608.
50. Mondal, A.; Lipps, P.; Jawahar, C.V. IIIT-AR-13K: A New Dataset for Graphical Object Detection in Documents. In Proceedings of the International Workshop on Document Analysis Systems, Wuhan, China, 26–29 July 2020.
51. Li, M.; Xu, Y.; Cui, L.; Huang, S.; Wei, F.; Li, Z.; Zhou, M. DocBank: A Benchmark Dataset for Document Layout Analysis. *arXiv* **2020**, arXiv:2006.01038.
52. Sohn, K.; Zhang, Z.; Li, C.; Zhang, H.; Lee, C.; Pfister, T. A Simple Semi-Supervised Learning Framework for Object Detection. *arXiv* **2020**, arXiv:2005.04757.
53. Powers, D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* **2020**, arXiv:2010.16061.
54. Lin, T.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S.J. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; IEEE Computer Society: Piscataway, NJ, USA, 2017; pp. 936–944. [[CrossRef](#)]
55. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
56. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021.
57. Nguyen, P.; Ngo, L.; Truong, T.; Nguyen, T.T.; Vo, N.D.; Nguyen, K. Page Object Detection with YOLOF. In Proceedings of the 2021 8th NAFOSTED Conference on Information and Computer Science (NICS), Hanoi, Vietnam, 21–22 December 2021; pp. 205–210. [[CrossRef](#)]