

# UnSupDLA: Towards Unsupervised Document Layout Analysis

Talha Uddin Sheikh<sup>\*1,2,3</sup>[0009-0004-9156-5679], Tahira Shehzadi<sup>\*1,2,3</sup>[0000-0002-7052-979X], Khurram Azeem Hashmi<sup>1,2,3</sup>[0000-0003-0456-6493], Didier Stricker<sup>1,2,3</sup>, and Muhammad Zeshan Afzal<sup>1,2,3</sup>[0000-0002-0536-6867]

<sup>1</sup> Department of Computer Science, Technical University of Kaiserslautern, Germany

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

<sup>3</sup> German Research Institute for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany

{firstname\_middlename.lastname@dfki.de}

**Abstract.** Document layout analysis is a key area in document research, involving techniques like text mining and visual analysis. Despite various methods developed to tackle layout analysis, a critical but frequently overlooked problem is the scarcity of labeled data needed for analyses. With the rise of internet use, an overwhelming number of documents are now available online, making the process of accurately labeling them for research purposes increasingly challenging and labor-intensive. Moreover, the diversity of documents online presents a unique set of challenges in maintaining the quality and consistency of these labels, further complicating document layout analysis in the digital era. To address this, we employ a vision-based approach for analyzing document layouts designed to train a network without labels. Instead, we focus on pre-training, initially generating simple object masks from the unlabeled document images. These masks are then used to train a detector, enhancing object detection and segmentation performance. The model's effectiveness is further amplified through several unsupervised training iterations, continuously refining its performance. This approach significantly advances document layout analysis, particularly precision and efficiency, without labels.

**Keywords:** Unsupervised Learning · Document Segmentation · Document Object Detection · Document Layout Analysis.

## 1 Introduction

Document layout analysis (DLA) has always been a key challenge in computer vision and document understanding. Historically, the field has developed diverse methodologies [1], ranging from traditional classical techniques [2,3,4] to more contemporary, learning-based models [5,6]. The advancement of technologies

---

\* These authors contributed equally to this work

such as convolutional neural networks (CNNs) has marked a notable improvement in the precision and functionality of these models, showing a significant evolution in the approach to DLA [7,8,9,10,11,12,13]. As technology advances, there has been a corresponding change in the complexity of documents, particularly in the digital domain. This shift is most evident in business environments, where documents come in increasingly varied and complex formats [1,14]. These developments present a new set of challenges, requiring models that are accurate and adaptable enough to adjust to a wide range of document types and layouts. In response to this dynamic landscape, the strategies employed in DLA have been continuously refined and improved. The focus has expanded to include the accuracy of analysis and the adaptability to handle the diverse array of modern document formats. This ongoing advancement in DLA methods underscores the importance and persistent relevance of the field in the broader context of document understanding and computer vision research. As documents continue to evolve, so will the techniques and technologies in DLA, ensuring that it remains an essential and ever-progressing study area [15].

Previously, classical rule-based methods were employed for document layout analysis [16,17,18,19,20]. More recently, it’s been approached as a Document Object Detection (DOD) problem, employing vision-based object detection models [21,19,20,22,23,24,25,26]. Researchers have also combined sequence and language models with object detection for better accuracy [5]. However, there’s an overlooked issue. Unconventional document formats require labor-intensive annotation for traditional supervised methods. So, unsupervised approaches have become important. Implementing unsupervision in DOD is challenging because images contain multiple document objects of different classes, and treating each image as a class isn’t effective. However, it’s worth noting that these salient object detection methods [27] are specifically designed to locate a single object, typically the most prominent one, and may not be suitable for handling real-world document images containing multiple objects and complex layouts. This raises questions about the effectiveness of unsupervision in document segmentation.

In this paper, we identify and localize graphical elements within documents without labels. In the initial phase of unsupervised training, We use unlabeled data, which lacks specific information about the locations and types of objects in the documents. We generate initial layout masks based on features from a self-supervised DINO [28]. We analyze patch-wise similarities for images with multiple objects and use Normalized Cuts (NCut) to isolate a mask for each object, repeating this multiple times for multiple objects. Later, we apply a loss drop strategy in the detector training to improve performance. The model undergoes several iterations of unsupervised training for further refinement. Previous research has shown self-supervised vision-based methods [28,29] to be less effective for DLA tasks because they require direction from learned text and layout embeddings. Yet, we propose that unsupervised learning employs visual representation. The visual features generate masks that provide a preliminary idea of where objects might be located within the documents, serving as a start-

ing point for further analysis. In short, our approach does not rely on layout information from pre-trained text recognition models. Instead, we use the inherent visual information within documents as a layout guide for learning visual representations.

In summary, the contributions of our paper are as follows:

- A vision-based unsupervised learning framework aims to train the detector to perform document layout analysis. This approach recognizes and analyzes the layout of documents autonomously.
- A layout-guided strategy that generates initial layout masks using visual features for document segmentation.
- An efficient unsupervised learning approach that learns about different document objects to minimize data use. It can be used as a pre-training model for document analysis.

We organize the content of the paper as follows. We begin with a thorough review of existing literature in Section 2. Then, in Section 3, we detail the methodology. Section 4 is dedicated to the discussion of our experiments and the results obtained. In Section 5, we conduct an ablation analysis. Finally, we conclude our paper in Section 6 with our final thoughts and findings.

## 2 Related Work

### 2.1 Fully-Supervised Document Understanding

Recent advancements in deep learning methodologies have broadened their applications, extending from healthcare [30,31], traffic analysis [32], to document analysis [33,34,35,36,37,38]. In recent years, the idea of Document Understanding (DU) has expanded to include many different challenges and tasks related to Document Intelligence systems [39]. This includes, but is not limited to, Key Information Extraction [40,41,42], Document Classification [43], Document Layout Analysis [44,45], Question Answering [46,47], and Machine Reading Comprehension [48], particularly when dealing with Visually Rich Documents (VRDs) as opposed to simple text or basic image-text combinations. Leading DU systems predominantly utilize extensive pre-training to merge visual and textual elements [49,5,29,50,51]. However, methods like Donut [52] and Dessurt [53] focus more on enhancing visual features using synthetic generation techniques [54,55,56] for effective layout representation during document pre-training.

### 2.2 Fully-Supervised Document Layout Analysis

DLA has emerged as a key application in data utilization, focusing on optimizing storage and handling of vast amounts of information [1]. The field has transformed with the introduction of deep learning and Convolutional Neural Networks (CNN), leading to a shift in document layout segmentation [57,6,58,59,60] towards a Document Object Detection. The development of extensive DLA

benchmarks [44,45] has made it easier for deep learning techniques to be applied in this field. Biswas et.al [61] has considered DLA as an instance-level segmentation task that is crucial for identifying bounding boxes and segmentation masks in pages with overlapping elements. Transformer-based methods [5,62] have recently achieved improved results in DLA, particularly for large-scale document datasets, though they still face challenges in smaller datasets. Innovative language-based methods like LayoutLMv3 [5] and UDoc [50] have shown impressive results on the PubLayNet benchmark but struggle with more complex layouts and smaller data samples.

### 2.3 Advancements in Self-Supervised Learning

In the evolving field of computer vision, researchers have been concentrating on understanding complex visual details from different images. This led to the development of data-driven machine learning models, for extracting and correlating features, to meet increasingly complex demands. Advanced networks require a lot of data. This makes data annotation very important, leading to many self-supervised learning strategies. MoCo [63] introduced a novel approach in contrastive learning settings, utilizing exponential moving averages and large memory banks for weight updates. Building on this, SimCLR [64] proposed using larger batch sizes as an alternative to memory banks. DINO [28] brought the concept of self-supervision to vision transformers [65]. MoCov2 [66] and SwAV [67] subsequently achieved remarkable results within this self-supervised framework. Alternatively, BYOL [68] and SimSiam [69] approached the problem by treating different sections of the same image as analogous pairs, moving away from traditional contrastive learning. Additionally, masked autoencoders [70] have revitalized classic autoencoder techniques by incorporating a masking strategy for learning representations through reconstruction.

Despite the remarkable success of supervised object detection techniques such as Mask RCNN [71], Yolo [72], Retinanet [73], and DETR [74], their self-supervised alternatives have been somewhat limited in scope until recently. Recent advancements have seen the development of end-to-end self-supervised object detection models like UP-DETR [75] and DETReg [76], as well as backbone pre-training strategies such as Self-EMD [77] and Odin [78]. While significant research has been done on self-supervised learning, unsupervised methods still need to be explored. While some attempts have been made at unsupervised document analysis [79,80], these methods have yet to improve effectively. This paper aims to fill this gap by introducing

## 3 Methodology

In our research, we focus on applying unsupervised learning to document layout segmentation and object detection domains, as shown in Fig. 1. Our primary data, denoted as  $\mathcal{D}$ , consists of a comprehensive collection of RGB document images. To align with the unsupervised learning framework, which emphasizes

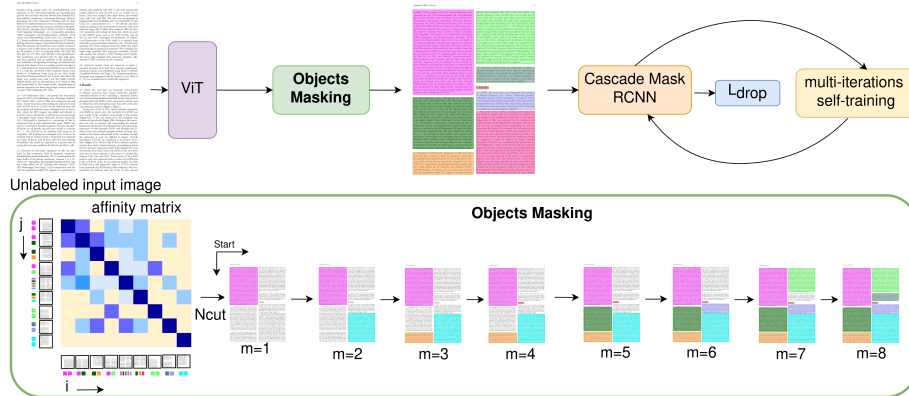


Fig. 1: Overview of our unsupervised training module: It takes unlabeled data to train models for object detection and instance segmentation. Then, Objects Masking [27] generates rough object masks utilizing the features of self-supervised DINO [28]. We employ a patch-wise similarity matrix for multiple object masks in an unlabeled image. Applying Normalized Cuts (Ncut) to this matrix, we initially extract a mask for a single foreground object. This procedure is repeated, altering the affinity matrix each time, allowing Objects Masking to discover multiple object masks in one image, demonstrated here with eight iterations.

learning from unlabeled data, we derive an unlabeled dataset  $\mathcal{D}_u = \{x_u^i\}_{i=1}^{N_u}$  from  $\mathcal{D}$ , where  $N_u$  represents the total number of images in  $\mathcal{D}_u$ . It does not contain traditional annotations or labels usually associated with supervised learning tasks, such as explicit object categories, locations, or dimensions.

Initially, we employ a mask generation technique following [81,27,28] that creates several binary masks for each document image utilizing unsupervised features derived from DINO [28]. The approach for extracting this mask is detailed in Section 3.1, highlighting the extraction process that emphasizes the document’s physical layout. Furthermore, as outlined in Section 3.2, we employ a dynamic loss reduction approach to effectively train a detector using the initial masks generated previously while simultaneously prompting the model to identify object masks that may have been overlooked. Lastly, as explained in Section 3.3, we enhance our method’s effectiveness by implementing several iterations of unsupervised training.

### 3.1 Layout Mask Generation for Multiple Objects

Generating the layout masks is crucial in our approach, as our unsupervised framework relies on them for visual guidance. For input document image  $x$ , we create multiple object masks within an image without the need for any manual annotations. In our approach, we initially partition the input document image

into smaller image patches. We create a patch-wise similarity matrix to analyze the relationships between these patches. The crucial aspect here is using a self-supervised DINO [28], which extracts meaningful features from these patches without needing labeled data. These extracted features are then employed to determine the similarity between each pair of patches, resulting in the formation of the similarity matrix as follows:

$$W_{ij} = \frac{F_i F_j}{\|F_i\|^2 \|F_j\|^2} \quad (1)$$

where  $F_i$  and  $F_j$  represent the key features of patch  $i$  and patch  $j$ , respectively. The diagonal elements in the patch-wise similarity matrix have the highest values because they represent the same patch overlapping with itself, making them inherently identical and, therefore, maximally similar as shown by arrows around infinity matrix in Fig 1. This matrix is a fundamental component in our pipeline, facilitating subsequent analysis and tasks by capturing the visual relationships within the document image. We then employ the Normalized Cuts algorithm [82] on the similarity matrix, generating a single mask that highlights the primary foreground object within the image. Normalized Cuts (NCut) approaches consider image segmentation a problem of dividing a graph into meaningful parts. To do this, we create a fully interconnected and undirected graph, representing each image patch as a node. Edges between nodes are established with weights, denoted as  $W_{ij}$ , which quantify how similar the connected nodes are. NCut aims to find the optimal way to split this graph into two distinct sub-graphs, essentially forming a bipartition. It is achieved by solving a generalized eigenvalue system, minimizing the overall cost of this partitioning process as follows:

$$(D^m - W)x^m = \lambda D^m x^m \quad (2)$$

where  $x^m$  is the eigenvector associated with the second smallest eigenvalue  $\lambda$  at stage  $m$ . Here,  $D^m$  represents a diagonal matrix of size  $N \times N$ , with  $d(i) = \sum_j W_{ij}$ , and  $W$  is a symmetrical matrix of size  $N \times N$ . One crucial aspect of this approach is determining which group of patches corresponds to the foreground, a fundamental step in object mask generation. For this, we employ two specific criteria. Firstly, we identify the patch with the highest absolute value in the second smallest eigenvector of the binary mask  $M^m$ . This selection intuitively represents the most prominent part of the foreground, enhancing object detection. Secondly, we incorporate a straightforward yet empirically effective prior: the foreground group should not contain two of the four input image corners. These criteria help ensure accurate identification of the foreground and background regions. The generated mask for a single document object is as follows:

$$M_{ij}^m = \begin{cases} 1, & \text{if } M_{ij}^m \geq \text{mean}(x^m) \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

where, If  $M_{ij}^m$  is greater than or equal to the average value of  $x^m$ , it sets  $M_{ij}^m$  to 1, effectively marking that element in the mask. If  $M_{ij}^m$  is less than the mean of  $x^m$ ,

it sets  $M_{ij}^m$  to 0, indicating that the element is not part of the mask. In this way, it generates a mask that identifies elements belonging to the foreground. If we don't meet certain criteria previously explained, as if there are two input image corners in the current foreground, We reverse the foreground and background as  $M_{ij}^m = 1 - M_{ij}^m$ . Moreover, we set values of  $W_{ij}$  less than  $\tau_t$  to  $1 \times 10^{-5}$  and values greater than or equal to  $\tau_t$  to 1.

**Mask Pooling:** To ensure that each object in the sequence receives a distinct mask, focusing on different data or image areas. We exclude nodes previously identified as part of the foreground. This exclusion ensures that the mask generation process remains consistent with the specific characteristics of each object, leading to accurate mask generation. For this, we obtain the mask for the  $(m+1)_{th}$  object by updating the node similarity  $W_{ij}^{m+1}$  and excluding the nodes corresponding to the foreground in previous stages as follows:

$$W_{ij}^{m+1} = \frac{(F_i \prod_{l=1}^m \hat{M}_{ij}^l)(F_j \prod_{l=1}^m \hat{M}_{ij}^l)}{\|F_i\|_2 \|F_j\|_2} \quad (4)$$

were,  $\hat{M}_{ij}^l = 1 - M_{ij}^l$ . Here, masking by excluding the nodes of previously masked foreground enables our approach to uncover multiple object masks within a single image. In document mask generation, we've set  $m$  to 10. We can vary this according to maximum possible objects in the document image. In Fig 1 we adept at generating up to six distinct object masks in the image. This strategic masking enables the uncovering of multiple object masks within a single image. Employing the updated similarity matrix  $W_{m+1,ij}$ , we iterate through Eqs. 1 and 2 to derive a new mask denoted as  $M^{m+1}$ . This innovative pipeline allows us to reveal and distinguish various objects within the same image without manual supervision or annotations.

**Augmentation:** In our training process, we incorporate copy-paste augmentation approach, following [83,84]. However, we modify this technique to enhance our model's ability to segment small objects precisely. Traditionally, copy-paste augmentation involves taking a portion of an image and placing it elsewhere within the same image or in another image. Instead of following this conventional approach, we introduce an additional step. When we copy a portion of the mask, we randomly reduce its size by a certain factor. This reduction is determined by a scalar value that we randomly select from a uniform distribution between 0.3 and 1.0. For small objects, we downsizing the mask this way to effectively replicate scenarios where objects are small. This adjustment aids the model in becoming more proficient at handling and accurately segmenting these smaller objects throughout its training process, leading to an overall enhancement in its performance.

### 3.2 Loss Reduction for Exploring Object Regions

In standard object detection, the loss function penalizes predictions  $p_j$  that do not align with the actual ground-truth. However, in our unsupervised setting,

we consider the previously generated mask as the ground-truth that may overlook certain instances, making it essential to extend beyond the standard loss to enable the detector to identify new, unlabeled instances effectively. To address this challenge, we employ  $L_{\text{drop}}$ , which selectively ignores the loss for predicted regions ( $p_j$ ) that exhibit minimal overlap with the masked ground-truth. During training, we drop the loss for each predicted region ( $p_j$ ) if its maximum Intersection over Union (IoU) with any masked ground-truth instance is below a threshold of  $\tau_i = 0.01$ , as described by the equation:

$$L_{\text{drop}}(p_j) = \begin{cases} L_{\text{det}}(p_j) & \text{if } \text{IoU}_j^{\text{max}} > \tau_i = 0.01 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Here,  $\text{IoU}_j^{\text{max}}$  represents the highest IoU of  $p_j$  with all generated masked instances, and  $L_{\text{det}}$  denotes the conventional loss function used in detectors. By implementing  $L_{\text{drop}}$ , the model avoids penalties for detecting objects not present in the previously generated mask, allowing it to focus on exploring various image regions.

### 3.3 Multi-Iterations Unsupervised Training

Our experiments show that as we train detection models, they become surprisingly good at improving the quality of the masks they generate. Even when they start with rough masks, the models gradually make them better. It, along with  $L_{\text{drop}}$  strategy, helps the models find new object masks effectively. To improve performance, we employ multiple rounds of unsupervised training. We take the masks and proposals generated in the previous round in each round, but only if they have a confidence score exceeding  $0.75 - 0.5$  from the  $m$ -th round. These become annotations for the next round ( $m + 1$ )-th, helping the model learn more about the objects in the data. To avoid feeding the network redundant information, we skip ground-truth masks that have IoU greater than 0.5 with the predicted masks. We aim to avoid redundancy in the model’s learning process to ensure efficiency. Our experiments have shown that doing this training process three times works well. With each round, the model has more high-quality mask examples to learn from, making it better at generating object masks in complex scenes.

## 4 Experimental Setup

### 4.1 Datasets

We employ several specialized datasets such as PubLayNet [44], DocLayNet [85], and TableBank [86] for our document unsupervised detection and segmentation framework. DocLayNet [85] dataset includes 69,375 training images, 6,489 validation images, and 4,999 test images across six domains, each annotated for 11 classes. PubLayNet [44], a large public dataset, contains 335,703 training,



11,240 validation, and 11,405 test images, with annotations for figures, lists, titles, tables, and texts in academic images. TableBank [86] dataset is designed to identify tables in scientific documents and contains 417,000 document images from the arXiv database. It classifies tables into LaTeX, Word, and combined categories and includes table structure recognition data. However, we only used the training images without ground-truth labels during the training.

## 4.2 Evaluation Metrics

We evaluate our unsupervised document analysis approach using the following metrics:  $mAP^{box}$ ,  $AP_{50}^{box}$ ,  $AP_{75}^{box}$ ,  $mAP^{mask}$ ,  $AP_{50}^{mask}$ , and  $AP_{75}^{mask}$ . The mean Average Precision  $mAP^{box}$  calculates the average precision of bounding box detections.  $AP_{50}^{box}$  and  $AP_{75}^{box}$  extend this evaluation to specific IoU thresholds of 50% and 75%, respectively. Similarly,  $mAP^{mask}$  measures the precision of object segmentation masks, while  $AP_{50}^{mask}$  and  $AP_{75}^{mask}$  assess this precision at the same IoU thresholds. These metrics provide a comprehensive assessment of the model’s capability in accurately detecting and segmenting objects with varying degrees of precision.

## 4.3 Implementation Details

Our approach employs Document analysis dataset, without utilizing any annotations during training. For image processing, Objects Masking is employed in three stages. Images are resized to  $480 \times 480$  pixels, and a patch-wise similarity matrix is generated using the ViT-B/8 DINO model. Post-processing of masks is conducted using a Conditional Random Field (CRF) to calculate their bounding boxes. We employ Cascade Mask R-CNN [87] starting with initial masks and bounding boxes for  $150k$  iterations. Specifically, when leveraging a ResNet-50 backbone [88], the model is initially equipped with weights from a self-supervised pretrained DINO model [28]. We train our network on 2 GPUs RTX A6000 for around 8 hours. The detector is optimized over  $150k$  iterations using Stochastic Gradient Descent (SGD). It begins with a learning rate of 0.005, which is decreased by 5 times after  $80k$  iterations. The training uses batches of 16, a weight decay of  $5 \times 10^{-5}$ , and a momentum of 0.9.

## 4.4 Performance Analysis

The effectiveness of our unsupervised training method is evaluated in Table 1. It shows unsupervised performance for object detection and instance segmentation on different datasets, PubLayNet, DocLayNet, and TableBank. TableBank outperforms PubLayNet and DocLayNet due to its single-class focus on tables, making the task simpler. Consequently, TableBank achieves significantly higher accuracy in both bounding box and mask predictions. We initialize the backbone with DINO network [28] and employ cascade Mask RCNN as the detector. TableBank shows high AP and mAP scores, indicating precise detection and segmentation capabilities without table labels.



Fig. 2: Comparative visual analysis of unsupervised learning on the PubLayNet dataset: top-predicted layouts; bottom-corresponding ground-truth layouts. The model’s proficiency in detecting details overlooked by human annotators is also highlighted, marked by red arrows.

Table 1: Quantitative analysis of unsupervised detection and segmentation in document datasets such as PubLayNet, DocLayNet, and TableBank. We discuss the effectiveness of detection and segmentation, focusing on the detection method and backbone initialization (Init) with DINO [28]. The term ‘Cascade’ here represents the Cascade Mask R-CNN network [87].

Dataset	Unsup-train	Detector	Init.	Performance					
				$mAP^{box}$	$AP_{50}^{box}$	$AP_{75}^{box}$	$mAP^{mask}$	$AP_{50}^{mask}$	$AP_{75}^{mask}$
PubLayNet				28.7	43.1	30.0	29.3	44.1	30.5
DocLayNet	✓		‘Cascade’ DINO	22.4	37.5	23.1	24.2	38.7	24.8
TableBank				88.6	91.2	89.7	88.8	91.2	89.7

TableBank has mAP of 88.6% for detection and 88.8% for segmentation on unsupervised training. Fig. 2 shows the performance of our unsupervised learning approach on the PubLayNet dataset. The analysis includes the unsupervised model’s predicted layouts against the ground-truth layouts. Notably, the model demonstrates an improved ability to recognize various elements within a document, such as footers. It also excels in precisely segmenting smaller components like text blocks. A key aspect of this analysis is the model’s remarkable performance in identifying fine details within the layouts, some of which might even

be missed by human annotators. These instances, where the model’s predictions positively diverge from human annotations, are specifically highlighted with red arrows. It highlights the model’s advanced capability in document object detection and segmentation in unsupervised settings.

Table 2: Merged Results for PubLayNet, TableBank, and DocLayNet

Methods	PubLayNet		TableBank		DocLayNet	
	$mAP^{\text{box}}$	$mAP^{\text{mask}}$	$mAP^{\text{box}}$	$mAP^{\text{mask}}$	$mAP^{\text{box}}$	$mAP^{\text{mask}}$
<b>Fully-supervised methods:</b>						
V+BERT-12L [58]	96.5	-	-	-	81.0	-
VGT [59]	96.2	-	-	-	-	-
SwinDocSegmenter [60]	-	93.72	-	-	98.04	-
TRDLU [89]	95.95	-	-	-	-	-
VSR [90]	95.7	-	-	-	-	-
CDeC-Net [7]	96.7	-	89.8	-	-	-
DocSegTr [8]	-	-	-	93.3	-	-
Layout LMv3 [9]	-	-	-	92.9	-	-
GLAM + YOLOv5x6 [91]	-	-	-	-	-	80.8
Mask R-CNN [92]	-	-	-	-	-	78.0
<b>Unsupervised methods:</b>						
Our	28.7	29.3	88.6	88.8	22.4	24.2

Table 2 compares our unsupervised approach with previous fully supervised approaches, highlighting the effectiveness of the unsupervised approach in object detection and segmentation tasks within document layout analysis. Supervised methods, which have the advantage of learning from labeled data, generally yield high precision scores; for instance, SwinDocSegmenter [60] achieves an impressive 93.72  $AP_{\text{box}}$  on TableBank, indicating its strong capability to identify and localize objects accurately. However, the unsupervised method is particularly noteworthy, achieving an  $AP_{\text{mask}}$  of 88.8 on TableBank without the aid of labeled training data. This high score in segmentation precision suggests that our approach can predict the shapes and boundaries of document elements, such as tables or text blocks, almost as effectively as its supervised approaches. The ability of our approach to perform so well in an unsupervised manner is significant as it implies a considerable reduction in the dependency on costly and time-consuming data labeling processes. It also opens up new possibilities for analyzing documents in domains where obtaining labeled data is difficult, thus expanding the applicability of unsupervised learning in document analysis. Therefore, it provides a performance benchmark for current methods and the possibility of unsupervised learning approaches in real-world document layout understanding tasks.

Size $\rightarrow$	240	360	480	640	$\tau_t \rightarrow$	0	0.1	0.15	0.2
$AP_{50}^{mask}$	86.4	87.5	88.6	88.7	$AP_{50}^{mask}$	88.2	88.5	88.6	88.5
(a) Image size.					(b) $\tau_t$ for Objects Masking.				
$N \rightarrow$	5	10	15		$\tau_i \rightarrow$	0	0.01	0.1	0.2
$AP_{50}^{mask}$	88.1	88.6	88.6		$AP_{50}^{mask}$	88.3	88.6	85.5	82.9
(c) # masks per image.					(d) $\tau_i$ for $L_{drop}$ .				

Table 3: Ablations for mask generation and loss reduction for exploring object regions. This study examines the impact of different parameters on unsupervised training performance using the TableBank dataset. The parameters varied include: (a) image size, (b) the threshold value  $\tau_t$  which determines the sparsity level of the affinity matrix in Normalized Cuts, (c) the number of masks generated by Objects Masking, and (d) the threshold  $\tau_i$  in  $L_{drop}$ , which is the maximum allowable overlap between predicted regions and ground-truth before excluding loss for those regions. Default parameter settings are indicated in gray.

## 5 Ablation Study

**Design choices of unsupervised training parameters.** This study conducts an ablation analysis on the design choices of unsupervised training parameters in the context of mask generation and loss reduction for exploring object regions, as shown in Table 3. The research is centered on utilizing the TableBank dataset to evaluate the impact of various parameters on unsupervised training performance. The parameters under scrutiny encompass: (a) the image size, (b) the threshold value  $\tau_t$ , which plays a crucial role in determining the sparsity of the affinity matrix within the Normalized Cuts method, (c) the quantity of masks generated through the Objects Masking technique, and (d) the threshold  $\tau_i$  in  $L_{drop}$ , which dictates the maximum allowable overlap between predicted regions and ground-truth before dismissing the loss for those regions. A key aspect of this analysis is identifying default parameter settings, which are distinctly highlighted in gray for reference. Understanding the influence of unsupervised training parameters in object region exploration is important for optimizing mask generation and loss reduction efficiency and accuracy. By varying these parameters and assessing their effects on performance, this research provides the best results for enhancing the overall performance. The study’s insights can aid in fine-tuning unsupervised training processes, ensuring more precise and effective results in tasks related to document analysis and object recognition.

**Effectiveness of unsupervised training iterations.** Multiple rounds of unsupervised training effectively enhance the quality and quantity of object masks, as indicated in Table 4. Through iterative refinement, the model progressively improves the precision of object masks, even when starting with rough initial

predictions. This process generates more masks, aiding the model’s training. Combining these masks with the  $L_{\text{drop}}$  strategy, which focuses on uncertain predictions, helps the model target areas where it initially struggles, improving mask accuracy. Our experiments suggest that performing unsupervised training three times provides a balance between generating high-quality masks and avoiding overfitting, making it particularly valuable for handling even small and complex document objects in document analysis data.

Table 4: Analysis of training iterations in unsupervised learning. Here, analysis shows that three iterations provide the best results using Cascade Mask RCNN on the TableBank dataset.

Iteration	$mAP^{box}$	$AP_{50}^{box}$	$AP_{75}^{box}$	$mAP^{mask}$	$AP_{50}^{mask}$	$AP_{75}^{mask}$
1	86.2	89.5	88.4	88.2	89.6	88.9
2	88.3	90.7	89.1	88.5	90.8	89.4
3	88.6	91.2	89.7	88.8	91.2	89.7
4	88.6	91.0	89.5	88.7	91.2	89.7

**Effectiveness of quantity of pre-training data.** The quantity of unsupervised training data significantly influences the effectiveness of our unsupervised approach. Essentially, the larger the dataset we have for training, the better our model tends to perform in terms of its ability to generalize and achieve higher performance. This relationship between data quantity and model performance is demonstrated in Table 5. Using only 10% of the data for unsupervised training, we achieved an mAP of 82.9 for detection and 85.2 for segmentation. However, when we utilized the full 100% of the available data, our performance improved significantly to an mAP of 88.6 for detection and 88.8 for segmentation.

Table 5: Performance analysis of Cascade Mask RCNN unsupervised training with varying percentages of data utilized in TableBank dataset.

% data	$mAP^{box}$	$AP_{50}^{box}$	$AP_{75}^{box}$	$mAP^{mask}$	$AP_{50}^{mask}$	$AP_{75}^{mask}$
10%	82.9	88.9	86.0	85.2	88.9	86.6
30%	85.4	89.3	87.2	86.2	89.3	87.2
50%	85.8	90.5	88.1	87.3	90.5	88.2
100%	88.6	91.2	89.7	88.8	91.2	89.7

**Effectiveness of cross-data unsupervised learning.** Moreover, in Table 6, we examine the impact of training data on the efficiency of unsupervised training.

Specifically, we investigate the performance differences when a network is unsupervisedly trained sequentially on two distinct datasets. Initially, the network undergoes unsupervised training for 150k iterations exclusively on just PubLayNet dataset. In second experiment, the network is first unsupervisedly trained on the TableBank dataset for 75k iterations. Following this, the network undergoes an

Table 6: Impact of Dataset Selection on Cross Unsupervised Training. We explore how different datasets affect cross unsupervised training results.

Cross Unsup-training	$mAP^{box}$	$AP_{50}^{box}$	$AP_{75}^{box}$	$mAP^{mask}$	$AP_{50}^{mask}$	$AP_{75}^{mask}$
PubLayNet	28.7	43.1	30.0	29.3	44.1	30.5
TableBank + PubLayNet	65.6	84.8	71.2	65.3	85.2	71.5

additional 75k iterations of unsupervised training on the PubLayNet dataset. Our findings reveal a significant performance improvement when cross-training is employed. Specifically, training solely on the PubLayNet dataset resulted in a mAP of 28.7 for document object detection. In contrast, the cross-data training approach, involving both TableBank and PubLayNet datasets, yields a substantially higher mAP of 65.6. Our experiments show that unsupervised training the network on multiple datasets, rather than just one, significantly improves its performance.

## 6 Conclusion

In conclusion, the paper presents a significant advancement in the field of document layout analysis by introducing a vision-based approach that effectively addresses the challenges of limited labeled data and the diversity of documents online. This method diverges from traditional techniques that rely heavily on labeled data, which are increasingly impractical due to the massive volume of documents on the internet. The proposed approach begins with pre-training that generates simple object masks from unlabeled document images, bypassing the need for extensive labeling. These masks are then employed to train a detector, leading to improved object detection and segmentation precision. The model’s performance is further enhanced through multiple training iterations, allowing for continuous refinement. This approach offers a more efficient, accurate, and flexible way for analyzing document layouts, making a major improvement in the field of document research. In the future research, we intend to investigate how unsupervised techniques can be utilized to improve Document Layout Analysis.

## Acknowledgements

The work leading to this publication has been partially funded by the EU Horizon Europe Project AIRISE (<https://airise.eu/>) under grant agreement 101092312.

## References

1. G. M. Binmakhshen and S. A. Mahmoud, "Document layout analysis: A comprehensive survey," *ACM Comput. Surv.*, vol. 52, no. 6, oct 2019. [Online]. Available: <https://doi.org/10.1145/3355610>
2. M. Agrawal and D. S. Doermann, "Voronoi++: A dynamic page segmentation approach based on voronoi and docstrum features," *2009 10th International Conference on Document Analysis and Recognition*, pp. 1011–1015, 2009. [Online]. Available: <https://api.semanticscholar.org/CorpusID:3355513>
3. S. Marinai, M. Gori, and G. Soda, "Artificial neural networks for document analysis and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 23–35, 2005.
4. J. Fang, L. Gao, K. Bai, R. Qiu, X. Tao, and Z. Tang, "A table detection method for multipage pdf documents via visual seperators and tabular structures," in *2011 International Conference on Document Analysis and Recognition*, 2011, pp. 779–783.
5. Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "Layoutlmv3: Pre-training for document ai with unified text and image masking," 2022. [Online]. Available: <https://arxiv.org/abs/2204.08387>
6. Z. Shen, R. Zhang, M. Dell, B. C. G. Lee, J. Carlson, and W. Li, "Layoutparser: A unified toolkit for deep learning based document image analysis," in *Document Analysis and Recognition – ICDAR 2021*, J. Lladós, D. Lopresti, and S. Uchida, Eds. Cham: Springer International Publishing, 2021, pp. 131–146.
7. M. Agarwal, A. Mondal, and C. V. Jawahar, "Cdec-net: Composite deformable cascade network for table detection in document images," *CoRR*, vol. abs/2008.10831, 2020. [Online]. Available: <https://arxiv.org/abs/2008.10831>
8. D. Prasad, A. Gadpal, K. Kapadni, M. Visave, and K. Sultanpure, "Cascadetabnet: An approach for end to end table detection and structure recognition from image-based documents," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.
9. Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "Layoutlmv3: Pre-training for document ai with unified text and image masking," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 4083–4091.
10. T. Shehzadi, K. A. Hashmi, D. Stricker, M. Liwicki, and M. Z. Afzal, "Bridging the performance gap between detr and r-cnn for graphical object detection in document images," *arXiv preprint arXiv:2306.13526*, 2023.
11. T. Shehzadi, D. Stricker, and M. Z. Afzal, "A hybrid approach for document layout analysis in document images," 2024.
12. T. Shehzadi, S. Sarode, D. Stricker, and M. Z. Afzal, "Towards end-to-end semi-supervised table detection with semantic aligned matching transformer," 2024.
13. I. Ehsan, T. Shehzadi, D. Stricker, and M. Z. Afzal, "End-to-end semi-supervised approach with modulated object queries for table detection in documents," *arXiv preprint arXiv:2405.04971*, 2024.
14. J. Bhatt, K. A. A. Hashmi, M. Z. Afzal, and D. Stricker, "A survey of graphical page object detection with deep neural networks," *Applied Sciences*, vol. 11, no. 12, 2021. [Online]. Available: <https://www.mdpi.com/2076-3417/11/12/5344>
15. L. Markewich, H. Zhang, Y. Xing, N. Lambert-Shirzad, Z. Jiang, R. K.-W. Lee, Z. Li, and S.-B. Ko, "Segmentation for document layout analysis: not dead yet," *International Journal on Document Analysis and Recognition (IJDAR)*, Jan 2022. [Online]. Available: <https://doi.org/10.1007/s10032-021-00391-3>

16. B. Coüiasnon and A. Lemaitre, "Recognition of tables and forms," in *Handbook of Document Image Processing and Recognition*, 2014.
17. R. Zanibbi, D. Blostein, and J. R. Cordy, "A survey of table recognition," *Document Analysis and Recognition*, vol. 7, no. 1, pp. 1–16, 2004.
18. A. M. Jorge, L. Torgo *et al.*, "Design of an end-to-end method to extract information from tables," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 8, no. 2, pp. 144–171, 2006.
19. S. Khusro, A. Latif, and I. Ullah, "On methods and tools of table detection, extraction and annotation in pdf documents," *Journal of Information Science*, vol. 41, no. 1, pp. 41–57, 2015.
20. D. W. Embley, M. Hurst, D. Lopresti, and G. Nagy, "Table-processing paradigms: a research survey," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 8, no. 2, pp. 66–86, 2006.
21. F. Cesarini, S. Marinai, L. Sarti, and G. Soda, "Trainable table location in document images," in *2002 International Conference on Pattern Recognition*, vol. 3, 2002, pp. 236–240 vol.3.
22. T. Shehzadi, K. A. Hashmi, D. Stricker, and M. Z. Afzal, "Object detection with transformers: A review," 2023.
23. X. Yang, M. E. Yümer, P. Asente, M. Kralej, D. Kifer, and C. L. Giles, "Learning to extract semantic structure from documents using multimodal fully convolutional neural network," *CoRR*, vol. abs/1706.02337, 2017. [Online]. Available: <http://arxiv.org/abs/1706.02337>
24. T. Shehzadi, K. A. Hashmi, A. Pagani, M. Liwicki, D. Stricker, and M. Z. Afzal, "Mask-aware semi-supervised object detection in floor plans," *Applied Sciences*, vol. 12, no. 19, 2022.
25. D. He, S. Cohen, B. Price, D. Kifer, and C. L. Giles, "Multi-scale multi-task fcn for semantic page segmentation and table detection," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, 2017, pp. 254–261.
26. T. Shehzadi, K. A. Hashmi, D. Stricker, and M. Z. Afzal, "Sparse semi-detr: Sparse learnable queries for semi-supervised object detection," *arXiv preprint arXiv:2404.01819*, 2024.
27. Y. Wang, X. Shen, Y. Yuan, Y. Du, M. Li, S. X. Hu, J. L. Crowley, and D. Vaufreydaz, "TokenCut: Segmenting objects in images and videos with self-supervised transformer and normalized cut," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
28. M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.
29. P. Li, J. Gu, J. Kuen, V. I. Morariu, H. Zhao, R. Jain, V. Manjunatha, and H. Liu, "Selfdoc: Self-supervised document representation learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5648–5656.
30. T. Shehzadi, A. Majid, M. Hameed, A. Farooq, and A. Yousaf, "Intelligent predictor using cancer-related biologically information extraction from cancer transcriptomes," in *2020 International Symposium on Recent Advances in Electrical Engineering & Computer Sciences (RAEE & CS)*, vol. 5, 2020, pp. 1–5.
31. A. Yousaf, T. Shehzadi, A. Farooq, and K. Ilyas, "Protein active site prediction for early drug discovery and designing," *International Review of Applied Sciences and Engineering*, vol. 13, no. 1, pp. 98 – 105, 2021.



32. W. Saeed, M. S. Saleh, M. N. Gull, H. Raza, R. Saeed, and T. Shehzadi, "Geometric features and traffic dynamic analysis on 4-leg intersections," *International Review of Applied Sciences and Engineering*, 2023.
33. M. Minouei, K. A. Hashmi, M. R. Soheili, M. Z. Afzal, and D. Stricker, "Continual learning for table detection in document images," *Applied Sciences*, vol. 12, no. 18, 2022. [Online]. Available: <https://www.mdpi.com/2076-3417/12/18/8969>
34. A. Kölsch, M. Z. Afzal, M. Ebbecke, and M. Liwicki, "Real-time document image classification using deep cnn and extreme learning machines," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, 2017, pp. 1318–1323.
35. S. Sinha, K. A. Hashmi, A. Pagani, M. Liwicki, D. Stricker, and M. Z. Afzal, "Rethinking learnable proposals for graphical object detection in scanned document images," *Applied Sciences*, vol. 12, no. 20, 2022. [Online]. Available: <https://www.mdpi.com/2076-3417/12/20/10578>
36. S. Naik, K. A. Hashmi, A. Pagani, M. Liwicki, D. Stricker, and M. Z. Afzal, "Investigating attention mechanism for page object detection in document images," *Applied Sciences*, vol. 12, no. 15, 2022. [Online]. Available: <https://www.mdpi.com/2076-3417/12/15/7486>
37. K. A. Hashmi, A. Pagani, M. Liwicki, D. Stricker, and M. Z. Afzal, "Cascade network with deformable composite backbone for formula detection in scanned document images," *Applied Sciences*, vol. 11, no. 16, 2021. [Online]. Available: <https://www.mdpi.com/2076-3417/11/16/7610>
38. K. A. Hashmi, D. Stricker, M. Liwicki, M. N. Afzal, and M. Z. Afzal, "Guided table structure recognition through anchor optimization," *CoRR*, vol. abs/2104.10538, 2021. [Online]. Available: <https://arxiv.org/abs/2104.10538>
39. Łukasz Borchmann, M. Pietruszka, T. Stanisławek, D. Jurkiewicz, M. Turski, K. Szyndler, and F. Graliński, "Due: End-to-end document understanding benchmark," in *NeurIPS Datasets and Benchmarks*, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:244906279>
40. G. Jaume, H. K. Ekenel, and J.-P. Thiran, "Funsd: A dataset for form understanding in noisy scanned documents," in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 2. IEEE, 2019, pp. 1–6.
41. S. Park, S. Shin, B. Lee, J. Lee, J. Surh, M. Seo, and H. Lee, "Cord: A consolidated receipt dataset for post-ocr parsing," 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:207900784>
42. T. Stanisławek, F. Graliński, A. Wróblewska, D. Lipiński, A. Kaliska, P. Rosalska, B. Topolski, and P. Biecek, "Kleister: Key information extraction datasets involving long documents with complex layouts," in *Document Analysis and Recognition – ICDAR 2021*, J. Lladós, D. Lopresti, and S. Uchida, Eds. Cham: Springer International Publishing, 2021, pp. 564–579.
43. A. W. Harley, A. Ufkes, and K. G. Derpanis, "Evaluation of deep convolutional nets for document image classification and retrieval," *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 991–995, 2015. [Online]. Available: <https://api.semanticscholar.org/CorpusID:2760893>
44. X. Zhong, J. Tang, and A. J. Yepes, "Publaynet: largest dataset ever for document layout analysis," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, Sep. 2019, pp. 1015–1022.
45. Z. Shen, K. Zhang, and M. Dell, "A large dataset of historical japanese documents with complex layouts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 548–549.

46. M. Mathew, D. Karatzas, and C. Jawahar, “Docvqa: A dataset for vqa on document images,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 2200–2209.
47. R. Tito, D. Karatzas, and E. Valveny, “Hierarchical multimodal transformers for multipage docvqa,” *Pattern Recognition*, vol. 144, p. 109834, 2023.
48. R. Tanaka, K. Nishida, and S. Yoshida, “Visualmrc: Machine reading comprehension on document images,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, 2021, pp. 13 878–13 888.
49. S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha, “Docformer: End-to-end transformer for document understanding,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 993–1003.
50. J. Gu, J. Kuen, V. I. Morariu, H. Zhao, R. Jain, N. Barmpalios, A. Nenkova, and T. Sun, “Unidoc: Unified pretraining framework for document understanding,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 39–50, 2021.
51. A. Gemelli, S. Biswas, E. Civitelli, J. Lladós, and S. Marinai, “Doc2graph: A task agnostic document understanding framework based on graph neural networks,” in *Computer Vision – ECCV 2022 Workshops*, L. Karlinsky, T. Michaeli, and K. Nishino, Eds. Cham: Springer Nature Switzerland, 2023, pp. 329–344.
52. G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park, “Ocr-free document understanding transformer,” in *European Conference on Computer Vision*. Springer, 2022, pp. 498–517.
53. B. Davis, B. Morse, B. Price, C. Tensmeyer, C. Wigington, and V. Morariu, “End-to-end document recognition and understanding with dessurt,” in *Computer Vision – ECCV 2022 Workshops*, L. Karlinsky, T. Michaeli, and K. Nishino, Eds. Cham: Springer Nature Switzerland, 2023, pp. 280–296.
54. S. Biswas, P. Riba, J. Lladós, and U. Pal, “Docsynth: a layout guided approach for controllable document image synthesis,” in *International Conference on Document Analysis and Recognition*. Springer, 2021, pp. 555–568.
55. M. Yim, Y. Kim, H.-C. Cho, and S. Park, “Synthtiger: Synthetic text image generator towards better text recognition models,” in *International Conference on Document Analysis and Recognition*. Springer, 2021, pp. 109–124.
56. L. Kang, P. Riba, M. Rusiñol, A. Fornés, and M. Villegas, “Content and style aware generation of text-line images for handwriting recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 8846–8860, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:239999745>
57. S. Schreiber, S. Agne, I. Wolf, A. Dengel, and S. Ahmed, “Deepdesrt: Deep learning for detection and structure recognition of tables in document images,” in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, 2017, pp. 1162–1167.
58. Z. Zhong, J. Wang, H. Sun, K. Hu, E. Zhang, L. Sun, and Q. Huo, “A hybrid approach to document layout analysis for heterogeneous document images,” in *Document Analysis and Recognition - ICDAR 2023*, G. A. Fink, R. Jain, K. Kise, and R. Zanibbi, Eds. Cham: Springer Nature Switzerland, 2023, pp. 189–206.
59. C. Da, C. Luo, Q. Zheng, and C. Yao, “Vision grid transformer for document layout analysis,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 19 462–19 472.
60. A. Banerjee, S. Biswas, J. Lladós, and U. Pal, “Swindocsegmenter: An end-to-end unified domain adaptive transformer for document instance segmentation,” in *Document Analysis and Recognition - ICDAR 2023*, G. A. Fink, R. Jain, K. Kise, and R. Zanibbi, Eds. Cham: Springer Nature Switzerland, 2023, pp. 307–325.

61. S. Biswas, P. Riba, J. Lladós, and U. Pal, “Beyond document object detection: instance-level segmentation of complex layouts,” *International Journal on Document Analysis and Recognition (IJ DAR)*, vol. 24, pp. 269 – 281, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:237309680>
62. T. Shehzadi, K. Azeem Hashmi, D. Stricker, M. Liwicki, and M. Zeshan Afzal, “Towards end-to-end semi-supervised table detection with deformable transformer,” in *Document Analysis and Recognition - ICDAR 2023*, G. A. Fink, R. Jain, K. Kise, and R. Zanibbi, Eds. Cham: Springer Nature Switzerland, 2023, pp. 51–76.
63. K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.
64. T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
65. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: <https://arxiv.org/abs/2010.11929>
66. X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” *arXiv preprint arXiv:2003.04297*, 2020.
67. M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, “Unsupervised learning of visual features by contrasting cluster assignments,” *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.
68. J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Dohersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, “Bootstrap your own latent a new approach to self-supervised learning,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS’20. Red Hook, NY, USA: Curran Associates Inc., 2020.
69. X. Chen and K. He, “Exploring simple siamese representation learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 750–15 758.
70. K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. B. Girshick, “Masked autoencoders are scalable vision learners,” *CoRR*, vol. abs/2111.06377, 2021. [Online]. Available: <https://arxiv.org/abs/2111.06377>
71. K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.
72. Y. Fang, B. Liao, X. Wang, J. Fang, J. Qi, R. Wu, J. Niu, and W. Liu, “You only look at one sequence: Rethinking transformer in vision through object detection,” *CoRR*, vol. abs/2106.00666, 2021. [Online]. Available: <https://arxiv.org/abs/2106.00666>
73. T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” *CoRR*, vol. abs/1708.02002, 2017. [Online]. Available: <http://arxiv.org/abs/1708.02002>
74. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European conference on computer vision*. Springer, 2020, pp. 213–229.
75. Z. Dai, B. Cai, Y. Lin, and J. Chen, “UP-DETR: unsupervised pre-training for object detection with transformers,” *CoRR*, vol. abs/2011.09094, 2020. [Online]. Available: <https://arxiv.org/abs/2011.09094>

76. A. Bar, X. Wang, V. Kantorov, C. J. Reed, R. Herzig, G. Chechik, A. Rohrbach, T. Darrell, and A. Globerson, “Detreg: Unsupervised pretraining with region priors for object detection,” *CoRR*, vol. abs/2106.04550, 2021. [Online]. Available: <https://arxiv.org/abs/2106.04550>
77. S. Liu, Z. Li, and J. Sun, “Self-emd: Self-supervised object detection without imagenet,” *arXiv preprint arXiv:2011.13677*, 2020.
78. O. J. Hénaff, S. Koppula, E. Shelhamer, D. Zoran, A. Jaegle, A. Zisserman, J. Carreira, and R. Arandjelović, “Object discovery and representation networks,” in *European Conference on Computer Vision*. Springer, 2022, pp. 123–143.
79. H. Davoudi, M. Fiorucci, and A. Traviglia, “Ancient document layout analysis: Autoencoders meet sparse coding,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 5936–5942.
80. X. Wu, L. Xiao, X. Du, Y. Zheng, X. Li, T. Ma, and L. He, “Cross-domain document layout analysis via unsupervised document style guide,” *CoRR*, vol. abs/2201.09407, 2022. [Online]. Available: <https://arxiv.org/abs/2201.09407>
81. X. Wang, R. Girdhar, S. X. Yu, and I. Misra, “Cut and learn for unsupervised object detection and instance segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3124–3134.
82. J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.
83. G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph, “Simple copy-paste is a strong data augmentation method for instance segmentation,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2917–2927.
84. D. Dwibedi, I. Misra, and M. Hebert, “Cut, paste and learn: Surprisingly easy synthesis for instance detection,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1301–1310.
85. B. Pfizmann, C. Auer, M. Dolfi, A. S. Nassar, and P. Staar, “Doclaynet: A large human-annotated dataset for document-layout segmentation,” in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 3743–3751.
86. M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, and Z. Li, “Tablebank: A benchmark dataset for table detection and recognition,” 2019.
87. Z. Cai and N. Vasconcelos, “Cascade R-CNN: delving into high quality object detection,” *CoRR*, vol. abs/1712.00726, 2017. [Online]. Available: <http://arxiv.org/abs/1712.00726>
88. C. Szegedy, S. Ioffe, and V. Vanhoucke, “Inception-v4, inception-resnet and the impact of residual connections on learning,” *CoRR*, vol. abs/1602.07261, 2016. [Online]. Available: <http://arxiv.org/abs/1602.07261>
89. H. Yang and W. Hsu, “Transformer-based approach for document layout understanding,” in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 4043–4047.
90. P. Zhang, C. Li, L. Qiao, Z. Cheng, S. Pu, Y. Niu, and F. Wu, “VSR: A unified framework for document layout analysis combining vision, semantics and relations,” *CoRR*, vol. abs/2105.06220, 2021. [Online]. Available: <https://arxiv.org/abs/2105.06220>
91. J. Wang, M. Krundick, B. Tong, H. Halim, M. Sokolov, V. Barda, D. Vendryes, and C. Tanner, “A graphical approach to document layout analysis,” in *International Conference on Document Analysis and Recognition*. Springer, 2023, pp. 53–69.

92. B. Pfitzmann, C. Auer, M. Dolfi, A. S. Nassar, and P. Staar, “Doclaynet: A large human-annotated dataset for document-layout segmentation,” in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '22. ACM, Aug. 2022. [Online]. Available: <http://dx.doi.org/10.1145/3534678.3539043>