

# Palmira: A Deep Deformable Network for Instance Segmentation of Dense and Uneven Layouts in Handwritten Manuscripts

S P Sharan<sup>[0000–0002–6298–6464]</sup>, Sowmya Aitha<sup>[0000–0003–2266–9947]</sup>, Amandeep Kumar<sup>[0000–0002–9292–9778]</sup>, Abhishek Trivedi<sup>[0000–0002–6763–4716]</sup>, Aaron Augustine<sup>[0000–0003–3022–9714]</sup>, and Ravi Kiran Sarvadevabhatla  
 <sup>[0000–0003–4134–1154]</sup>

Centre for Visual Information Technology  
 International Institute of Information Technology, Hyderabad – 500032, INDIA  
<https://ihdia.iiit.ac.in/Palmira>  
 ravi.kiran@iiit.ac.in

**Abstract.** Handwritten documents are often characterized by dense and uneven layout. Despite advances, standard deep network based approaches for semantic layout segmentation are not robust to complex deformations seen across semantic regions. This phenomenon is especially pronounced for the low-resource Indic palm-leaf manuscript domain. To address the issue, we first introduce Indiscapes2, a new large-scale diverse dataset of Indic manuscripts with semantic layout annotations. Indiscapes2 contains documents from four different historical collections and is 150% larger than its predecessor, Indiscapes. We also propose a novel deep network PALMIRA for robust, deformation-aware instance segmentation of regions in handwritten manuscripts. We also report Hausdorff distance and its variants as a boundary-aware performance measure. Our experiments demonstrate that PALMIRA provides robust layouts, outperforms strong baseline approaches and ablative variants. We also include qualitative results on Arabic, South-East Asian and Hebrew historical manuscripts to showcase the generalization capability of PALMIRA.

**Keywords:** instance segmentation · deformable convolutional network · historical document analysis · document image segmentation · dataset

## 1 Introduction

Across cultures of the world, ancient handwritten manuscripts are a precious form of heritage and often serve as definitive sources of provenance for a wide variety of historical events and cultural markers. Consequently, a number of research efforts have been initiated worldwide [6,23,20,11] to parse image-based versions of such manuscripts in terms of their structural (layout) and linguistic (text) aspects. In many cases, accurate layout prediction greatly facilitates downstream processes such as OCR [18]. Therefore, we focus on the problem of obtaining high-quality layout predictions in handwritten manuscripts.

Among the varieties of historical manuscripts, many from the Indian sub-continent and South-east Asia are written on palm-leaves. These manuscripts pose significant and unique challenges for the problem of layout prediction. The digital versions often reflect multiple degradations of the original. Also, a large variety exists in terms of script language, aspect ratios and density of text and non-text region categories. The Indiscapes Indic manuscript dataset and the deep-learning based layout parsing model by Prusty et. al. [23] represent a significant first step towards addressing the concerns mentioned above in a scalable manner. Although Indiscapes is the largest available annotated dataset of its kind, it contains a rather small set of documents sourced from two collections. The deficiency is also reflected in the layout prediction quality of the associated deep learning model.

To address these shortcomings, we introduce Indiscapes2 dataset as an expanded version of Indiscapes (Sec. 3). Indiscapes2 is **150%** larger compared to its predecessor and contains two additional annotated collections which greatly increase qualitative diversity (see Fig. 2, Table 1a). In addition, we introduce a novel deep learning based layout parsing architecture called Palm leaf Manuscript Region Annotator or PALMIRA in short (Sec. 4). Through our experiments, we show that PALMIRA outperforms the previous approach and strong baselines, qualitatively and quantitatively (Sec. 6). Additionally, we demonstrate the general nature of our approach on out-of-dataset historical manuscripts. Intersection-over-Union (IoU) and mean Average Precision (AP) are popular measures for scoring the quality of layout predictions [10]. Complementing these area-centric measures, we report the boundary-centric Hausdorff distance and its variants as part of our evaluation approach (Sec. 6).

The source code, pretrained models and associated material are available at this link: <https://ihdia.iit.ac.in/Palmira>.

## 2 Related Work

Layout analysis is an actively studied problem in the document image analysis community [6,14,4]. For an overview of approaches employed for historical and modern document layout analysis, refer to the work of Prusty et. al. [23] and Liang et. al. [16]. In recent times, large-scale datasets such as PubLayNet [30] and DocBank [15] have been introduced for document image layout analysis. These datasets focus on layout segmentation of modern language printed magazines and scientific documents.

Among recent approaches for historical documents, Ma et. al. [18] introduce a unified deep learning approach for layout parsing and recognition of Chinese characters in a historical document collection. Alaasam et. al. [2] use a Siamese Network to segment challenging historical Arabic manuscripts into main text, side text and background. Alberti et. al. [3] use a multi-stage hybrid approach for segmenting text lines in medieval manuscripts. Monnier et. al. [20] introduce docExtractor, an off-the-shelf pipeline for historical document element extraction from 9 different kinds of documents utilizing a modified U-Net [25]. dhSegment

Table 1: Document collection level, region level statistics of Indiscapes2 dataset.

(a) Collection level stats.				(b) Region count statistics.														
	Train	Validation	Test	Total	Indiscapes (old)				Character Line Segment (CLS)	Character Component (CC)	Hole (Virtual)	Hole (Physical)	Page Boundary (PB)	Library Marker (LM)	Decorator / Picture (D/P)	Physical Degradation (PD)	Boundary Line (BL)	
PIH	285	70	94	449	193				PIH	5105	1079	—	9	610	52	153	90	724
BHOOMI	408	72	96	576	315				BHOOMI	5359	524	8	737	547	254	8	2535	80
ASR	36	11	14	61	—				ASR	673	59	—	—	52	41	—	81	83
JAIN	95	40	54	189	—				JAIN	1857	313	93	38	166	7	—	166	292
Total	824	193	258	1275	508				Combined	12994	1975	101	784	1375	354	161	2872	1179

[21] is a similar work utilizing a modified U-Net for document segmentation of medieval era documents. Unlike our instance segmentation formulation (i.e. a pixel can simultaneously have two distinct region labels), existing works (except dhSegment) adopt the classical segmentation formulation (i.e. each pixel has a single region label). Also, our end-to-end approach produces page and region boundaries in a single stage end-to-end manner without any postprocessing.

Approaches for palm-leaf manuscript analysis have been mostly confined to South-East Asian scripts [29,22] and tend to focus on the problem of segmented character recognition [19,24,11]. The first large-scale dataset for palm leaf manuscripts was introduced by Prusty et. al. [23], which we build upon to create an even larger and more diverse dataset.

Among deep-learning based works in document understanding, using deformable convolutions [7] to enable better processing of distorted layouts is a popular choice. However, existing works have focused only on tabular regions [26,1]. We adopt deformable convolutions, but for the more general problem of multi-category region segmentation.

### 3 Indiscapes2

We first provide a brief overview of Indiscapes dataset introduced by Prusty et. al. [23]. This dataset contains 508 layout annotated manuscripts across two collections - Penn-in-Hand (from University of Pennsylvania’s Rare Book Library) and Bhoomi (from libraries and Oriental Research Institutes across India). The images span multiple scripts, exhibit diversity in language and text line density, contain multiple manuscripts stacked in a single image and often contain non textual elements (pictures and binding holes).

Although Indiscapes was the first large-scale Indic manuscript dataset, it is rather small by typical dataset standards. To address this limitation and enable advanced layout segmentation deep networks, we build upon Indiscapes to create Indiscapes2. For annotation, we deployed an instance of HInDoLA [28] - a multi-feature annotation and analytics platform for historical manuscript layout processing. The fully automatic layout segmentation approach from Prusty et. al. [23] is available as an annotation feature in HInDoLA. The annotators utilize the same to obtain an initial estimate and edit the resulting output, thus minimizing the large quantum of labour involved in pure manual annotation.



Fig. 1: Representative manuscript images from Indiscapes2 - from newly added ASR collection (top left, pink dotted line), Penn-in-Hand (bottom left, blue dotted line), Bhoomi (green dotted line), newly added Jain (brown dotted line). Note the diversity across collections in terms of document quality, region density, aspect ratio and non-textual elements (pictures).

HInDoLA also provides a visualization interface for examining the accuracy of annotations and tagging documents for correction.

In the new dataset, we introduce additional annotated documents from the Penn-in-Hand and Bhoomi book collections mentioned previously. Above this, we also add annotated manuscripts from two new collections - ASR and Jain. The ASR documents are from a private collection and contain 61 manuscripts written in Telugu language. They contain 18 – 20 densely spaced text lines per document (see Fig. 2). The Jain collection contains 189 images. These documents contain 16 – 17 lines per page and include early paper-based documents in addition to palm-leaf manuscripts.

Altogether, Indiscapes2 comprises of 1275 documents - a 150% increase over the earlier Indiscapes dataset. Refer to Tables 1a,1b for additional statistics related to the datasets and Fig. 2 for representative images. Overall, Indiscapes2 enables a greater coverage across the spectrum of historical manuscripts - qualitatively and quantitatively.

## 4 Our Layout Parsing Network (Palmira)

In their work, Prusty et. al. [23] utilize a modified Mask-RCNN [10] framework for the problem of localizing document region instances. Although the introduced framework is reasonably effective, the fundamental convolution operation throughout the Mask-RCNN deep network pipeline operates on a fixed, rigid spatial grid. This rigidity of receptive fields tends to act as a bottleneck in obtaining precise boundary estimates of manuscript images containing highly deformed regions. To address this shortcoming, we modify two crucial stages of the Mask R-CNN pipeline in a novel fashion to obtain our proposed architecture

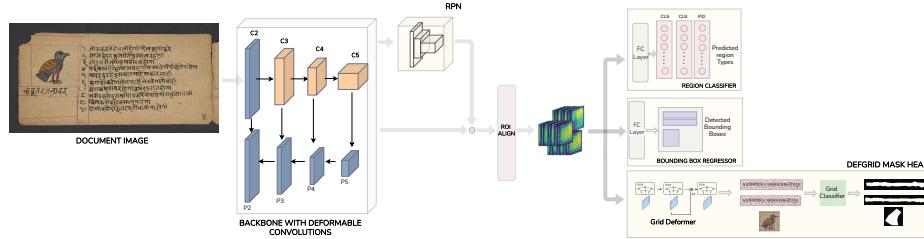


Fig. 2: A diagram illustrating PALMIRA’s architecture (Sec. 4). The orange blocks in the backbone are deformable convolutions (Sec. 4.2). Refer to Fig. 3b, Sec. 4.3 for additional details on Deformable Grid Mask Head which outputs region instance masks.

(see Fig. 2). To begin with, we briefly summarize the Mask R-CNN approach adopted by Prusty et. al. We shall refer to this model as the Vanilla Mask-RCNN model. Subsequently, we shall describe our novel modifications to the pipeline.

#### 4.1 Vanilla Mask-RCNN

Mask R-CNN [10] is a three stage deep network for object instance segmentation. The three stages are often referred to as Backbone, Region Proposal Network (RPN) and Multi-task Branch Networks. One of the Branch Networks, referred to as the Mask Head, outputs individual object instances. The pipeline components of Mask-RCNN are modified to better suit the manuscript image domain by Prusty et al [23]. Specifically, the ResNet-50 used in Backbone is initialized from a Mask R-CNN trained on the MS-COCO dataset. Within the RPN module, the anchor aspect ratios of 1:1,1:3,1:10 were chosen keeping the peculiar aspect ratios of manuscript images in mind and the number of proposals from RPN were reduced to 512. The various thresholds involved in other stages (objectness, NMS) were also modified suitably. Some unique modifications were included as well – the weightage for loss associated with the Mask head was set to twice of that for the other losses and focal-loss [17] was used for robust labelling.

We use the modified pipeline described above as the starting point and incorporate two novel modifications to Mask-RCNN. We describe these modifications in the sections that follow.

#### 4.2 Modification-1: Deformable Convolutions in Backbone

Before examining the more general setting, let us temporarily consider 2D input feature maps  $\mathbf{x}$ . Denote the 2D filter operating on this feature map as  $\mathbf{w}$  and the convolution grid operating on the feature map as  $\mathcal{R}$ . As an example, for a  $3 \times 3$  filter, we have:

$$\mathcal{R} = \left\{ \begin{matrix} (-1, -1) & (-1, 0) & (-1, 1) \\ (0, -1) & (0, 0) & (0, 1) \\ (1, -1) & (1, 0) & (1, 1) \end{matrix} \right\} \quad (1)$$

Let the output feature map resulting from the convolution be  $\mathbf{y}$ . For each pixel location  $\mathbf{p}_0$ , we have:

$$\mathbf{y}(\mathbf{p}_0) = \sum_{\mathbf{p}_n \in \mathcal{R}} \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_n) \quad (2)$$

where  $n$  indexes the spatial grid locations associated with  $\mathcal{R}$ . The default convolution operation in Mask R-CNN operates via a fixed 2D spatial integer grid as described above. However, this setup does not enable the grid to deform based on the input feature map, reducing the ability to better model the high inter/intra-region deformations and the features they induce.

As an alternative, Deformable Convolutions [7] provide a way to determine suitable local 2D offsets for the default spatial sampling locations (see Fig. 3a). Importantly, these offsets  $\{\Delta\mathbf{p}_n; n = 1, 2, \dots\}$  are adaptively computed as a function of the input features for each reference location  $\mathbf{p}_0$ . Equation 2 becomes:

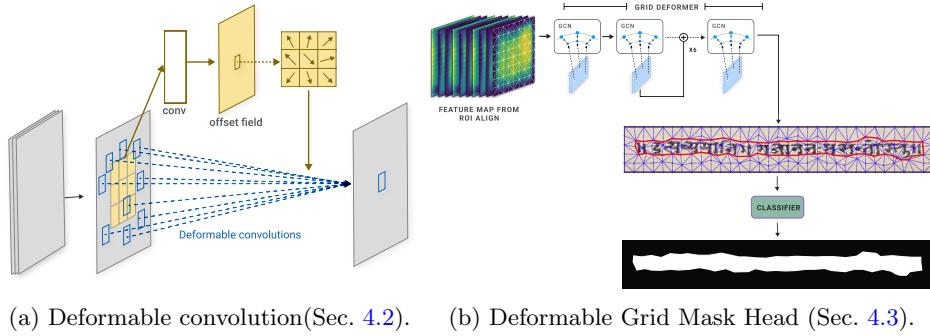
$$\mathbf{y}(\mathbf{p}_0) = \sum_{\mathbf{p}_n \in \mathcal{R}} \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_n + \Delta\mathbf{p}_n) \quad (3)$$

Since the offsets  $\Delta\mathbf{p}_n$  may be fractional, the sampled values for these locations are generated using bilinear interpolation. This also preserves the differentiability of the filters because the offset gradients are learnt via backpropagation through the bilinear transform. Due to the adaptive sampling of spatial locations, broader and generalized receptive fields are induced in the network. Note that the overall optimization involves jointly learning both the regular filter weights and weights for a small additional set of filters which operate on input to generate the offsets for input feature locations (Fig. 3a).

### 4.3 Modification-2: Deforming the spatial grid in Mask Head

The ‘Mask Head’ in Vanilla Mask-RCNN takes aligned feature maps for each plausible region instance as input and outputs a binary mask corresponding to the predicted document region. In this process, the output is obtained relative to a  $28 \times 28$  regular spatial grid representing the entire document image. The output is upsampled to the original document image dimensions to obtain the final region mask. As with the convolution operation discussed in the previous section, performing upsampling relative to a uniform (integer) grid leads to poorly estimated spatial boundaries for document regions, especially for our challenging manuscript scenario.

Similar in spirit to deformable convolutions, we adopt an approach wherein the output region boundary is obtained relative to a deformed grid [8] (see Fig. 3b). Let  $F \in \mathbb{R}^{256 \times 14 \times 14}$  be feature map being fed as input to the Mask



(a) Deformable convolution(Sec. 4.2). (b) Deformable Grid Mask Head (Sec. 4.3).

Fig. 3: Our novel modifications to the Vanilla Mask-RCNN framework (Sec. 4).

Head. Denote each of the integer grid vertices that tile the  $14 \times 14$  spatial dimension as  $v_i = [x_i, y_i]^T$ . Each grid vertex is maximally connected to its 8-nearest neighbors to obtain a grid with triangle edges (see ‘Feature Map from ROI Align’ in Fig. 3b). The Deformable Grid Mask Head network is optimized to predict the offsets of the grid vertices such that a subset of edges incident on the vertices form a closed contour which aligns with the region boundary. To effectively model the chain-graph structure of the region boundary, the Mask Head utilizes six cascaded Residual Graph Convolutional Network blocks for prediction of offsets. The final layer predicts binary labels relative to the deformed grid structure formed by the offset vertices (i.e.  $v_i + [\Delta x_i, \Delta y_i]$ ). The resulting deformed polygon mask is upsampled to input image dimensions via bilinear interpolation to finally obtain the output region mask.

#### 4.4 Implementation Details

*Architecture:* The Backbone in PALMIRA consists of a ResNet-50 initialized from a Mask R-CNN network trained on the MS-COCO dataset. Deformable convolutions (Sec. 4.2) are introduced as a drop-in replacement for the deeper layers C3-C5 of the Feature Pyramid Network present in the Backbone (see Fig. 3a). Empirically, we found this choice to provide better results compared to using deformable layers throughout the Backbone. We use 0.5, 1, 2 as aspect ratios with anchor sizes of 32, 64, 128, 256, 512 within the Region Proposal Network. While the Region Classifier and Bounding Box heads are the same as one in Vanilla Mask-RCNN (Sec. 4), the conventional Mask Head is replaced with the Deformable Grid Mask Head as described in Sec. 4.3.

*Optimization:* All input images are resized such that the smallest side is 800 pixels. The mini-batch size is 4. During training, a horizontal flip augmentation is randomly performed for images in the mini-batch. To address the imbalance in the distribution of region categories (Table 1a), we use repeat factor sampling [9] and oversample images containing tail categories. We perform data-parallel optimization distributed across 4 GeForce RTX 2080 Ti GPUs for a total of 15000

iterations. A multi-step learning scheduler with warmup phase is used to reach an initial learning rate of 0.02 after a linear warm-up over 1000 iterations. The learning rate is decayed by a factor of 10 at 8000 and 12000 iterations. The optimizer used is stochastic gradient descent with gamma 0.1 and momentum 0.9.

Except for the Deformable Grid Mask Head, other output heads (Classifier, Bounding Box) are optimized based on choices made for Vanilla Mask-RCNN [23]. The optimization within the Deformable Grid Mask Head involves multiple loss functions. Briefly, these loss functions are formulated to (i) minimize the variance of features per grid cell (ii) minimize distortion of input features during differentiable reconstruction (iii) avoid self-intersections by encouraging grid cells to have similar area (iv) encourage neighbor vertices in region localized by a reference central vertex to move in same spatial direction as the central vertex. Please refer to Gao et. al. [8] for details.

## 5 Experimental Setup

### 5.1 Baselines

Towards fair evaluation, we consider three strong baseline approaches.

Boundary Preserving Mask-RCNN [5], proposed as an improvement over Mask-RCNN, focuses on improving the mask boundary along with the task of pixel wise segmentation. To this end, it contains a boundary mask head wherein the mask and boundary are mutually learned by employing feature fusion blocks.

CondInst [27] is a simple and effective instance segmentation framework which eliminates the need for resizing and RoI-based feature alignment operation present in Mask RCNN. Also, the filters in CondInst Mask Head are dynamically produced and conditioned on the region instances which enables efficient inference.

In recent years, a number of instance segmentation methods have been proposed as an alternative to Mask-RCNN’s proposal-based approach. As a representative example, we use PointRend [12] - a proposal-free approach. PointRend considers image segmentation as a rendering problem. Instead of predicting labels for each image pixel, PointRend identifies a subset of salient points and extracts features corresponding to these points. It maps these salient point features to the final segmentation label map.

### 5.2 Evaluation Setup

We partition Indiscapes2 dataset into training, validation and test sets (see Table 1a) for training and evaluation of all models, including PALMIRA. Following standard protocols, we utilize the validation set to determine the best model hyperparameters. For the final evaluation, we merge training and validation set and re-train following the validation-based hyperparameters. A one-time evaluation of the model is performed on the test set.

Table 2: Document-level scores for various performance measures. The baseline models are above the upper separator line while ablative variants are below the line. PALMIRA’s results are at the table bottom.

Model	Add-On	HD ↓	$HD_{95} \downarrow$	Avg. HD ↓	IoU ↑	AP ↑	$AP_{50} \uparrow$	$AP_{75} \uparrow$
PointRend [12]	-	252.16	211.10	56.51	69.63	41.51	66.49	43.49
CondInst [27]	-	267.73	215.33	54.92	69.49	42.39	62.18	43.03
Boundary Preserving MaskRCNN [5]	-	261.54	218.42	54.77	69.99	42.65	68.23	<b>44.92</b>
Vanilla MaskRCNN [23]	-	270.52	228.19	56.11	68.97	41.46	68.63	34.75
Vanilla MaskRCNN	Deformable Convolutions	229.50	202.37	51.04	65.61	41.65	65.97	44.90
Vanilla MaskRCNN	Deformable Grid Mask Head	<b>179.84</b>	153.77	45.09	71.65	42.35	69.49	43.16
Palmira : Vanilla MaskRCNN	Deformable Conv., Deformable Grid Mask Head	184.50	<b>145.27</b>	<b>38.24</b>	<b>73.67</b>	<b>42.44</b>	<b>69.57</b>	42.93

### 5.3 Evaluation Measures

Intersection-over-Union (IoU) and Average Precision (AP) are two commonly used evaluation measures for instance segmentation. IoU and AP are area-centric measures which depend on intersection area between ground-truth and predicted masks. To complement these metrics, we also compute boundary-centric measures. Specifically, we use Hausdorff distance (HD) [13] as a measure of boundary precision. For a given region, let us denote the ground-truth annotation polygon by a 2D point set  $\mathcal{X}$ . Let the prediction counterpart be  $\mathcal{Y}$ . The Hausdorff Distance between these point sets is given by:

$$HD = d_H(\mathcal{X}, \mathcal{Y}) = \max \left\{ \max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} d(x, y), \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} d(x, y) \right\} \quad (4)$$

where  $d(x, y)$  denotes the Euclidean distance between points  $x \in \mathcal{X}$ ,  $y \in \mathcal{Y}$ . The Hausdorff Distance is sensitive to outliers. To mitigate this effect, the Average Hausdorff Distance is used which measures deviation in terms of a symmetric average across point-pair distances:

$$\text{Avg. HD} = d_{AH}(\mathcal{X}, \mathcal{Y}) = \left( \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} d(x, y) + \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} d(x, y) \right) / 2 \quad (5)$$

Note that the two sets may contain unequal number of points ( $|\mathcal{X}|, |\mathcal{Y}|$ ). Additionally, we also compute the 95<sup>th</sup> percentile of Hausdorff Distance ( $HD_{95}$ ) to suppress the effect of outlier distances.

For each region in the test set documents, we compute HD,  $HD_{95}$ , IoU, AP at IoU thresholds of 50 ( $AP_{50}$ ) and 75 ( $AP_{75}$ ). We also compute overall AP by averaging the AP values at various threshold values ranging from 0.5 to 0.95 in steps of 0.05. We evaluate performance at two levels - document-level and region-level. For a reference measure (e.g. HD), we average its values across all regions of a document. The resulting numbers are averaged across all the test documents to obtain document-level score. To obtain region-level scores, the measure values are averaged across all instances which share the same region label. We use document-level scores to compare the overall performance of models. To examine

Table 3: Document-level scores summarized at collection level for various performance measures.

Collection name	# of test images	HD ↓	$HD_{95} \downarrow$	Avg. HD ↓	IoU ↑	AP ↑	$AP_{50} \uparrow$	$AP_{75} \uparrow$
PIH	94	66.23	46.51	11.16	76.78	37.57	59.68	37.63
BHOOMI	96	220.38	175.52	46.75	69.83	30.40	50.53	29.03
ASR	14	629.30	562.19	169.03	67.80	51.02	73.09	64.27
JAIN	54	215.14	159.88	38.91	76.59	48.25	70.15	50.34
OVERALL	258	184.50	145.27	38.24	73.67	42.44	69.57	42.93

Table 4: PALMIRA’s overall and region-wise scores for various performance measures. The HD-based measures (smaller the better) are separated from the usual measures (IoU, AP etc.) by a separator line.

Metric	Overall	Character Line Segment (CLS)	Character Component (CC)	Hole (Virtual) (Hv)	Hole (Physical) (Hp)	Page Boundary (PB)	Library Marker (LM)	Decorator / Picture (D/P)	Physical Degradation (PD)	Boundary Line (BL)
$HD_{95}$	<b>171.44</b>	34.03	347.94	70.79	88.33	52.01	289.81	593.99	851.02	111.97
AVG HD	<b>45.88</b>	8.43	103.98	18.80	16.82	13.19	73.23	135.46	255.86	17.95
IoU (%)	<b>72.21</b>	78.01	54.95	74.85	77.21	92.97	67.24	50.57	27.68	61.54
AP	<b>42.44</b>	58.64	28.76	45.57	56.13	90.08	27.75	32.20	03.09	39.72
$AP_{50}$	<b>69.57</b>	92.73	64.55	81.20	90.53	93.99	55.18	54.23	12.47	81.24
$AP_{75}$	<b>42.93</b>	92.74	64.55	81.20	90.52	93.99	55.18	54.24	12.47	81.24

the performance of our model for various region categories, we use region-level scores.

## 6 Results

The performance scores for our approach (PALMIRA) and baseline models can be viewed in Table 2. Our approach clearly outperforms the baselines across the reported measures. Note that the improvement is especially apparent for the boundary-centric measures ( $HD$ ,  $HD_{95}$ , Avg. HD). As an ablation study, we also evaluated variants of PALMIRA wherein the introduced modifications were removed separately. The corresponding results in Table 2 demonstrate the collective importance of our novel modifications over the Vanilla Mask-RCNN model.

To understand the results at collection level, we summarize the document-level scores of PALMIRA in Table 3. While the results across collections are mostly consistent with overall average, the scores for ASR are suboptimal. This is due to the unusually closely spaced lines and the level of degradation encountered for these documents. It is easy to see that reporting scores in this manner is useful for identifying collections to focus on, for improvement in future.

We also report the performance measures for PALMIRA, but now at a per-region level, in Table 4. In terms of the boundary-centric measures ( $HD_{95}$ , Avg. HD), the best performance is seen for the most important and dominant region

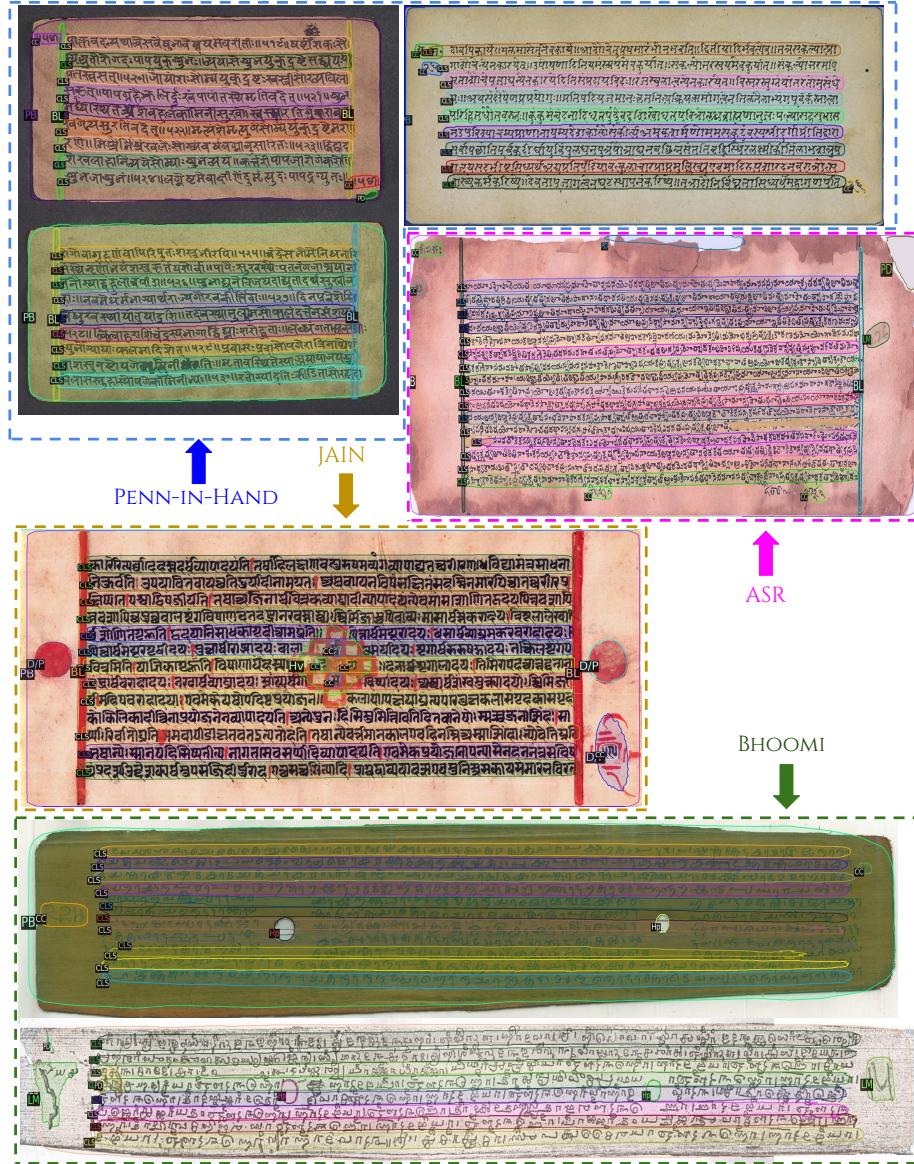


Fig. 4: Layout predictions by PALMIRA on representative test set documents from Indiscapes2 dataset. Note that the colors are used to distinguish region instances. The region category abbreviations are present at corners of the regions.

category - Character Line Segment. The seemingly large scores for some categories ('Picture/Decorator', 'Physical Degradation') are due to the drastically small number of region instances for these categories. Note that the scores for



Fig. 5: A comparative illustration of region-level performance. PALMIRA’s predictions are in red. Predictions from the best model among baselines (Boundary-Preserving Mask-RCNN) are in green. Ground-truth boundary is depicted in white.

other categories are reasonably good in terms of boundary-centric measures as well as the regular ones (IoU, AP).

A qualitative perspective on the results can be obtained from Figure 4. Despite the challenges in the dataset, the results show that PALMIRA outputs good quality region predictions across a variety of document types. A comparative illustration of region-level performance can be viewed in Figure 5. In general, it can be seen that PALMIRA’s predictions are closer to ground-truth. Figure 6 shows PALMIRA’s output for sample South-East Asian, Arabic and Hebrew historical manuscripts. It is important to note that the languages and aspect ratio (portrait) of these documents is starkly different from the typical landscape-like aspect ratio of manuscripts used for training our model. Clearly, the results demonstrate that PALMIRA readily generalizes to out-of-dataset manuscripts without requiring additional training.

## 7 Conclusion

There are three major contributions from our work presented in this paper. The *first* contribution is the creation of Indiscapes2, a new diverse and challenging dataset for handwritten manuscript document images which is 150% larger than its predecessor, Indiscapes. The *second* contribution is PALMIRA, a novel deep network architecture for fully automatic region-level instance segmentation of handwritten documents containing dense and uneven layouts. The *third* contribution is to propose Hausdorff Distance and its variants as a boundary-aware measure for characterizing the performance of document region boundary prediction approaches. Our experiments demonstrate that PALMIRA generates accurate layouts, outperforms strong baselines and ablative variants. We also demonstrate PALMIRA’s out-of-dataset generalization ability via predictions on



Fig. 6: Layout predictions by PALMIRA on out-of-dataset handwritten manuscripts.

South-East Asian, Arabic and Hebrew manuscripts. Going ahead, we plan to incorporate downstream processing modules (e.g. OCR) for an end-to-end optimization. We also hope our contributions assist in advancing robust layout estimation for handwritten documents from other domains and settings.

## Acknowledgment

We wish to acknowledge the efforts of all annotators who contributed to the creation of Indiscapes2.

## References

1. Agarwal, M., Mondal, A., Jawahar, C.: Cdec-net: Composite deformable cascade network for table detection in document images. ICPR (2020) [3](#)
2. Alaasam, R., Kurar, B., El-Sana, J.: Layout analysis on challenging historical arabic manuscripts using siamese network. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 738–742. IEEE (2019) [2](#)
3. Alberti, M., Vöglin, L., Pondenkandath, V., Seuret, M., Ingold, R., Liwicki, M.: Labeling, cutting, grouping: An efficient text line segmentation method for medieval manuscripts. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1200–1206. IEEE (2019) [2](#)
4. Barman, R., Ehrmann, M., Clematide, S., Oliveira, S.A., Kaplan, F.: Combining visual and textual features for semantic segmentation of historical newspapers. arXiv preprint arXiv:2002.06144 (2020) [2](#)
5. Cheng, T., Wang, X., Huang, L., Liu, W.: Boundary-preserving mask r-cnn. In: European Conference on Computer Vision. pp. 660–676. Springer (2020) [8](#), [9](#)
6. Clausner, C., Antonacopoulos, A., Pleitschacher, S.: Icdar2019 competition on recognition of documents with complex layouts-rdcl2019. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1521–1526. IEEE (2019) [1](#), [2](#)
7. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017) [3](#), [6](#)
8. Gao, J., Wang, Z., Xuan, J., Fidler, S.: Beyond fixed grid: Learning geometric image representation with a deformable grid. In: European Conference on Computer Vision. pp. 108–125. Springer (2020) [6](#), [8](#)
9. Gupta, A., Dollár, P., Girshick, R.B.: LVIS: A dataset for large vocabulary instance segmentation. CoRR [abs/1908.03195](#) (2019), <http://arxiv.org/abs/1908.03195> [7](#)
10. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask r-cnn. ICCV pp. 2980–2988 (2017) [2](#), [4](#), [5](#)
11. Kesiman, M.W.A., Pradnyana, G.A., Maysanjaya, I.M.D.: Balinese glyph recognition with gabor filters. Journal of Physics: Conference Series **1516**, 012029 (apr 2020). <https://doi.org/10.1088/1742-6596/1516/1/012029> [1](#), [3](#)
12. Kirillov, A., Wu, Y., He, K., Girshick, R.: Pointrend: Image segmentation as rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9799–9808 (2020) [8](#), [9](#)
13. Klette, R., Rosenfeld, A. (eds.): Digital Geometry. The Morgan Kaufmann Series in Computer Graphics, Morgan Kaufmann, San Francisco (2004) [9](#)
14. Lee, J., Hayashi, H., Ohyama, W., Uchida, S.: Page segmentation using a convolutional neural network with trainable co-occurrence features. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1023–1028. IEEE (2019) [2](#)
15. Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., Zhou, M.: Docbank: A benchmark dataset for document layout analysis. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 949–960 (2020) [2](#)
16. Liang, J., Hu, Q., Zhu, P., Wang, W.: Efficient multi-modal geometric mean metric learning. Pattern Recognition **75**, 188–198 (2018) [2](#)

17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2980–2988 (2017) [5](#)
18. Ma, W., Zhang, H., Jin, L., Wu, S., Wang, J., Wang, Y.: Joint layout analysis, character detection and recognition for historical document digitization. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 31–36. IEEE (2020) [1](#), [2](#)
19. Made Sri Arsa, D., Agung Ayu Putri, G., Zen, R., Bressan, S.: Isolated handwritten balinese character recognition from palm leaf manuscripts with residual convolutional neural networks. In: 2020 12th International Conference on Knowledge and Systems Engineering (KSE). pp. 224–229 (2020). <https://doi.org/10.1109/KSE50997.2020.9287584> [3](#)
20. Monnier, T., Aubry, M.: docExtractor: An off-the-shelf historical document element extraction. In: ICFHR (2020) [1](#), [2](#)
21. Oliveira, S.A., Seguin, B., Kaplan, F.: dhsegment: A generic deep-learning approach for document segmentation. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 7–12. IEEE (2018) [3](#)
22. Paulus, E., Suryani, M., Hadi, S.: Improved line segmentation framework for sundanese old manuscripts. In: Journal of Physics: Conference Series. vol. 978, p. 012001. IOP Publishing (2018) [3](#)
23. Prusty, A., Aitha, S., Trivedi, A., Sarvadevabhatla, R.K.: Indiscapes: Instance segmentation networks for layout parsing of historical indic manuscripts. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 999–1006. IEEE (2019) [1](#), [2](#), [3](#), [4](#), [5](#), [8](#), [9](#)
24. Puarungroj, W., Boonsirisumpun, N., Kulna, P., Soontarawirat, T., Puarungroj, N.: Using deep learning to recognize handwritten thai noi characters in ancient palm leaf manuscripts. In: International Conference on Asian Digital Libraries. pp. 232–239. Springer (2020) [3](#)
25. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015) [2](#)
26. Siddiqui, S., Malik, M., Agne, S., Dengel, A., Ahmed, S.: Decnt: Deep deformable cnn for table detection. IEEE Access **6**, 74151–74161 (2018) [3](#)
27. Tian, Z., Shen, C., Chen, H.: Conditional convolutions for instance segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 282–298. Springer International Publishing, Cham (2020) [8](#), [9](#)
28. Trivedi, A., Sarvadevabhatla, R.K.: Hindola: A unified cloud-based platform for annotation, visualization and machine learning-based layout analysis of historical manuscripts. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 2, pp. 31–35. IEEE (2019) [3](#)
29. Valy, D., Verleysen, M., Chhun, S., Burie, J.C.: Character and text recognition of khmer historical palm leaf manuscripts. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 13–18. IEEE (2018) [3](#)
30. Zhong, X., Tang, J., Yepes, A.J.: Publaynet: largest dataset ever for document layout analysis. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1015–1022. IEEE (Sep 2019). <https://doi.org/10.1109/ICDAR.2019.00166> [2](#)