SelfDocSeg: A Self-Supervised vision-based Approach towards Document Segmentation

Subhajit Maity*[1] **Sanket Biswas***[2] Siladittya Manna[3] Ayan Banerjee[2]
Josep Lladós[2] Saumik Bhattacharya[4] Umapada Pal[3]

[1]Technology Innovation Hub, Indian Statistical Institute, Kolkata, India

[2]Computer Vision Center, Universitat Autònoma de Barcelona, Spain

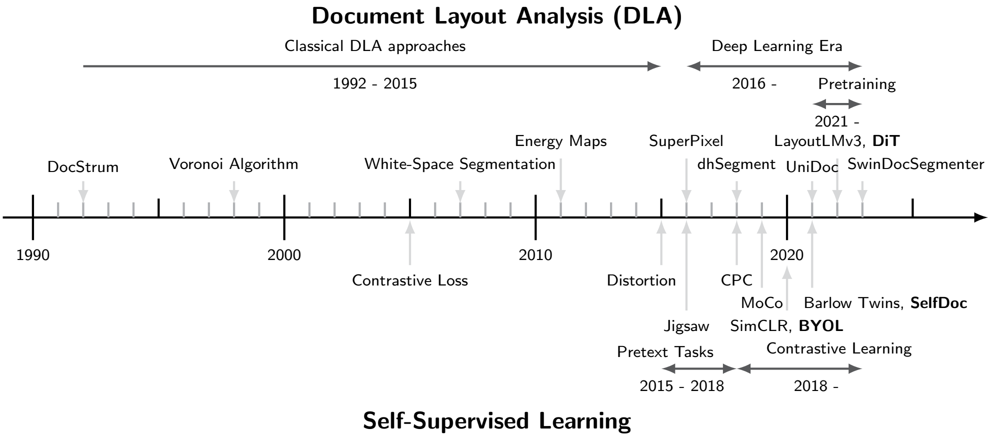
[3]CVPR Unit, Indian Statistical Institute, Kolkata, India

[4]Electronics and Electrical Communication Engineering, Indian Institute of Technology Kharagpur, India

SAN JOSE, CALIFORNIA, USA 2023

Outline

Going Back in Time



Coming back to the Present Day

Where DLA and Self-supervised Learning stand today?

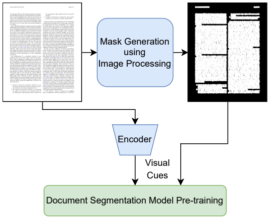
- The classical image processing techniques for DLA had long been lost.
- Modern self-supervision is bringing the classical algorithms, like Felzenszwalb, Normalized Cuts *etc*. back for guidance.
- The current document pre-training techniques for DLA use *multi-modal* approaches and *large-scaled* datasets.

What do we need?

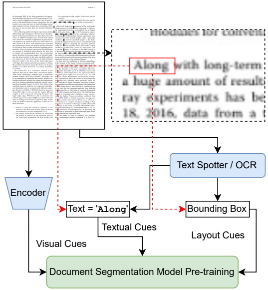
- A strategy to use the document images without annotations for DLA.
- A *data-efficient* pre-training strategy with *unlabelled* document images to alleviate the usual requirement of computational resources.

What's different from prior state-of-the-arts (SOTAs)?

- No textual cue or layout cue from trained OCR. Only visual cues are enough to capture global and local context.
- Superior data efficiency



(a) SelfDocSeg



(b) Existing Self-supervised SOTAs

Contributions

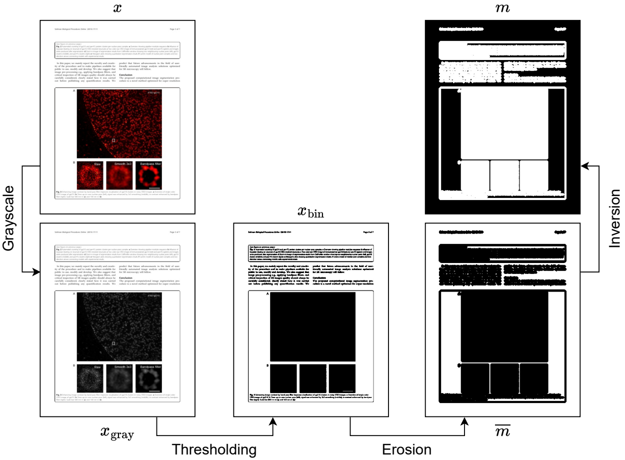
- A novel *vision-based self-supervised* framework, specifically designed to pre-train an image encoder for DLA task.
- A *pseudo physical layout guided strategy* for self-supervision in the region of interest localization for document segmentation.
- A *data-efficient pre-training strategy* to learn multiple document object representations simultaneously in the self-supervised setting

Overall Algorithmic Overview of SelfDocSeg

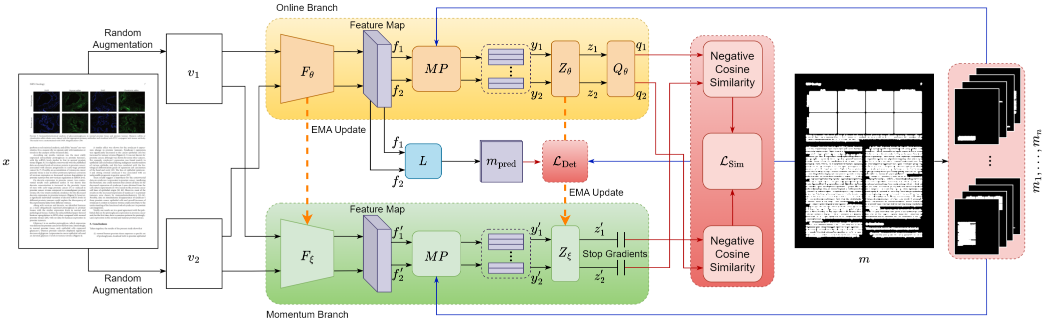
1. First, an approximate layout mask m is generated.
2. The image encoder is pre-trained with the help of m as visual cues in a non-weight-shared two-branch network similar to BYOL [28].
3. The encoder pre-training is done with two objectives - *(a) localization* and *(b) representation* of layout objects.
4. Localization objective is optimized using a detection loss $\mathcal{L}_{\text{Det}}$.
5. Layout object representations are extracted by *mask pooling* operation on encoded feature-maps for each object.
6. Representation objective is learnt via a similarity loss $\mathcal{L}_{\text{Sim}}$.

Mask Generation

1. **Grayscale Conversion:** CIE grayscale conversion on the document image x
2. **Thresholding:** Global thresholding with value 239 for 8-bit integer pixels
3. **Erosion:** 5×5 rectangular kernel
4. **Inversion:** pixel values subtracted from 255 to get the layout mask m



Architecture



Architecture Details

Online Branch

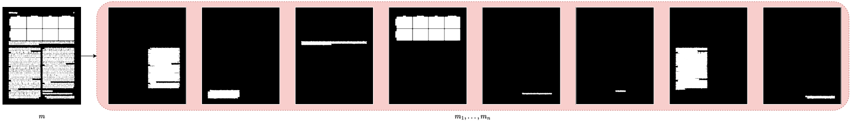
The online branch has an encoder F_θ, a projector Z_θ and a predictor Q_θ. It is updated using backpropagation.

Momentum Branch

The momentum branch has an encoder F_ξ and a projector Z_ξ. It is updated using an exponentially moving average (EMA).

Layout Objects for Mask Pooling

The layout object masks $m_1, \ldots, m_n$ are separated from mask m as separate contours.



Preliminaries

Mask Pooling

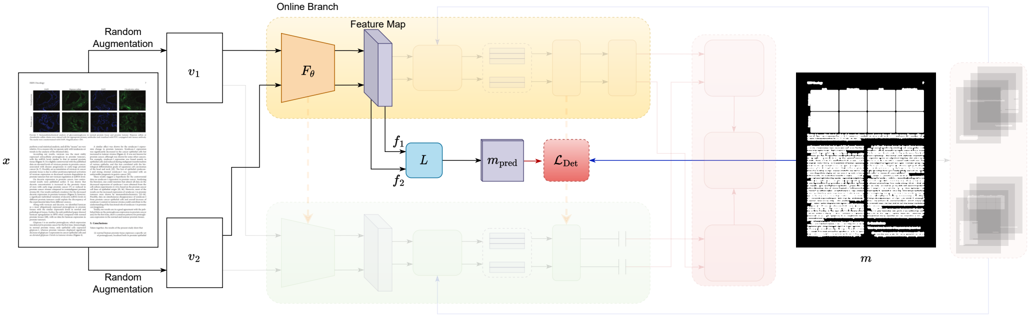
Mask pooling [35] is used for extracting layout object representation from encoded feature maps. It is just an average pooling inside each contour masked by m.

$$y^{(k)} = \frac{1}{\sum_{i,j} m_k[i,j]} \sum_{i,j} m_k[i,j] f[i,j]$$

Weight Updates

$$\textbf{Online Branch:} \quad \theta \leftarrow \text{optimizer}\left(\theta, \nabla_\theta \mathcal{L}_{\text{total}}, \eta\right)$$
$$\textbf{Momentum Branch:} \quad \xi \leftarrow \tau\xi + (1-\tau)\,\theta$$

Learning to Localize Layout Objects



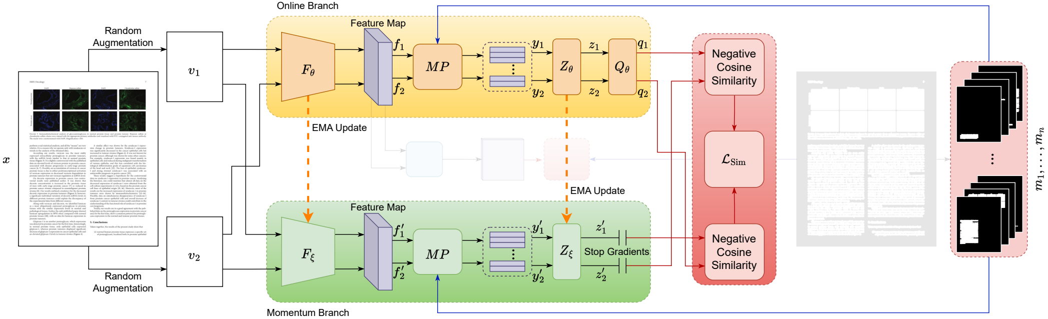
Localization Pre-training Objective

- The mask predictor module L is used to predict the layout mask m_{pred}.
- m_{pred} translates to the pixel-level probability of a layout object being present.
- We formulate this prediction objective as an imbalanced classification task and use Focal Loss [43] $\mathcal{L}_{\text{Det}}$ for the same.

Detection Loss

$$\mathcal{L}_{\text{Det}} = -\frac{\alpha}{\sum_{i,j} m[i,j]} \cdot \sum_{i,j} (m[i,j](1 - m_{\text{pred}}[i,j])^{\gamma} \log m_{\text{pred}}[i,j]$$
$$+ (1 - m[i,j]) m_{\text{pred}}[i,j]^{\gamma} \log(1 - m_{\text{pred}}[i,j]))$$

Learning to Recognize Layout Objects



Representation Pre-training Objective

- Representation of each layout object is extracted from feature maps of encoders F_θ and F_ξ in the online and the momentum branches using masked pooling.
- The encoder is trained using cosine similarity loss [28] $\mathcal{L}_{\text{Sim}}$ between online predictor Q_θ and momentum projector Z_ξ.

Similarity Loss

$$\mathcal{L}_{\text{Sim}} = 4 - 2\left(\frac{\langle q_1, z_2' \rangle}{\|q_1\|_2 \cdot \|z_2'\|_2} + \frac{\langle q_2, z_1' \rangle}{\|q_2\|_2 \cdot \|z_1'\|_2}\right)$$

Experimental Setup

Pre-training Dataset:

- DocLayNet [51]

Fine-tuning Dataset:

- DocLayNet [51]
- PubLayNet [63]
- PRImA [2]
- Historic Japanese [54]

Our Backbone:

- ResNet 50 [34]

Competitors:

- DocSegTr [8]
- LayoutParser [55]
- Biswas *et al.* [9]
- Mask RCNN [33] (Fully supervised)
- LayoutLMv3$_{\text{Base}}$ [36]
- UDoc [29]
- DiT$_{\text{Base}}$ [40]

Our Detector:

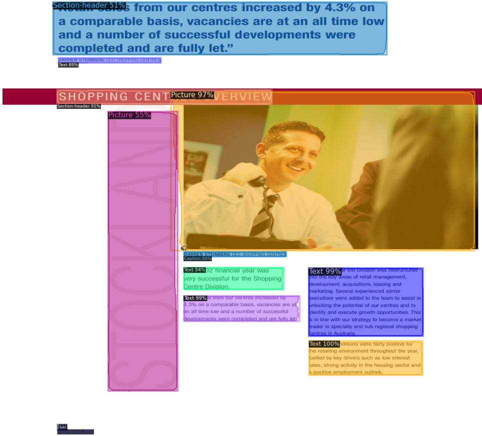
- Mask RCNN [33]

Frameworks: PyTorch, OpenCV, PyTorch Lightning, Lightly, Detectron2
Source code: `https://github.com/MaitySubhajit/SelfDocSeg`

Results - Quantitative

Methods	Cues V	Cues L	Cues T	# Data	DocLayNet mAP	PubLayNet mAP	PRImA mAP	HJ mAP
DocSegTr [8]	✓	✗	✗	-	-	90.4	42.5	**83.1**
LayoutParser [55]	✓	✓	✓	-	-	86.7	**64.7**	81.6
Biswas *et al.* [9]	✓	✗	✗	-	-	89.3	56.2	82.0
Mask RCNN [33]	✓	✗	✗	-	72.4	88.6	56.3	80.1
$\text{LayoutLMv3}_{\text{Base}}$ [36]	✓	✓	✓	11M	-	**95.1**	40.3	82.7
UDoc[29]	✓	✓	✓	1M	-	93.9	-	-
DiT_{Base} [40]	✓	✗	✗	42M	-	93.5	-	-
BYOL [28]	✓	✗	✗	81k	63.5	79.0	28.7	59.8
SelfDocSeg	✓	✗	✗	81k	**74.3**	89.2	52.1	78.8

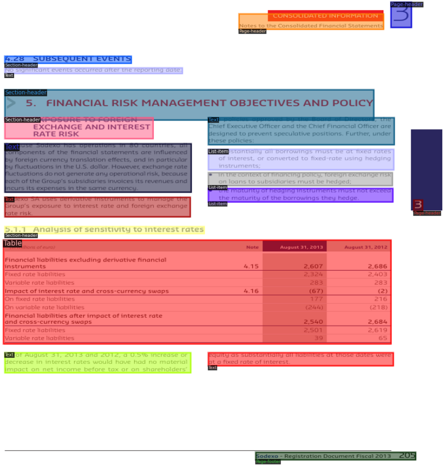
Results - Visual Comparison with Ground-truth



Left: Prediction

Right: Ground-truth

Results - Visual Comparison with Ground-truth



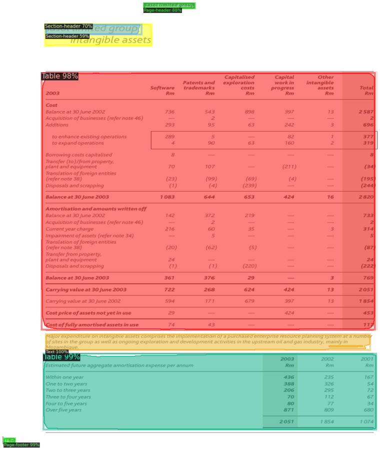
Left: Prediction

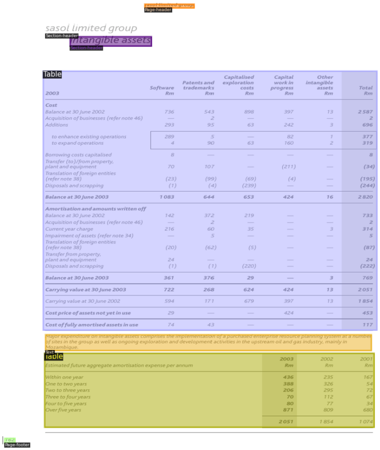
Right: Ground-truth

Results - Visual Comparison with Ground-truth

Left: Prediction

Right: Ground-truth





Left: Prediction

Right: Ground-truth

Discussion & Further Insights

Generalizability & Effectiveness of Pre-training

Experimented with fewer annotations in a semi-supervised setting for fine-tuning.

% Annotations	mAP
10%	41.3
50%	65.1
100%	74.3

Contribution of Detection & Representation Objectives

Experimentation with $\mathcal{L}_{\text{Det}}$ and $\mathcal{L}_{\text{Sim}}$ in strip-down style.

Loss	mAP
w/o $\mathcal{L}_{\text{Sim}}$	39.1
w/o $\mathcal{L}_{\text{Det}}$	69.7
Combined ($\mathcal{L}_{\text{total}}$)	74.3

The Story so far

- **SelfDocSeg** being a *OCR-free pertaining strategy* is computationally less expensive, as large-scale OCR systems are difficult and time-consuming to train.

- Ideally **SelfDocSeg** provides a *wide range of flexibility with backbones and detectors* and can achieve better performance with a stronger backbone.

- **SelfDocSeg** is *data-efficient* compared to state-of-the-art pre-training methods for DLA and can *generalize quickly over different domains* in the fine-tuning stage.

What's Next?

- Although being a superior pre-training strategy in terms of data efficiency, **SelfDocSeg** is not yet tested for data-hungry models like LayoutLMv3 or DiT.
- Following the recent trends to promote performance improvement, the scope of exploration remains in multi-domain and multi-lingual pre-training for introducing data variety.
- **SelfDocSeg** is particularly designed to cater to DLA. Thus, pre-training and fine-tuning for multi-task settings in downstream with vision-based self-supervision can be explored.

Conclusion

- **SelfDocSeg** is *self-supervised* and *vision based* and is motivated by recent self-supervised works in computer vision.
- Self-supervision for documents is possible in a unimodal setting without dependency on OCR or textual knowledge.
- The complete visual representation approach of **SelfDocSeg** facilitates superior understanding and encoding of the visual modality in the layout-guided document understanding paradigm.

To Know More

arXiv: `https://arxiv.org/abs/2305.00795`

Code: `https://github.com/MaitySubhajit/SelfDocSeg`

Project Page: `https://maitysubhajit.github.io/SelfDocSeg`